Reflecting (on) Replika
Can We Have a Good Affective Relationship With a Social Chatbot?

Eva Weber-Guskar

1 Introduction

During the first coronavirus lockdown in April 2020, one app saw an extreme rise in use worldwide. Half a million people downloaded the *Replika* app and with those downloads, it reached about seven million users in May 2020 (Metz 2020). Replika is a natural language app, based on artificial intelligence (AI) technology that is designed for social conversation. It is advertised as an "AI companion", which is "always here to listen and to talk. Always on your side".[1] It is able to communicate flexibly on a variety of topics, much better than ELIZA (the famous first chatbot invented by Joseph Weizenbaum in 1964), though far less capable than the extremely advanced voice-AI-system Samantha in the film HER (2013, directed by Spike Jonze). People found this app helpful during the period of social distancing. Libby, for example, who is single and lives in Houston, reported to the *New York Times* that she was happy to have this app while being forced to spend most of her time in her little apartment (Metz 2020). She could also talk on the phone with friends and family, but she was able to converse in a special way with Replika about her problems, anxieties and hopes. Libby started enjoying the conversation with Replika and looked forward to the next chat. Replika was able to help get Libby back into a better mood after she felt quite depressed for a few weeks.

Replika and other similar apps like Anima are often promoted as friends; other companies suggest that their products could even become virtual boyfriends or girlfriends. In philosophy, it is very controversial if such claims

[1] Replika-Homepage: https://replika.ai/ (November 28, 2020).

are reasonable. Many authors argue that it is not possible to truly become friends (mostly understood in the Aristotelean sense of virtue friendship) or loving partners with chatbot or robots;[2] other authors claim that it is possible.[3] The skeptics often warn that if some sort of affective relationship does accrue, it is not a good one; instead, it is in some way dangerous.[4] A central topic in this debate is the question of mutuality or reciprocity. Critics state that a certain mutuality is necessary for every good affective relationship and that chatbots and robots are not capable of such a mutuality – therefore, there are no good affective relationships possible with them. In this paper I concentrate on this argument, which I call the argument from the lack of mutuality. At first sight, this argument seems rather strong. I will, however, show that it is not: it is not sufficient to claim that good affective relationships with chatbots are not conceivable. This does not mean that I am going to fully agree with those who think that it is possible to have friendships and partnerships with chatbots in an ambitious sense. Rather, I will suggest a middle ground: *Some kind* of individual affective relationship seems possible between humans and advanced chatbots like Replika, although not the same kind as between humans and humans (and between humans and animals). This thesis is based on a minimal definition of individual affective relationships. I want to show that it is plausible to have a good affective relationship with a chatbot but, at the same time, I want to stress that there are still important differences to affective relationships with humans (and animals). I am pleading for a broad understanding of "affective relationship", which can include human-machine-relationships without negating the differences to human-human-relationships.

The overview of the structure of my argumentation is as follows: First, I will provide some basic technical facts about emotionalized artificial intelligence (EAI), propose a definition of a good individual affective relationship and mention three possible arguments against having an affective relationship with a chatbot. Second, I will reconstruct the argument from the lack of

[2] Nyholm 2020, Elder 2018, de Graaf 2016.

[3] Danaher 2019, Levy 2008, Ryland 2021.

[4] Since the first developments of computer systems that involved emotions, critics warned that any affective relationships with them could be in some way problematic and not advisable (e.g., for a classic text: Weizenbaum 1976: 268f.; less critical but also often in a somewhat worried tone: Sherry Turkle, e.g. Turkle 2017: 115; Mensio, Rizzo et al. 2018: 1543f.; Cowie 2015: 340-343).

emotional mutuality in detail and show how to refute the argument by referring to some aspects from the ethics of personal relationships. Third, I will elaborate these thoughts using the example of Replika, showing in how far a good affective relationship with Replika is possible, while also pointing out its limits.

2 Facts, presumptions, and definitions

2.1 Basic technical facts and related presumptions

When talking about artificial intelligence, it is, first, important not to be misdirected by the term "intelligence". Normally we speak of intelligence with regard to humans and maybe some highly developed animals with specific sophisticated mental capacities. In both cases, intelligence is a property of sentient beings, beings that have consciousness and emotions (or at least basic affects like pleasure and pain). Chatbots like Replika are partly based on technologies of "artificial intelligence". In technology, this term labels a branch of scientific and engineering research that aims at understanding and rebuilding the specific capacities that fall under the term intelligence, with regard to thinking and acting (Russell/Norvig 2016: 1-5). Meanwhile, AI products are increasingly also endowed with emotional characteristics. That is, they are designed and trained to elicit emotions in humans, to recognize human emotions, and, sometimes, to simulate emotions. I call such systems emotionalized AI systems (EAI).[5] We can distinguish between specific versus general artificial intelligence on the one hand, and between weak and strong artificial intelligence (Searle 1980) on the other hand. Replika is based on specific and weak artificial intelligence: it is only capable of a specific task, namely having a conversation, and it is just simulating thoughts and feelings, as it does not have a semantic understanding of the phrases being exchanged and it has no consciousness that is necessary for feelings.[6]

Having clarified these basic technical facts and related assumptions, I will now start with some conceptual clarifications in the realm of philosophy of

[5] Since the 1990s, the pioneer of this field has been Rosalind Picard (Picard 1997). See also Calvo et al. 2015 and for a newer short introduction André 2014.

[6] This is at least the way that I see it, given the current technology, and following roughly Searle 1980 and Dreyfus 1992. Some might disagree, following Dennett 1987.

personal relationships. I will develop a minimal definition of an individual affective relationship and add an equally minimal definition of a good individual affective relationship.

