Empathy and Dyspathy with Androids: Philosophical, Fictional and (Neuro-)Psychological Perspectives

Catrin Misselhorn
University of Tübingen

The fact that we develop feelings towards androids, i.e., objects with a humanlike appearance, has fascinated people since ancient times. However, as a short survey of the topic in history, science fiction literature and film shows, our emotional reactions towards them are ambivalent. On the one hand, we can develop feelings of empathy almost as we do with real human beings; on the other hand, we feel repulsion or dyspathy when those creatures show a very high degree of human likeness. Recently, Japanese roboticist Masahiro Mori coined the term "uncanny valley" to refer to this effect. The aim of this essay is, first, to give an explanation as to why we feel empathy towards androids although we know that they do not have feelings themselves. This presupposes a perception-based concept of empathy which is going to be developed on the basis of some of Theodor Lipps' ideas. The second question to be answered is why empathy with androids turns into dyspathy when they become very humanlike. As I will argue, this is due to a particular kind of interference between perception and the imagination when confronted with very humanlike objects. This makes androids quite special objects right at the divide between humans and non-humans. They are non-human, but we feel ill at ease when treating them as mere objects.

1. Emotional responses towards androids

The human capacity to feel emotions towards inanimate objects that are humanlike – emotions which we normally only feel towards real humans—has fascinated people since ancient times. An example is provided by the myth of Pygmalion who falls in love with a female statue he created himself. However, the improvement of mechanics in the 18th century also opened up much more realistic possibilities to construct humanlike automata, backed up by philosophical mechanism. In 1738, Jacques de Vaucanson caused a sensation with the construction of a mechanical flute player, and other stunning automata followed. This development was also reflected in literary fiction which was often concerned with the emotional interaction of humans and automata. Well-known examples are E.T.A. Hoffmann's "Der Sandmann" (*The Sandman*) where a young man falls in love with a mechanical doll, or Mary Shelley's "Frankenstein" where a man-made creature starts to develop an emotional life of it's own.

In the 19th century the term "android" became prevalent to designate humanlike automata. The novel "L'Eve future" (*Future Eve*) by Auguste Villiers de l'Ilse-Adam which originally appeared in 1886 contributed to its popularization. In the plot, a female robot is constructed as a life-like substitute for one of the protagonist's facile fiancées with the more general aim of finding an alternative to real women, with their bodily deficiencies and weaknesses of character. In the following, I want to use the term *android* in a very broad sense comprising all kinds of objects which although not human, are humanlike and have the capacity to elicit emotions normally only shown towards our conspecifics. Empathy plays a particular role in this context, a topic which was the subject of several literary works and movies (often inspired by a novel or short story) of the last hundred years, e.g., Stanislav Lem's "Solaris" (filmed by Tarkowskij and Soderbergh), Ridley Scott's "Blade Runner" after a novel of Philip K. Dick, or Steven Spielberg's "A.I. – Artificial Intelligence" (inspired by Brian Aldiss' short story). An interesting point raised by a number of these works is that humans who are incapable of displaying empathy towards androids are themselves deficient. They suffer from an impaired emotional life, which amounts to a kind of moral failure.

The idea that a lack of empathy towards humanlike but inanimate objects is morally questionable is not restricted to fiction. Impressive examples are provided by Milgram-style experiments with avatars, i.e., virtual characters.[1] In the experiments the participants had to punish an avatar with electroshocks if she did not accomplish certain linguistic tasks. The results show that in spite of the fact that all participants knew for sure that neither the subject nor the shocks were real, the participants who saw and heard the avatar tended to respond to the situation at a subjective, behavioral and physiological level as if confronted with a real person.[2] Yet, as in the original experiment carried out by Milgram, this did not necessarily mean that people backed out of the experiment. Most people, however, did feel ill at ease when punishing the avatar, and if one watches the video of the experiment, one gets the impression that the suppression of this feeling would have bad effects on our capacity to develop morally relevant feelings towards real humans.

Yet, in the history of the topic there is also a second strand that highlights the pathological aspect of the emotional involvement of humans with androids. A well-known example from literature is Nathanel's love for the mechanical doll "Olimpia"—jan android *avant la letter*—in E.T.A. Hoffmann's short story "The Sandman." To the other characters in the novel this is a pathological response, and they—as well as the reader—get(s) an eerie feeling when confronted with Olimpia. I want to call this feeling of eeriness *dyspathy*. Dyspathy is more than just apathy—a lack of feeling; it is a distinctly negative, aversive feeling towards androids. Recently, in the film industry, dyspathy has become a problem with respect to the development of new technologies in the realm of animated movies. Very elaborate techniques like motion-capture or morphing have been developed to produce life-like movements and facial expressions in film characters with the aim of eliciting in the spectator the same empathy for non-human characters as for human characters. Despite these elaborate technologies, some recent films that were entirely created with their help have been accused of a severe failure in this respect, e.g., "Polar Express" (US/2004) by Robert Zemeckis. The spectators did not empathize with the small children who are the protagonists of the movie, but rather showed signs of dyspathy.[3] Paradoxically, dyspathy occurs particularly with characters that show a very high degree of humanlikeness. Therefore, the team of the movie "Shrek" decided to decrease anthropomorphism in the character of Princess Fiona, because "she was beginning to look too real, and the effect was getting distinctly unpleasant."[4] If we take this evidence seriously, an explanation of our emotional involvement with androids has to elucidate two things: *first* of all, why we feel empathy with androids although we know that they do not have feelings, and *secondly* why not all androids are capable of eliciting empathy, but some of them rather produce dyspathy.