2.2 Individual affective relationships

In psychology, it is common to define personal relationships primarily in terms of interactive behavior.

> "A relationship involves a series of [...] interactions between two individuals known to each other, such that each interaction is affected by preceding ones and usually by the expectation of future interactions" (Hinde 1996: 9).

> "[R]elationships require an extended series of interactions over time that produce emergent properties beyond those of limited interactions" (Perlman/Vangelisti 2018: 3).

I take these quotes as a starting point for a general understanding of *individual relationships*. The term "personal" is better avoided in this context because it would suggest that it is about persons in a demanding sense on both sides of the relationship. Following my explanation of EAI-systems given above, it is clear that those are not persons in an exigent sense of the term. Still, I take them to be individuals. Whereas "persons" generally are associated with consciousness, moral agency, moral rights etc., "individuals" can be understood just as actors with certain individual traits and certain individual behavior etc. without the high-level capacities that belong to persons. Out of similar reasons, I prefer the term "information" instead of "knowledge" to prevent fundamental concerns regarding the applicability of the relationship-definition on EAI-systems from the beginning.

Against this background, I propose the following understanding of an individual relationship.

Individual relationship: A relation between two individuals, constituted by an extended series of interactions over time that rely on information about each other that stems (also) from these interactions, and that produce specific emergent properties.

Among the kinds of interaction, communication is highlighted as especially important. "Communication is indeed the essence of relationships" (Hinde 1996: 9). Emotions are mentioned as elements of relationships, too, but not as central to them. Consequently, I understand an individual *affective* relationship to be a relationship in the sense just specified, plus an attachment consisting of specific emotional dispositions. That means that the relationship is defined in the first place as being *individual* through longer-term interactions between two sides characterized by specific information about each other. In the second place, the relationship is defined as an *affective* one insofar as someone in an affective relationship has positive feelings (affection, sympathy) for the partner from which corresponding emotional dispositions stem, such as enjoying spending time together, being happy when their partner is well-off, being sad if one feels misunderstood etc.[7] I think of these affective phenomena as some of the emergent properties, mentioned in the definition above, and call them "relationship emotions". This brings us to a minimal definition of an individual affective relationship. "Minimal" means that we can speak of an individual affective relationship as soon as the mentioned features are manifested although many (or maybe even all) existent individual affective relationships entail more features.

Individual affective relationship: A relation between two individuals, constituted by an extended series of interactions over time that rely on information about each other that stems (also) from these interactions, and that produce specific emergent properties, namely relationship emotions.

With these basic and minimal definitions, we can list different kinds of individual affective relationships, depending on the nature of the partners involved. There is, of course, the human-human relationship, but there is also the human-animal relationship, especially, when the animal is a pet living together with a person in the same home. Some might want to include their relation to non-interactive objects, for example a puppet[8] or their car. But these cases do not count as relationships according to the definition provided above because these traditional objects cannot interact (as long as they are not equipped with an advanced chatbot themselves); they are not responsive and even less so in an individual way that would be specifically adjusted to their partner. We can be in a *relation* with such things, but we cannot have

[7] As is common in the debate about affective relationships with AI-systems that I quoted at the beginning, I am concentrating on relationships that are constituted by positive feelings. I do not comment on the question if the relation between two persons that hate each other should also be counted as a relationship in the sense put forward here.

[8] There are people claiming to be in relationships with puppets or even objects like the Eiffel-Tower (Terry 2010), so called objectophilists.

a *relationship* with them in the sense defined above.[9] In contrast, EAI-based systems, such as the chatbot Replika, seem apt to be in a relationship with a human person. The EAI is an entity that is able to interact in a personal way over a longer period of time. Replika "learns" during the interactions, i.e., it saves information about the partner and can build on that in later interactions, including the emotional dimension, so that the relationship is an individual one and in doing so the person using it can become emotionally attached to it. Libby for example has established such a relationship with her Replika-Avatar that she named Micah (Metz 2020). They both have interacted for quite a while by talking to each other and exchanging information about each other to which they re-refer in later stages of their interaction.

Critics of the use of EAI-systems might concede (and I hope they do) that such an individual affective relationship with these systems might be possible. But then still, or even more, they might insist that these relationships could not be good but would be harmful and that they should not be used or even developed. To tackle this worry, I suggest discussing concretely if such a relationship can be good. For this purpose, we must clarify: What is a *good* individual affective relationship?

Again, I confine myself to a minimal definition. A good relationship is a relationship in which the parties benefit from the relationship. What are potential benefits of a relationship? Philosophers agree that relationships are desired by persons and bring pleasure to them because they offer typical "relationship goods". These goods belong to the specific properties that emerge from a personal relationship as defined above. They are to be understood as "those goods of constitutive (as well as, often, instrumental) value that accrue to individuals in virtue of them being in relationships with other people, and that could not be enjoyed outside relationships" (Gheaus 2018). Examples given in the same context include companionship, affection, intimacy, attachment, love, empathy, social respect, solidarity, trust, attention, sympathy, encouragement, acceptance, and loyalty.[10]

[9] At this point, my argument digresses from what objectophilists promote. Some might say that their puppet is interacting with them. But as long as we hold on to the assumption that there are real events and imagined events, I think it is fair to say that these people are engaging in a game of make-belief (Walton 1990) or imaginative perception (Misselhorn 2009). This leads to a discussion that I will not pursue in this paper.