The two kinds of the emotional responses of humans towards androids were brought together in a hypothesis by the Japanese roboticist Masahiro Mori.[5] He conjectured that humanlike objects elicit emotional responses similar to real humans proportionate to their degree of human resemblance: The more humanlike a robot or another object is made, the more empathetic emotional responses from human beings it will elicit. Yet, if a certain degree of similarity is reached, repulsion is suddenly the typical emotional response. Only when the object in question becomes almost indistinguishable from real humans do the responses become once again empathic. The relationship between humanlikeness and emotional response can also be illustrated graphically (see fig. 1). Mori called the emerging gap in the figure, where positive empathic responses suddenly turn into repulsion, the "uncanny valley."[6] The term "uncanny" refers to a famous notion introduced to psychology by Ernst Jentsch in "Über die Psychologie des Unheimlichen" (*On the Psychology of the Uncanny*) from 1906, which was taken up by Freud in his essay "Das Unheimliche" (*The Uncanny*) from 1919. In our context the term is primarily used to express that very humanlike objects do not just fail to elicit empathy, but produce a sensation of eeriness[7] which I called dyspathy. Supposedly, the effect is amplified if the object is able to perform movements by itself. Since androids are very humanlike robots, they often fall into the uncanny valley, as long as they are not perfect copies of humans.

Mori's hypothesis has a lot of intuitive plausibility, but one has to keep in mind that he did nt test it empirically. Although there has recently been increasing research activity[8] on this subject, the phenomenon is still contested. Nevertheless, I wish to propose my own explanation. A satisfying explanation has to consider three general approaches to the topic: a broadly *phenomenological* one that reflects on the phenomenon from the first-person perspective, an *empirical* one based on empirical research in psychology and the neurosciences, and a *philosophical* one which provides the conceptual tools to bridge the gap between the other two approaches.[9] Moreover, the philosophical approach should provide a more unified perspective for the diverse strands of empirical research by embedding them in a more view of how the mind works. The aim of the explanation along these three lines is to give us a better understanding of the two kinds of emotional responses towards androids—empathy and dyspathy. More precisely, I will shed light on the following questions: Why do we feel empathy towards androids at all? Why do we cease to feel empathy when they become exceedingly humanlike? And why do we not just stop feeling empathy, but start to respond with dyspathy? To answer these questions, we first have to develop an appropriate account of empathy. It has to explain why we feel empathy with androids at all, even if we *know* that they do not have feelings. This looks at first glance like a paradox. Is *empathy* not just defined as *feeling someone else's emotions*? One might be tempted by this paradox to think that empathy with androids relies on an illusion which makes us accept the android as a real human being. Yet this strongly contradicts phenomenology, so I will suggest a different explanation. Its core is a kind of imaginative perception which is involved in empathic responses to androids. Finally, the results will be put to use in order to explain the dyspathic effect of the "uncanny valley".

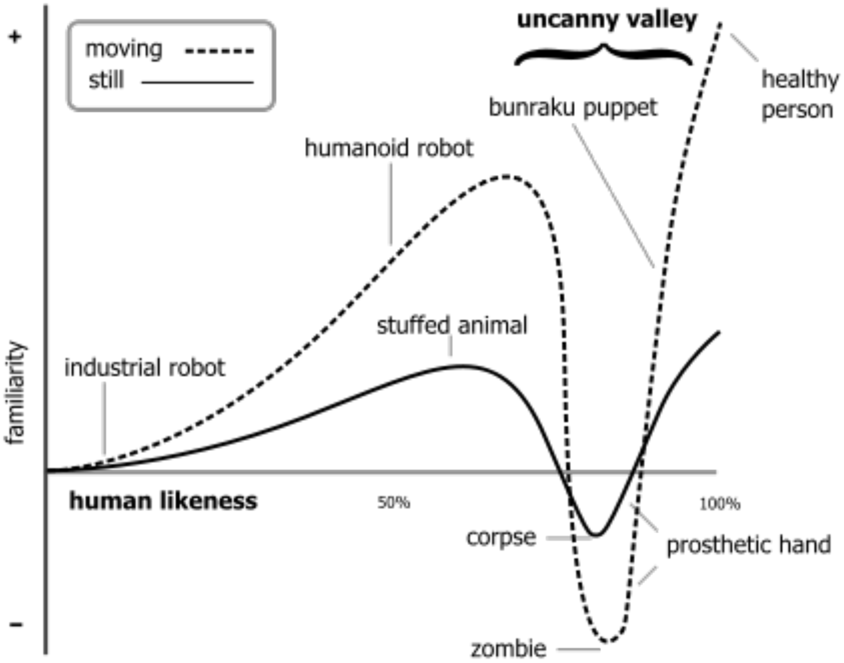

Fig. 1: Simplified version of Mori's original graph (MacDorman and Minato 2005)