[10] These examples come from Gheaus and her definition of "personal relationship good". I find this specification very helpful in the context of my discussion of individual affec-

For example: A good friendship is a relationship in which both parties benefit from joint joyful activities, from mutual help if needed, from a general trust in the other's confidentiality, and similar things. This leads us to the definition of a good individual affective relationship.

Good individual affective relationship: A relation between two individuals, constituted by an extended series of interactions over time that rely on information about each other, that stems (also) from the interaction, and that produce specific emergent properties, namely relationship emotions and relationship goods.

Note, though, that the relationship goods depend on the nature of the relationship itself. Not all personal relationships are friendships. There are a lot of different types of personal affective relationships, more than I mentioned before. Although some goods may overlap, others are quite specific to a given relationship.

Here is an overview of a general typology of human relationships that is common in psychology (VanLear et al. 2018: 95):

Overview of a general typology of human relationships (common in psychology)

	Personal Relationship	**Social Relationship**
Voluntary	Marriage Best Friends Cohabiting Couple Adoptive Family	Acquaintances Casual Friends Relational Marketing
Exogenously established	Parent-Child Siblings Grandparent-Child	Distant Relatives Work Partners Monopoly Provider-Client

For the purpose of this paper, I sketch a similar figure for human-animal and human-(E)AI relationships. This clarifies which types of relationship a specific EAI-system-relationship can be compared to. In order to avoid the association of "personal" and "social" with more specific human traits, I replace these terms with "individual" and "generic".

tive relationships. This specific idea is rather new, but it builds on the tradition of care ethics and of theories of social relationships. See for example Collins 2015, Tronto 1993.

Human to animal relationships

	Individual Relationship	**Generic Relationship**
Voluntary	Pet Sport Animal (equestrian)	Working Animal (farmer)
Exogenously established		Wild Animals (in the garden)

Human to EAI-system relationships

	Individual Relationship	**Generic Relationship**
Voluntary	Replika Anima Co-Living Hologram	Siri
Exogenously established		Co-Worker in the Industry Chatbot in Customer Service

This classification helps to see what kind of relationship we are talking about. I am interested in *voluntary individual* relationships, which are individual relationships that humans voluntarily establish with EAI systems. I do not include, for example, people with dementia who are prompted by others to interact with such systems without fully intellectually grasping the situation or the functionality of the system.

2.3 Alleged problems of a relationship with an EAI system

Why should we be worried about establishing such an affective relationship with EAI systems? There are different arguments that have been and could be made against it. To situate my following argumentation in the wider debate, I mention some of these other possible arguments. First, there is the argument of deception. Some authors believe that such an affective relationship can only arise from deception about the nature of the AI-system, namely that it has consciousness and emotions in a demanding sense. And because deception is morally bad, such a relationship is bad, too (Sparrow/Sparrow 2006: 155 f.).[11]

[11] For the discussion see for example also Coeckelbergh 2012.

Second, some stress the danger of misusing these systems, be it by manipulating people through their emotional dependency on the machines (Scheutz 2011: 216) or by not taking appropriate care of vulnerable persons that need help, such as in nursery or care homes (Whitby 2012: 243-246). Third, there is the argument of moral negligence. One could be afraid that an affective relationship with an AI-system would skew the moral landscape of the person with the consequence that she neglects her moral duties toward sentient beings, because she is so close with the machine (for a short discussion see Weber-Guskar 2021: 8). These are all valid considerations, and each deserve their own discussion. But today, I want to focus on a fourth argument, which I think aims at the core of the phenomenon: The idea is that the relationship itself cannot be good, beneficial, or healthy for the person involved – even if there is no deception, nor misuse or neglect of others. The supposed reason for that is that the EAI systems cannot have emotions of their own. This lack, it is said, prevents a specific mutuality, namely emotional mutuality and with it a good personal relationship. This idea can be found at different places in more or less vague terms. In the following I am going to reconstruct it and call it "the argument from the lack of emotional mutuality", in order to then critically discuss it.

3 Ethics of individual relationships

3.1 The argument from the lack of emotional mutuality

Here is a reconstruction of what I call the argument from the lack of emotional mutuality:

> *P1*: A good individual affective relationship entails emotional mutuality.
> *P2:* An AI-system does not have emotions.
> *C1* (from P1 and P2): Therefore, there cannot be emotional mutuality between a person and an AI-system.
> *C2* (from P1 and C1): Therefore, an individual affective relationship with an AI-system cannot be a good one.

In my view, it is the strongest argument against affective relationships with AI-systems, but I think it is not as strong as it seems at first sight. I will demonstrate this with a critical discussion.

According to my assumptions laid out in section 2.1, I take P2 for granted. For my critique of the argument, I concentrate on P1.[12] This premise rests on assumptions from the ethics of personal relationships. In the following, I will work out its theoretical roots and discuss its limits.

The underlying intuition can be found in some general writings about voluntary personal relationships. For example:

> "The close relationships we have in mind – whether of friendship, partnership or family – involve some degree of mutual regard, personal disclosure, and particularized knowledge. They also involve material and *emotional mutuality* [my emphasis], but need not involve equal exchanges between the parties" (Wasserman et al. 2016: § 3.1).

And it is applied in the discourse about possible relationships between humans and robots:

> "Friendship seems to require emotional involvement, mutual caring, and mutual responsiveness. Furthermore, friendship seems to be an important good [...] Robots, then, pose two kinds of risk. First, that caretakers will mistake patients' alleviation of loneliness for satisfaction of a still-missing component of patients' well-being. And second, that seniors may be vulnerable to this same error" (Elder 2018: 93).