II. Perception-based Empathy in the spirit of Theodor Lipps

In ordinary language, *empathy* is often characterized as the ability to "put oneself into another's shoes." This involves three dimensions which are not clearly distinguished: knowing what a person is feeling, feeling what another person is feeling, and responding compassionately to another person's distress. The third dimension is sometimes called sympathy since it involves a different feeling than the one of the target. However, in cases where sympathy depends on empathy, i.e., on literally feeling the other person's distress, the term empathy might as well refer to the whole emotional state.10 Let us therefore distinguish a narrow and a broad use of the concept. In the narrow sense empathy only includes sharing another person's emotional state; in the wide sense (which comes, from my point of view, closer to its ordinary meaning) it includes a compassionate response. I will be using the term in its broad sense including all dimensions. Since I assume that we know that androids do not really have feelings, the first dimension does not play a role in empathy with androids. Nevertheless, there is a way to make sense of the other two dimensions: even though we are well aware than the android does not have feelings, we can, in a way, come to feel what we perceive an android as feeling, and respond compassionately to it. Despite this, many approaches to empathy exclude empathy with non-sentient beings right from start. I will use as an example an influential formula created by Elliott Sober and David Wilson. They provide the following definition of empathy:

"S empathizes with O's experience of emotion E if and only if O feels E, S believes that O feels E, and this causes S to feel E for O."11

Obviously, this is not an eligible concept of empathy for our purposes. The demand that there be a belief that O feels E conceptually excludes the possibility of our feeling empathy with androids, since we do not believe that they really have feelings.
There are, however, other concepts of empathy which have a less intellectual and more perceptual basis. They were even prevalent at the beginning of the concept's historical evolution, which is rather short in comparison to the history of other philosophical concepts. Although the word has its roots in ancient Greek, the term was introduced into psychology and philosophy only at the turn of the 19th to the 20th century.12 The English term "empathy" goes back to the psychologist Edward Titchener (1867-1927) who used it in 1909 as a translation of the German term Einfühlung, which played a major role in philosophical aesthetics by the end of the 19th century.13 One of the most interesting thinkers dealing with empathy was Theodor Lipps (1851-1914). Like most people at the time, he studied empathy mainly as an ingredient in aesthetic experience. More precisely, he saw aesthetic experiences as a specific kind of empathetic perception of external states of affairs in analogy to the model of human empathy. The notion of an expressive movement (Ausdrucksbewegung) is of particular importance for his approach. This is a bodily movement or sound that serves the expression of a person's emotional state of mind.14 As Lipps states, the emotion and its

expressive movement are not just causally connected, but rather we can perceive the former somehow directly in the latter. How is it possible to see an emotion in a physical object, and why does this involve some kind of empathic feeling (miterleben)?

It is not enough that we ourselves show the same expressive behavior when we are undergoing an emotion of the same type as the emotion that we see the other person as having. The reason for this is that experiencing an emotion (including the corresponding expression) is different from *seeing* another person's expressive movements. In the latter case we are entertaining a visual representation, in the former we are feeling muscular tension and skin sensations, but we do not *see* our own expressive behavior, or feel the other person's. For that reason, our own emotional experiences as such cannot establish a connection between the feeling and the visual representation of the expressive behavior. Another explanatory approach discussed by Lipps—which would probably be adopted by many philosophers today—suggests that the link is established by empirical association: We have observed a connection between an emotion and a certain type of expressive behavior numerous times; therefore, we infer that the emotion is present if we see someone behaving in a specific way.[15] Although this kind of inference might be relevant in some cases (for instance, learning that wagging its tail is a sign of a dog's happiness), it cannot account for all kinds of perceptually based empathy. As Lipps was well aware, infants respond much earlier to their mother's faces' emotional expression than they can be credited with the capacity to draw inferences.