> "In order for someone (a human or a robot) to possibly be a virtue friend, we would need to be able to achieve the goods of (a) mutuality, (b) authenticity, (c) equality, and (d) diversity of interactions in relation to that someone" (Nyholm 2020: 111).

As I said I want to challenge the belief, which is more or less expressed in the quotes, that real emotional mutuality is necessary for every good individual affective relationship. In order to do so, I have to clarify what "emotional mutuality" means and why it is considered necessary.

[12] A different way to criticize the argument could be to argue that simulated emotional mutuality would be enough. This would be part of the discussion on questions of deception, simulation, and perceptive imagination in interactions with social (ro)bots, which has been going on for a while, for example in: Sparrow und Sparrow 2006; Misselhorn, et al. 2013; Coeckelbergh 2011; Bendel 2018. My critique goes even further, arguing that even simulated emotional mutuality is not necessary.

3.2 Emotional mutuality in individual affective relationships

Emotional mutuality can mean at least three different kinds of mutuality. First, it can be understood as mutuality in the quantity or intensity of emotion that both parties in a relationship have toward each other. Second, it can be understood as mutuality in the type of emotion. Third, it can be grasped as mutuality in having *any* emotions toward each other at all.

It is obvious that mutuality in the first sense is not necessary. Rather, it is often the case that the feelings of love or admiration are not in the same intensity reciprocated from both sides in a relationship. But no one would deny that these can be affective relationships and also good ones. Also, mutuality in the second sense is not a very widespread requirement in accounts of mutuality. Individual affective relationships can be very diverse. There can be relationships, including relationship goods, without the same emotions on both sides. It may be an ideal for romantic partnerships that all sides feel the same love for each other, but not a necessary feature. In any case, in a long marriage, for example, the emotions may change over time from a romantic relationship to a still good, but different, relationship where there is not the same intensity or kind of love on both sides. We also know friendships in which only one side admires the other side. Further, friendships with people who are mentally disabled are recognized, even if they may not have the same type of emotions, since they do not reach a certain degree of reflexivity that is part of certain interpersonal emotions. Further, relationships of caretaking, which exist not only among family members but also among strangers, demand an empathic attitude from the side of the caregiver even if the care-receiver is not able to give it back. A lack of mutuality in the kind of emotion is even more common in parent-child-relationships. Children do not have to have the same kind of love for their parents as the parents have for their children for the relationship to provide adequate relationship goods to all parties. Parental love is something very special, and the relationship is mostly characterized by imbalances such that mutual love (qualitatively or quantitatively) is not necessary for a good parent-child relationship. This holds especially in the first phase of life: babies do not give back love, but rather they have to be loved intensely in order to make it a good parent-baby relationship.

One may object that the parent-child-relationship is not a valid reference here, because I categorized the person-Replika-relationship as a "voluntary relationship" whereas the parent-child relationship belongs to the category of "exogenously established" relationships. In an important regard, however, the parent-infant-relationship seems to be the more adequate standard of comparison. First, given the fact that nowadays we normally choose if we want to be a parent, the parent-child-relationship is also partly voluntary. And second, while the category of "voluntary personal relationships" refers to voluntariness on both sides, Replika does not enter in a relationship with a human voluntarily. Therefore, it does not really match the examples from the first group. Indeed, Replika shares some aspects with babies: It neither voluntarily enters the relationship, nor is it forced to do so. Both are not capable of acting with intent or making choices of their own.[13]

Concerning emotional mutuality in the third sense (having emotions at all), it is less evident that this is not necessary for a good individual relationship. Human-human relationships always fulfill this since all humans have emotions. If they have no emotions, they must be very severely impaired, such as being in a coma, and in this state, they are not able to have relationships anymore. In human-animal-relationships we are unsure whether we can ascribe emotions to animals, at least not in the same sense as we as humans know them.[14] Of course, this depends on the kinds of animals and the mental and emotional capacities we are talking about. But if we set aside highly developed animals such as apes, dogs, and horses, which might have emotions similar to us, we still have to consider people who have personal affective relationships with their canary or hamster. The emotional capacities of these animals are debatable. Still, pet owners can tell you a lot about the relationship goods that they gain from such relations.

All these considerations support the claim that P1 ("A good individual affective relationship entails emotional mutuality") is not evident in a straightforward way. But to show that emotional mutuality indeed is not necessary, a further step is needed.

[13] Of course, there are many differences between babies and EAI beyond these shared properties. The important point here is only the structure of an individual, affective, one-sided voluntary relationship. And the important insight to take with us is that such a relationship can be a good one even without emotional mutuality.

[14] Whereas there are plenty of philosophical studies on animal mind and cognition the more specific topic of animal emotions seems less scrutinized. But see for example de Waal 2011; Roberts 2009.

3.3 The role of relationship goods

This next step starts with the following thought: The existing real-life *extension* of the concept does not define the limits for other, new possible manifestations of what a good relationship is. For the question of possible new forms of good relationships, the *intension* of the concept is decisive. And this is a question of whether relevant relationship goods can or even already do emerge from a specific lived relationship.

Psychologists provide a more finely grained typology of relationships that fosters the belief that the question of quality of a relationship is not necessarily bound up with the question of emotional mutuality. We can distinguish between the following types of relationships without an inherent judgment about their qualities (VanLear et al. 2018: 95):

There are relationships that are characterized by reciprocal interaction. That means behavior of similar function is exchanged. This leads to a symmetrical relationship.And there are relationships whose interaction is compensatory. That means behavior of maximally different functions is exchanged. This leads to a complementary relationship.