Therefore, Lipps concludes that the visual (or auditive) representation of expressive behavior has to trigger an unconscious simulation mechanism which he calls *innere Nachahmung*: a visual representation of expressive behavior produces a kinesthetic representation of the kind we experience when we are really executing that behavior. This idea is consistent with recent empirical research which shows that face-based emotion recognition, in particular, is dependent on some sort of simulation. It even occurs in neonates by mimicking facial expressions.[16] Further evidence is provided by studies finding deficits in face-based emotion recognition and emotion production in people with facial paralysis.[17]

At the neurological level, resonance phenomena studied in monkeys and humans were also brought together with Lipps' account of empathy.[18] So called mirror neurons are neurons that get activated when an individual either personally experiences mental events or when he or she observes a sign that another individual is experiencing or is about to experience the same type of event.[19] Yet, activation in the observer occurs mostly at a level below the threshold of consciousness. The trigger signs can consist of a behavioral manifestation, facial expression, or a stimulus that can be expected to produce the mental event in question.[20] For example, when one person is observing another, either wincing in pain, showing a face contorted with pain, or being stabbed with a knife, the observer's pain mirror neurons discharge. Therefore, the other personobserver feels the same emotions and may, as a consequence, react compassionately to the other person's distress. I do not want to discuss the neurological evidence in detail, but it seems to provide support of Lipps' view of "*innere Nachahmung*. Finally, I will define perception-based empathy based on Lipps' account:

(1) A person *S* empathizes with another person *O*'s experience of emotion *E* if *S* perceives *O* as undergoing *E*, and this perception causes *S* to feel *E* for *O*.

This definition avoids explicitly ascribing emotions to O, which distinguishes it from more cognitive definitions of empathy that involve imagining being in someone else's situation by simulating the person's thoughts and experiences. Yet, it is still not apt for our purposes, because we do not want to say that androids really undergo emotions. We do not literally *see* an inanimate object showing an emotion. We are even aware that androids belong to a category of inanimate objects that do not have emotions. Nevertheless, an inanimate object does not have to be very similar to a human being to illicit empathy. A rather low degree of human resemblance is sufficient, as in a rag doll, for example. Therefore, I am suggesting that the imagination must somehow be involved in the emergence of empathy with androids. This idea is captured by the following modification of the previous definition:

(2) A person *S* empathizes with an android *A*'s imagined experience of emotion *E* if *S* imaginatively perceives *A* as undergoing *E* and this imaginative perception causes *S* to feel *E* for *A*.

However, to fully understand what this definition amounts to, we have to obtain a better grasp of how imaginative perception works.

III. Imaginative Perception

The debate about pictorial representation is a helpful starting point. In this context the notion of imaginative perception has been most prominently discussed. One of the guiding intuitions of this discussion is that pictorial representation differs in specific ways from linguistic representation. Pictures seem to represent in a less conventional way than words; they somehow appear more similar to the depicted objects, and allow us to recognize them more easily without any special kind of semantic knowledge. To explain these facts, a variety of accounts of pictorial representation bring into play the concept of imaginative perception. The basic idea is that pictorial experience invokes a combination of perception and imagination: we simultaneously perceive the marks on the surface of a picture and imagine the depicted object.[21] Although this approach is aimed at pictorial representation, we can learn something from it about the interplay of perception and imagination in empathy with androids. This is not to say that we have to endorse it as the best account of pictorial representation. We are using the debate about pictorial representation only because the topic of imaginative perception has been studied most thoroughly in this context. It can help us to answer two questions: How does imaginative perception work? And how is it involved in empathy with androids?

The first thing to notice is that the imagining involved cannot just be propositional. Merely *supposing* that an android is having an emotion is not sufficient because there is no connection with perception, and perception seems to play a major role in this kind of empathy. Another straightforward suggestion can be discarded for similar reasons. Can we not simply say that we imagine the *android's* expressive behavior, facial expression or exposition to the relevant stimulus being a *human's* expressive behavior, facial expression or exposition to the relevant stimulus? Yet, this does not lead us far enough in the direction of perception, either. Let me explain why with the help of an example: we can imagine a banana being a telephone without imaginatively *perceiving* the banana as a telephone. We are just pretending that it is one by behaving in suitable ways, like putting one end to our ears and speaking into the other one. However, this is not the kind of imagination we are looking for. As we conceived the situation, it is the perception of the android that produces empathy in us, and not our pretending that it is human. This holds independently of how exactly one thinks of pretense.[22]

Yet, there is a more promising account of imaginative perception proposed by Kendall Walton.[23] He claims that someone who is looking at a picture, for instance, Meindert Hobbema's *Water Mill with the Great Red Roof* imagines that the object of his or her gaze is a mill. In other words: in seeing the canvas he or she imaginatively sees a mill. As I said before, I do not want to discuss whether this is an adequate account of pictorial representation. However, the idea of imagining that a perceptual experience of one thing is a perceptual experience of something else extends quite naturally to non-pictorial cases. Walton himself gives an example in a later article.[24] In this article he takes up a proposal by Patrick Maynard using Hitchcock's movie *Vertigo* in order to refute an objection. The protagonist of the movie, Scottie dresses up the girl Judy in order to enjoy a vivid imaginative experience of perceiving the now deceased woman he knew as Madeleine. In seeing the dressed-up Judy he imagines seeing Madeleine, and with the help of this imaginative perception he wants to produce the emotions he would have when seeing the real Madeleine. (Finally, Judy turns out to be Madeleine, but this does not matter here). Applied to our case, this is to say that in seeing the looks and the movements of the android and its reactions to an emotion-producing stimulus, we imagine perceiving a human expressive behavior, facial expression or reactions to the relevant stimulus.

But what is the relation between the two episodes of seeing something and thereby imagining that one is perceiving something else? One might suggest as a minimal condition that the two occur *simultaneously*. However, this is not a sufficient condition, since we can entertain an imaginary experience while at the same time having quite different perceptual experiences, e.g., when I imagine going out tonight while looking out of the window of my office. Yet, the way we are talking about imaginary perception as imagining that a perceptual experience of one thing is a perceptual experience of something else insinuates a stronger connection. It suggests that the two episodes are intentionally related: what I perceive has to be the object *of* my imagination. Moreover, as we have seen, imagining one experience to be another has to be something more experiential than supposing or pretending that one experience is the other.[25] One might be tempted by the latter desideratum to think that imaginary perception consists in an act of visualization. This seems to be a very natural suggestion with respect to imaginative perception in pictorial experience. According to this view seeing a picture amounts to a fusion of two things: the experience of seeing the surface of the picture and the experience of visualizing the scene depicted.[26]

However, at this point we have to leave the analogy with pictorial representation behind. The distinctive phenomenology of empathy with androids is not altogether identical with the one of pictorial experience. In particular, it does not involve visualization. We do not visualize a human undergoing an emotion when we imaginatively perceive an android as having an emotion. In order to come to terms with this problem we have to take a closer look at the structure of perceptual experience.

IV. Perceptual experience and imaginary perception

The debate about perceptual experience is vast, and I can hardly cover all the problems in this context. I will concentrate on the aspects which are relevant to the current topic and try to avoid unnecessary theoretical commitments. Perceptual experience has two general aspects: it has content, and it has phenomenal character (in the broadest sense of the term). Different kinds of perceptual experience can have the same content. I can, for instance, see someone walking down the street or hear someone walking down the street. Although the content of the perceptual experience is, in a certain sense, the same, both experiences do not have the same phenomenal character. The phenomenal aspect of a perceptual experience is what it is like for a subject to have it. Each sense modality has a particular phenomenal quality. There is also a phenomenal difference in perceptual experiences within every sensory modality, e.g., between seeing red and seeing blue.

With the help of these distinctions we can shed some more light on the functioning of imaginary perception. I am suggesting that imagining a perceptual experience of an (or the) object *F* as a perceptual experience of an (or the) object *G* involves the following: Because of certain salient similarities between the perceived object *F* and another object (or kind of objects) *G* the concept of a (or the) *G* is triggered. Yet, it is not applied to the perception, but entertained off-line, as it is commonly expressed in the language of contemporary theories of the imagination. This metaphor can be spelled out in the following way: a selected mental state or at least a rough facsimile of such a state is created by the faculty of the imagination. The real state and its imaginary counterpart are intrinsically very similar (they are wired into the same neural circuits, as, for instance, Goldman puts it),[27] but they differ with respect to their input and output conditions. Whereas the real state is triggered by the apprehension of facts, and normally leads to an action, this does not hold for its imaginary counterpart which is produced independently of the facts, and remains inert with respect to action.

I propose that a similar mechanism is also involved in empathy with androids. A concept is imaginatively or imaginarily triggered, but not applied to the situation; i.e., the subject does not have a disposition to *judge* that the present object is a human being. Despite its not being applied in the strict sense, the triggering of the concept does influence the perception. This might be described as a "blending" of perception and imagined concept.[28] Yet, the concept does not change the content of the perception – I am still seeing an *F* and not a *G* – but it does influence its phenomenal feel. Therefore, perceiving an (or the) *F* feels (to some extent) like perceiving an (or the) *G*. Of course, a concept can be more or less strongly triggered all the way up to full application. The more strongly it is triggered, the more vivid is the imagining, and the more the perception of *F* feels like a perception of *G*. How strongly a concept is triggered depends on several dimensions like the number of relevant features, as well as their typicality and salience. As a rule of thumb, the more features there are, and the more typical and salient they are, the stronger will the concept of a (or the) *G* be triggered, and the more vivid will the imagining and phenomenal feel be. This can be illustrated again with the help of Hitchcock's *Vertigo*. The perception of the dressed-up Judy triggers the concept of Madeleine in Scottie. As a result, perceiving Judy feels for him somehow like perceiving Madeleine. The more features both women share and the more typical and salient these are, the more the perception of Judy feels like a perception of Madeleine. This Madeleine-like perception of Judy then produces emotions (like hate or affection) the protagonist would have towards the real Madeleine, but not towards the real Judy.