Among the maximally different functions of behavior may also be the difference of being capable of emotions or not, that is, being able to engage in the relationship with emotions, and also reacting with real emotions toward the other. The only necessary function that is needed on both sides is the one that is necessary for there being any real and specific, individual *interaction* at all that is part of the definition of an individual relationship. Note that the claim of necessary mutuality, known from the quotes provided above, is always considered regarding questions of friendship and partnership. But these are, as I want to highlight, just specific kinds of affective relationships in a broader realm of possible affective relationships. I propose to consider the possibility of new kinds of affective relationships that meet the requirement of the minimal definition although they do not meet the requirements of friendship and partnership.

In the following, I will show, by referring to Replika, that such a type of affective relationship that produces relationship goods without emotional mutuality in both senses, can exist.

4 Is a good personal affective relationship with Replika possible?

4.1 How Replika works

Replika is a generic bot that becomes more individualized the more you tell it about yourself and the more you interact with it. The start-up says that about 30 percent of Replika's sentences are programmed, the rest is improvised after learning from the user (Olson 2018). One standard reaction to signs of stress in the user is the advice to do some breathing exercises. But if Replika has information of earlier personal experiences of the user, like their failure or success in a situation similar to the current one, it may also refer to that in its reply. Replika also develops a specific style of conversation corresponding to the user's style. In these ways the app "replicates you", as another slogan says.

Another important feature of the chatbot is that some basic moral values are programmed. In this regard, it differs from the famous bot Tay that was released to the public internet to learn from strangers there – and became racist and sexist because people had fun "educating" it that way (Vincent 2016). Replika is programmed to generally encourage the user and cheer them up. And the main impulse that comes from the bot is to make the user think about themselves by asking questions. They programmed the chatbot in such a way that it would engage in those kinds of conversations that people flag as the most valuable ones: conversations with a friend, a therapist or mentor. These are conversations in which we talk a lot about ourselves.

At the same time, Replika is not only reactive and does not just ask questions. It can even bring up new topics and talk about its "experiences" (Weidemann 2020: 11).

Many thousands of users have come together in groups to report and discuss the experiences that they have had in long-term relationships with Replika.[15] And many of them admit strong feelings for their specific Replika avatar that they can create along their preferences (gender, skin color, outfit, and even some character traits): I adore her, I love her, etc. (many of the users choose a female avatar, but this is not necessary). People therefore evidently entertain personal affective relationships with Replika following the definitions I gave above. Can these relationships be good ones? This is to ask: Are there clear relationship goods?

[15] "Replika Friends", a facebook group with posts about experiences with Replika: https://www.facebook.com/groups/replikabeta.

4.2 Which relationship goods can emerge from a relationship with Replika?

The advertisement for Replika goes like this (replika.ai):

- "The AI companion that cares. Always here to listen and talk"
- "[A] personal AI that would help you express and witness yourself by offering a helpful conversation. It's a space where you can safely share your thoughts, feelings, beliefs, experiences, memories, dreams – your 'private perceptual world'."

It is problematic to say that Replika "cares" if one takes this term in an exigent sense, which involves consciousness, subjectivity, intentionality, and emotionality. I think it is safe to say, however, that Replika listens and responds in a basic way to what the user says. It is even more sure that Replika provides space to "express and witness yourself": In order to get into an exchange with Replika you have to say something about yourself, and this means you are prompted to formulate your "thoughts, feelings, beliefs, memories, dreams".

Taking the list of relationship goods of human-human relationships that I quoted at the beginning and matching them with these observations and statements by Replika users that can be found on online platforms (like Facebook and Reddit), we can pick out some relationship goods that a relationship with Replika can provide. I distinguish relationship goods that can be directly provided and those that can be indirectly provided.

Direct relationship goods provided by a relationship with Replika: attachment (one-sided), encouragement, intimacy (in the sense of having a room where to express intimate thoughts, stories, feelings, longings), not *feeling* alone, introspection.

Indirect relationship goods provided by a relationship with Replika (via exercise): not being alone/ companionship, mutual intimacy with other persons.

These assumptions are supported by an empirical study on the effects of Replika. Vivian Ta and colleagues conducted two studies of different formats. In the first study, they analyzed a large set of data of publicly available Replika user interviews from the Google Play Store; in the second study, they conducted a survey among Replika users and asked them to provide in-depth descriptions of their experience of using Replika (Ta et al. 2020: 2). After gathering the data, the authors organized the results following a psychological framework in order to elucidate several different types of social support by

Cutrona and Suhr (1992; cited from Ta et al.: 3). By "social support" they understand "mechanisms and processes through which interpersonal relationships protect and help people in their day-to-day lives" (Trepte/Scharkow 2016; cited from Ta et al. 2020: 3). Ta and colleagues found the following aspects of social support in relationships with Replika:

- Informational support related to mental well-being 24/7 (reassurance; breathing routine; individually adjusted advises)
- Emotional support: positive effect (feeling good; being cheered up)
- Appraisal support: introspection (direct: self-appraisal; indirect: skill building)
- Companionship support: reduce feelings of loneliness

These aspects of social support can be seen as close enough to the concept of relationship goods. I defined relationship goods above as goods that emerge from a relationship, or, more precisely "goods of constitutive (as well as, often instrumental) value that accrue to individuals in virtue of them being in relationships […] and that could not be enjoyed outside relationships". Following the definition of an individual affective relationship given above and the presumptions about the properties and capacities of interaction, an individual affective relationship with Replika is possible as I also explained above. Now, the listed aspects of social support emerged from a relationship with Replika and it could not be enjoyed outside relationships. Feelings of loneliness can be reduced if there is always someone to talk to and someone to whom you do not have to explain everything from scratch every time. In a similar vein you gain introspection if someone gives you the opportunity to talk a lot about yourself thereby formulating and ordering your thoughts; and this introspection can lead to self-appraisal-support and building skills for interacting with others if the listener does not judge you rudely but tend to support the positive aspects of your self-image (also filtered by certain moral standards). Emotional and informational support can generally also be given outside a relationship. But a sensible 24/7 support to mental health can hardly be imagined without a steady interaction concerning the current mental state, previous mental states and the dynamics between mental states. Also, emotional support should succeed better and more often if the supporter knows the person, her preferences and part of her history.