Let us now apply this thought to the case of empathy with androids: Seeing the humanlike features *M* of an android triggers the concept of a human *N*. For that reason, seeing the android feels like seeing a human being, and its looks, movements and the exposition to a stimulus that can be expected to induce the emotion in question in humans feels (to some extent) like seeing a human being undergoing the corresponding emotion. Yet, as was argued above, the perception of these signs of an emotion in another can cause the same emotions in us, and lead to a compassionate response to the other's distress. Since these are the two dimensions of empathy we are interested in, we have by now arrived at the core of my explanation of how empathy with androids is possible. Moreover, we have a firmer grip on the intuition that a lack of empathy towards androids is an indicator of a more general deficit in the moral sensitivity of a person. If seeing an android in distress *feels* like seeing a human being in distress, a person who is too blunt to develop the morally relevant emotion of empathy in this case will also not respond with the appropriate emotion when confronted with real human suffering.

Although this explanation has some initial plausibility, one might, nevertheless, be troubled by some worries. *First*, this account seems to commit me to accept qualia, i.e. particular sensory qualities of experiences. Yet, I think my view is acceptable to anyone except eliminativists *tout court*.[29] The reason is that the account is compatible with the view that the qualitative character of a concept is determined by its intentional content, and I take it that this is a crucial question with respect to reductionism. That is to say, I am prepared to subscribe to a thesis which is called the *Phenomenology of Intentionality*.[30] It says that mental states of the sort commonly cited as paradigmatically intentional have phenomenal character that is inseparable from their intentional content. However, since we are just dealing with the phenomenal aspect of concepts here, I have to emphasize that I am only willing to accept the thesis of the Phenomenology of Intentionality with respect to them. It seems plausible that – if concepts do have a qualitative aspect at all – it has to be related to their intentional content.

This brings me to the *second* and more important objection. One might challenge my position since it deviates grossly from philosophical mainstream. I have to assume that concepts have phenomenal character, and this is not at all evident. A lot of people think that only perceptions or sensations can have a qualitative aspect. This objection depends on a certain wide-spread account of concepts in philosophy. In this view, concepts are analyzed as abstractions which are entirely detached from bodily aspects. A highly influential version of this approach was elaborated minutely by Jerry Fodor.[31] It assumes that concepts are abstract, amodal and arbitrary representations in a "language of thought" constituted of symbols that have subject-predicate structure and are manipulated by logical rules. Although this has been for a long time the predominant view, it has recently been challenged by evidence from clinical neuropsychology and cognitive neuroscience.[32] The evidence is supposed to show the following: distributed networks of discrete brain regions are active during object processing. The distribution of these categories varies as a function of semantic category, and the same regions are active, at least partly, when objects from a category are recognized, named, imagined, and when reading and answering questions about them. To accommodate this evidence, an account of concepts was proposed that differs profoundly from the standard view. It suggests that we see concepts as "embodied," i.e. (at least concrete perceptual) concepts are neural representations located in *sensory-motor areas* in the brain.[33] On this account concepts are not abstract, amodal and arbitrary, but involve the same neural activation pattern that is present in the perception, imagination and interaction with the relevant objects. If this is true, it seems quite natural to assume that these concepts can have a qualitative aspect. However, I cannot defend this claim here in detail. I just invoke it here to show that there are independent reasons to challenge the traditional paradigm of concepts. Finally, the reasons for seeing concepts as embodied coincide with the reasons to ascribe phenomenal character to them. However, if anyone who defends another theory of concepts is ready to grant that much, this would be sufficient to make my point.[34]

5. How does the uncanny valley emerge?

In the beginning I formulated three questions which I wanted to answer: Why do we feel empathy towards androids at all? Why do we stop feeling empathy when they become exceedingly humanlike? And why do we not just stop feeling empathy but start responding with dyspathy? So far we have mainly dealt with the first question. To provide an explanation of the uncanny valley we have to find satisfying answers to the remaining two. At the heart of my explanation lies the idea that in these cases the triggering of the concept gets so strong that it is about to turn into full-fledged concept application. Yet the attempt to apply the concept fails since the object of the perception is not accepted as an instance of the concept. For this reason the process leading to empathy is brusquely interrupted. However, because of the humanlikeness of the android the concept is triggered again and is repeatedly about to be elicited. This leads to a kind of very fast oscillation between four situations: the mere triggering of the concept, the reaching of the threshold of concept application, the failure of concept application resulting in a complete turning off of the concept, and the renewed triggering in continuing on to perceive the object.