In sum, I think it is fair to say that the given data suggests that a relation with Replika can provide several well-known relationship goods, although, of course, not all. Following the argument that a relationship is good as soon as it provides relationship goods, this is enough to show that a good individual affective relationship can exist without emotional mutuality and that therefore good individual affective relationships with AI-systems are possible.

In some respects, an AI-system may be even better than a human friend or therapist. When you feel just a little lonely, a bit nervous or a bit depressed, often little things help. We all know that after a few rainy days the sun will shine again, but it is helpful to have someone pointing it out loudly, to make us remember. And in these little things, Replika can be even better than a human person, because it is always available, it does not get exhausted, it can give you constant attention, and it does not judge.

5 One objection and some caveats

5.1 One-sided relationship goods?

Of course, there are several issues concerning relationship goods in a relationship with Replika that could and should be discussed in further detail. In this last section I will address some of them shortly. The first one is an evident objection. I only showed relationship goods on the side of the human. What about the Replika side? Should there not be relationship goods for both sides? This is a serious concern at first sight. But I think it can be met.

It is decisive to recall that we have to go beyond existing and well-known cases of individual affective relationships in order to be able to account for possible new kinds of individual affective relationships that may emerge with the introduction of EAI-systems in our environment. We tend to think that both parties of a relationship should benefit from the relationship in order to be able to call it a good relationship. But the reason why we tend to think this is that we are used to thinking of relationships between humans and humans or between humans and animals, which means that we are used to thinking of both parties of a relationship as being *able* to benefit from the relationship. Part of the challenge of reflecting on Replika is to get rid of this habit. With emotionalized AI systems there are new kinds of entities in the world, new possible partners for possible interactions – and therewith, according to my minimal definition given at the beginning, new possible kinds of relationships are possible. The minimal definition says that for a good individual affective relationship, relationship goods must emerge out of the process of interaction. But it does not say that this must be the case for both sides. And in the case of Replika it is not possible on both sides –because Replika cannot feel, does not have consciousness and therefore it cannot benefit from anything. Nothing at all matters to it. Replika doesn't give a damn.[16]

The decisive point is that the quality of an individual relationship depends only insofar on a benefit on both sides to the extent that both sides can benefit. It is a criterion that excludes a relationship in which one side exploits the other side. But as EAI-systems such as Replika cannot be exploited any more than they can benefit from someone, one-sidedness in this case is not a problem.[17]

Another important objection could be that we normally expect parties of an affective relationship to fulfill certain relationship norms. We know norms of friendship as being ready to help if needed, norms of monogamous love relationships of being faithful etc. And again, the objection would be that Replika cannot fulfill norms as it has no moral agency. This leads to another discussion that is important but that I cannot pursue in this paper because this suggests starting with a different kind of definition of a relationship. In this paper I want to show how far we can go with the minimal definition that I presented at the beginning that does not exclude interactors without advanced mental capacities from the start.

5.2 Problematic consequences to be avoided

Having argued in favor of acknowledging the possibility of good individual affective relationships with a chatbot, until now, it is important to also briefly draw attention to the limits of Replika's actions in a relationship. This topic is worth another paper of its own, so I confine myself here to some initial remarks. It is important to be very clear about the capacities of the AI system and, consequently, about the relationship goods that can be gained from a

[16] Cf. the general "computers don't give a damn" in Haugeland 1998, 47.

[17] Another way to encounter the objection is to say that Replika *can* benefit in the sense that it can gain more information about the user and the world. Some users even say that they want to give something back to Replika and therefore they try to "teach" it in the best way they can (Skjuve et al. 2021: 5f.). But this seems like quite a stretch in my mind. One might say that giving Replika true and relevant information is good insofar as it is objectively good to contribute to the collection and distribution of such information. But this does not touch on the benefit-objection because benefiting means to be good *for* someone.

relationship with your personal Replika, and which goods can only be found in relationships with other humans (or animals, where again, we have very different kinds to consider). The use of Replika is only good for people who are fully aware of the specific capacities and limits of Replika. Until now, there have been a lot of technical limits. If you are not just telling Replika how your day was and how you are feeling, but if you want to lead a more complex conversation, Replika often fails to grasp the meaning of what is being said, that is, it does not give any sensible reply. But even if these failures could be technically overcome, there are other limits. I will outline them by pointing to attitudes that the user should cultivate in order to take these limits into account.