This suggestion finds some empirical support by the study of vision.[35] There is research concerning the perception of ambiguous or fragmented figures which shows that, as the visual impression changes, so does the scan-path of the eye-movements (i.e. the sequence of fixations).[36] This suggests that the visual information leads to the generation of a hypothesis in the brain which, in turn, directs the eye-movements. In our case a constant alternation between two hypotheses of the kind "*a* is a human being" (concept triggering) and "*a* is not a human being" (concept deactivation) would take place. One would expect that this is correlated with incoherent eye-movements, and it can be anticipated that this will be accompanied by a feeling of confusion on the side of the subject. But does it explain the feeling of eeriness towards the entities falling into the uncanny valley? If it did, then all kinds of oscillations between different hypotheses should arouse that feeling.[37]

The proposed mechanism, therefore, seems to fall prey to a similar critique as Freud advanced against Ernst Jentsch's explanation of the uncanny.[38] Jentsch characterized it by doubts whether a lifeless object might be, as a matter of fact, animate, or whether an apparently animate being really is alive [39] My explanation of dyspathy with androids seems to come close to this intuition formulated in terms of the oscillation between the activation and deactivation of the concept of a human, i.e. a fortiori living being. The major difference between Jentsch's account and mine is that, from my point of view, these doubts take place at a sub-conscious level. They do not necessarily invoke a disposition to make the corresponding judgments. Freud criticized Jentsch's account as being too intellectual. Does not the same hold for my explanation, too? Is it not just a mere conceptual confusion?

This objection does not take into account that there is something phenomenally special about the concepts invoked in dyspathy with androids. Perceiving a human being has a very distinctive phenomenology, a kind of feeling that something is alive or endowed with a soul in the Aristotelian sense of the term.[40] In triggering the human concept this feeling gets activated and transferred to the perception of the inanimate object. However, if the threshold of concept application is reached, but it fails, this feeling is all of a sudden cut off. This is comparable to a sudden disenchantment: something that seemed to be alive and soulful a moment ago now appears cold and dead.[41] The alternation between these two states amounts to the feeling of eeriness. We are therefore now able to state the full-fledged explanation of the uncanny valley. The feeling of dyspathy with androids can be traced back to two elements: The *first* one is the oscillation between concept application prompted by imaginary perception and the failure of concept application since the object does not qualify as an instance of the concept. However, the first element just amounts to a feeling of conceptual confusion which also arises in other cases with a similar structure. The second element explains the specific character of eeriness which is characteristic of this particular kind of confusion. The triggering of the concept of a human being by an object involves seeing it in some sense as alive or endowed with a soul. The oscillation of concept activation and deactivation, therefore, also leads to an oscillation of the corresponding phenomenal feeling of seeing something as alive and seeing something as dead. Both elements taken together can account for the eerie feeling of dyspathy with androids.

I now want to close the circle by coming back to the use of androids in fiction discussed in the opening sections of the paper. As we saw, literary texts discovered the phenomenon of dyspathy with androids long before Mori formulated his hypothesis of the uncanny valley. On the other hand, it poses a practical threat to the use of androids in fiction, particularly movies, in so far as the recipient is called upon to empathize with them. It therefore looks as if we ought to supplement the three laws of robotics which guided Isaac Asimov in writing his short stories by a fourth one. Asimov's laws are:[42] (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. The suggested supplement would be something like: (4) A robot must not be too similar to a human being if it is supposed to elicit empathy.

[1] See Mel Slater et al. 2005. There is also a video of the experiment available online.

[2] Milgram-style experiments were also conducted with some more basic kinds of robots. In these cases people felt empathy, but not up to the level of real human beings (Bartneck et al. 2005).

[3] See, for instance, Paul Clinton's review of the movie for CNN (Clinton 2004).

[4] This is a statement of Lucia Modesto from PDI/Dreamworks quoted in Wischler (2002).

[5] See Mori 1970 and some more recent remarks in Mori 2005.

[6] There is, however, a discussion as to the shape of the function, see, for instance, Bartneck et al. 2007.

[7] For a study on the emotional valence of the expressions "eerie" or "creepy" in contrast to "strange", see Ho et al. 2008.

[8] The debate started, more or less, with the IEEE-RAS Humanoids Workshop: *Views on the Uncanny Valley*, 2005. A pioneer with respect to the empirical exploration of the valley was Karl F. MacDorman (see, for instance, MacDorman 2005a, 2005b, 2006, MacDorman et al. 2005, MacDorman/Ishiguro 2006). From the more practical side the work of Hiroshi Ishiguro (University of Osaka) had a great impact on the discussion. He constructed several stunning humanoid robots made out of silicone, among them a double of himself and one of his 5-year old daughter (film sequences are available on youtube, see for instance, URL: http://www.youtube.com/watch?v=JdHJ54fZ_Bw and http://www.youtube.com/watch?v=S9IrYL5IpG4&feature=related).

[9] However, "bridging the gap" does not necessarily amount to a reduction of one to the other.

[10] Empathy is not a necessary condition for sympathy, but both emotions have often been tied together in ethical and aesthetic contexts, e.g., in Lessing's theory of sympathy (*Mitleid*) in tragedy as developed in the *Hamburgische Dramaturgie*.

[11] Sober and Wilson 1998, 234f.