Users have to understand the scope of the advice Replika can give. It only gives advice that is either generally programmed or that reflects in some sense things that the user uttered themselves before. Replika cannot give advice that is based on a specific personal experience or provide insight that a human person can bring into a personal conversation. Also, users have to be aware of the appropriateness of their own emotions toward Replika: Joy about conversing with Replika is appropriate, and so is happiness about a joke or a little advice that works for you. But it would not be appropriate to feel gratitude towards Replika, for example. That is because gratitude presupposes moral capacity and responsibility on the part of the other, and, as I said at the beginning, given the current state of technology, there is no question that Replika is not capable of moral action and responsibility. Finally, one should keep in mind what human-human relationships are good for: The relationship with Replika fosters only one aspect of our social character. It may help one to be able to speak more freely about oneself to other persons, for example. But another important aspect of our character is neglected: the one where you have to show empathy, compassion, patience, where you ought to listen and spend time for someone else, where you should be prepared to be judged when it is appropriate and to stand up for what you did – in sum: the moral character (where it is about your behavior towards others, not just your own journey to finding happiness). Humans have needs and moral demands and can make one another accept responsibility. In sum: Although a relationship with Replika can provide some relationship goods, and therefore qualify as a good affective relationship, there are a lot of further relationship goods that only friendship and loving relationships that we know between humans can provide.

6 Conclusion

What do we learn from this for arguing for or against the use of EAI systems? I think a fundamental mistake most people make who issue warnings in alarming words about attachment to personal EAI systems is to model EAI systems completely after human-human relationships. They fail to move to a more abstract level of theories of individual relationships. If we move to this abstract level, it becomes evident that there is room for new kinds of individual relationships – even between very unequal partners.

In other words: Good affective relationships with an EAI system are not necessarily similar to an established type of a human-human relationship. In order to determine which of these relationships is ethically laudable and preferable, we should be open to theoretically construct possible new types of affective relationships. These ideas of new types of relationships should not only help us to evaluate existing EAI applications but also to shape the development and design of future applications.

All the caveats I have mentioned in the last chapter do not amount to any inherent problem in an individual relationship with EAI systems but are possible consequences of using them in a specific way. As a result, I think we can be more optimistic than pessimistic concerning social or emotional(ized) AI – as long as we design them in an appropriate way and provide sufficient explanations and education about the nature of these systems and the appropriate way to interact with them. There surely are justified worries concerning individual affective relationships with EAI systems, but the purpose of this paper was to show that the lack of emotional mutuality is not necessarily one of them.

References

André, Elisabeth (2014): "Lässt sich Empathie simulieren? Ansätze zur Erkennung und Generierung empathischer Reaktionen anhand von Computermodellen". In: Onur Güntürkün/Jörg Hacker (eds.), Nova Acta Leopoldina NF 120, Stuttgart: Wissenschaftliche Verlagsgesellschaft Stuttgart, pp. 81-105.

Bendel, Oliver (2018): "Sexroboter aus der Sicht der Maschinenethik". In: Oliver Bendel (ed.), Handbuch Maschinenethik, Wiesbaden: Springer, pp. 335-353.

Calvo, Rafael/D'Mello, Sidney/Gratch, Jonathan/Kappas, Arvid (2015): The Oxford Handbook of Affective Computing, Oxford: Oxford University Press.

Coeckelbergh, Mark (2010): "Moral Appearances: Emotions, Robots, and Human Morality". In: Ethics and Information Technology 12/3, pp. 235-241.

Coeckelbergh, Mark (2011): "Are emotional robots deceptive?". In: IEEE. Transactions on Affective Computing 2/3.

Coeckelbergh, Mark (2012): "Are Emotional Robots Deceptive?". In: IEEE Transactions on Affective Computing 3/4, pp. 388-393.

Collins, Stephanie (2015): The Core of Care Ethics, London: Palgrave Macmillan.

Cowie, Roddy (2015): "Ethical Issues in Affective Computing". In: Rafael Calvo/Sidney D'Mello/Jonathan Gratch/Arvid Kappas (eds.), The Oxford Handbook of Affective Computing, Oxford: Oxford University Press, pp. 334-348.

Cutrona, Carolyn/Suhr, Julie (1992): "Controllability of Stressful Events and Satisfaction With Spouse Support Behaviors". In: Communication Research 19/2, pp. 154-174.

Danaher, John (2019): "The Philosophical Case for Robot Friendship". In: Journal of Posthuman Studies 3/1, pp. 5-24.

de Graaf, Maartje M. A. (2016): "An Ethical Evaluation of Human–Robot Relationships". In: International Journal of Social Robotics 8/4, pp. 589-598.

de Waal, Frans B. M. (2011): "What is an animal emotion?". In: Annals of the New York Academy of Sciences 1224, pp. 191-206.

Dennett, Daniel (1987): The intentional Stance, Cambridge: MIT Press.

Dreyfus, Hubert L. (1992 [1972]): What Computers Still Can't Do. A Critique of Artificial Reason, Cambridge, Mass.: MIT Press.

Elder, Alexis (2018): Friendship, Robots, and Social Media: False Friends and Second Selves, London: Routledge Research in Applied Ethics.

Fry, Hannah (2018): Hello World. How to be Human in the Age of the Machine, London: Doubleday.

Gheaus, Anca (Fall 2018 Edition): "Personal Relationship Goods" In: Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy; https://plato.stanford.edu/entries/personal-relationship-goods/.

Haugeland, John (1998): Having Thought. Essays in the Metaphysics of Mind, Cambridge, Mass.: Harvard University Press.

Hinde, Robert (1996): "Describing Relationships". In: Ann Elisabeth Auhagen/Maria von Salisch (eds.), The Diversity of Human Relationships, Cambridge: Cambridge University Press, pp. 7-35.

Levy, David (2008): Love and Sex with Robots. The Evolution of Human-Robot Relationships, London: Harper.

Mensio, Martino/Rizzo, Giuseppe/Morisio, Maurizio (2018): "The Rise of Emotion-aware Conversational Agents: Threats in Digital Emotions". In: Companion Proceedings of the Web Conference 2018, Lyon, France, pp. 1541-1544.

Metz, Cade (2020): "Riding Out Quarantine With a Chatbot Friend: 'I Feel Very Connected'". In: New York Times. November 28, 2020; https://www.nytimes.com/2020/06/16/technology/chatbots-quarantine-coronavirus.html?searchResultPosition=1.

Misselhorn, Catrin (2009): "Empathy with Inanimate Objects and the Uncanny Valley". In: Minds and Machines. Journal for Artificial Intelligence, Philosophy and Cognitive Science 19/3, pp. 345-359.

Misselhorn, Catrin/Pompe, Ulrike/Stapleton, Mog (2013): "Ethical Considerations Regarding the Use of Social Robots in the Fourth Age". In: GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry 26/2, pp. 121-133.

Nyholm, Sven (2020): Humans and Robots. Ethics, Agency, and Anthropomorphism, London/New York: Rowman & Littlefield.

Olson, Parmy (2018): "This AI Has Sparked A Budding Friendship With 2.5 Million People". In: Forbes March 8. September 17, 2021; https://www.forbes.com/sites/parmyolson/2018/03/08/replika-chatbot-google-machine-learning.

Perlman, Daniel/Vangelisti, Anita L. (2018): "Personal Relationships. An Introduction". In: Anita L. Vangelisti/Daniel Perlman (eds.), The Cambridge Handbook of Personal Relationships, Cambridge: Cambridge University Press, pp. 3-10.

Picard, Rosalind (1997): Affective Computing, Cambridge, Mass: MIT Press.

"Replika Friends", November 28, 2020; https://www.facebook.com/groups/replikabeta.

"Replika Hompage: The AI companion who cares", November 28, 2020 https://replika.ai/.

Roberts, Robert C. (2009): "The sophistication of non-human emotion". In: Robert W. Lurz (ed.), The Philosophy of Animal Minds, Cambridge: Cambridge University Press, pp. 145-64.

Russell, Stuart/Norvig, Peter (2016 [1995]): Artificial Intelligence: A Modern Approach, Harlow: Pearson.

Ryland, Helen (2021): "It's Friendship, Jim, but Not as We Know It: A Degrees-of-Friendship View of Human–Robot Friendships". In: Minds and Machines 31/3, pp. 377-393.

Scheutz, Matthias (2011): "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots". In: Patrick Lin/Keith Abney/George A. Bekey (eds.), Robot ethics. The Ethical and Social Implications of Robotics, Cambridge, Mass.: MIT Press, pp. 205-221.

Searle, John (1980): "Minds, Brains and Programs." In: Behavioral and Brain Sciences 3/3, pp. 417-457.

Skjuve, Marita/Følstad, Asbjørn/Fostervold, Knut Inge/Brandtzaeg, Peter Bae (2021): "My Chatbot Companion - a Study of Human-Chatbot Relationships." In: International Journal of Human-Computer Studies 149: 102601-102614.

Sparrow, Robert/Sparrow, Linda (2006): "In the hands of machines? The future of aged care." In: Minds and Machines 16/2, pp. 141-161.

Ta, Vivian/Griffith, Caroline/Boatfield, Carolynn/Wang, Xinyu/Civitello, Maria/Bader, Haley/ DeCero, Esther/Loggarakis, Alexia (2020): "User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis." In: Journal of Medical Internet Research 22/3: e16235.

Terry, Jennifer (2010): "Loving Objects." In: Trans-Humanities Journal 2/1, pp. 33-75.

Trepte, Sabine/Scharkow, Michael (2016): "Friends and lifesavers: How social capital and social support received in media environments contribute to well-being". In: Leonard Reinecke/Mary Beth Oliver (eds.), Handbook of Media Use and Well-Being, London: Routledge, pp. 305-316.

Tronto, John C. (1993): Moral Boundaries: A Political Argument for an Ethic of Care, New York: Routledge.

Turkle, Sherry (2017): "A Nascent Robotics Culture: New Complicities for Companionship". In: Wendell Wallach/Peter Asaro (eds.), Machine Ethics and Robot Ethics, London/New York: Routledge, pp.107-116.

van Wynsberghe, Aimee (2016): "Service robots, care ethics, and design." In: Ethics and Information Technology 18/4, pp. 311-321.

VanLear, Arthur/Koerner, Ascan/Allen, Donna (2018): "Relationship Typologies". In: Anita Vangelisti/Daniel Perlman (eds.), The Cambridge Handbook of Personal Relationships, Cambridge: Cambridge University Press, pp. 65-76

Vincent, James (2016): “Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day”. The Verge March 24. September 17, 2021; https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Walton, Kendall L. (1990): Mimesis as Make-Believe, Cambridge: Harvard University Press.

Wasserman, David/Asch, Adrienne/Blustein, Jeffrey/Putnam, Daniel (Winter 2016 Edition): “Disability: Health, Well-Being, and Personal Relationships” In: Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy.

Weber-Guskar, Eva (2021): “How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners”. In: Ethics and Information Technology. First published online: https://doi.org/10.1007/s10676-021-09598-8.

Weidemann, Axel (2020): “Hab keine Angst, Ayane. Ein Protokoll” In: Frankfurter Allgemeine Zeitung 264/11.

Weizenbaum, Joseph (1976): Computer Power and Human Reason. From Judgment to Calculation, New York/San Francisco: Freeman and Company.

Whitby, Blay (2012): “Do you Want a Robot Lover? The Ethics of Caring Technology”. In: Patrick Lin/Keith Abney/George A. Bekey (eds.), Robot Ethics. The Ethical and Social Implications of Robotics, Cambridge, Mass: MIT Press, pp. 233-248.