[12] Nevertheless the concept is often traced back to romantic thinking (see Stueber, 2006, Chap. 1). However, the romantics used the term 'empathy' (*Einfühlung*) in a quite specific sense to designate a sort of panpsychic symbiotic feeling of identity with nature (see Perpeet 1966).

[13] It was Robert Vischer who originally coined the term as a noun and technical concept in his work "On the Optical Sense of Form: A Contribution to Aesthetics" (1873). Yet, there are earlier, less technical uses, for instance, in the aesthetics of Friedrich Theodor Vischer (Robert Vischer's father), and Hermann Lotze's writings (see Stüber 2006, Chap. 1).

[14] See Lipps 1914, Chap. 2.

[15] An inferential view was proposed, for instance, by Ickes 1993.

[16] See, for instance, Meltzoff and Moore 1983.

[17] It is, however, an intricate matter, since the evidence is not indisputable (see Goldman/Sripada 2005 who discuss the topic from a simulationist point of view).

[18] See, for instance, Gallese 2001, Stüber 2006, Chap. 4.

[19] See, for instance, Rizzolatti/Craighero 2004.

[20] See Goldman 2008.

[21] See, for instance, Walton 1990, 2002; Wollheim 1998; O'Shaughnessy 2002.

[22] For different accounts, see Leslie 1987, Currie 1995, Nichols and Stich 2000.

[23] See Walton 1990.

[24] See Walton 2002.

[25] See Wollheim 1998.

[26] See Budd 1992.

[27] See Goldman 2006, 48.

[28] This might involve something like a blending mechanism as described by Fauconnier and Turner (see Fauconnier and Turner 1998; 2002).

[29] Eliminativism amounts to the claim that there are no mental states as we conceive of them in folk psychology, just brain states. This includes the non-existence of qualia. As a consequence, we should stop talking the way we commonly do about mental states. Reductionism, in contrast, claims that the mental states we assume in folk psychology do exist, but are reducible to neuronal events. With respect to qualia reductionists typically believe that the qualitative aspect of perceptual states is reducible to their representational content, and representational content can be explained naturalistically.

Therefore, the aim of reductionism is not to substitute common-sense discourse about mental states, but to explain it.

[30] See Horgan and Tienson 2002.

[31] See, for instance, Fodor 1998.

[32] See, for instance, Damasio 1989, Martin/Chao 2001, Caramazza/Martin 2003, Caramazza/Rumiati 2005, Pecher/Zwaan 2005; and, from a more philosophical point of view, Prinz 2005.

[33] See, for instance, Gallese/Lakoff 2005, who even want to expand this view to all kinds of concepts.

[34] For a defense of the phenomenal aspects of concepts following Husserl see Soldati 2005.

[35] See Stark et al. 2001.

[36] It seems to be promising to pursue in this context the approach to study eye-movement in human-android interaction (see MacDorman et al. 2005, Minato et al. 2005).

[37] This position is held by Ramey 2005

[38] See Freud 1919.

[39] See Jentsch 1906.

[40] There are some interesting parallels between the perception of animacy and causality (Scholl and Tremoulet 2000).The particular problems autistic children have with the perception of animacy are also very instructive in this context (Rutherford et al. 2006).

[41] This gets us to the true core of the widespread intuition that the uncanny valley has something to do with the fear of death (this was empirically studied within the framework of terror management research by MacDorman 2005b).

[42] See Asimov 1990, 8.

Works Cited

Asimov, I. (1990). Robot Visions, London.

Bartneck, C. et al. (2005). Robot Abuse – a Limitation of the Media Equation. Proceedings of the Interact 2005 Workshop on Agent Abuse, Rome.

Bartneck, C. et al. (2007). Is the Uncanny Valley an Uncanny Cliff? Proceedings of the 16th IEEE, RO-MAN 2007, Jeju, Korea, pp. 368-373.

Budd, M. (1992). On the Foundations of Representational Arts. Mind 101, 195-198.

Caramazza, A.

--- / Martin, A. (eds.) (2003). The Organisation of Conceptual Knowledge in the Brain. Cognitive Neuropsychology 20 (special issue).

--- / Rumiati, R.I. (Eds.) (2005). The Multiple Functions of Sensory-motor Representation. Cognitive Psychology 22 (special issue).

Clinton, P. (2004): 'Polar Express.' A Creepy Ride. [Online]. Available from http://www.cnn.com/2004/SHOWBIZ/Movies/11/10/review.polar.express/index.html (accessed 12 October 2008)

Currie, G. (1995). Imagination and simulation: Aesthetics meets cognitive science. Mental simulation: Evaluations and applications, ed. by A. Stone and M. Davies, Oxford, 151-69.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. Cognition 33, 25-62.

Fauconnier, G./Turner, M.

(1998). Conceptual Integration Networks. Cognitive Science 22, 133-187.

(2002). Conceptual Blending and the Mind's Hidden Complexities. New York: Basic Books.

Fodor, J.A. (1998). Concepts: Where cognitive science when wrong. Oxford.

Freud, S. (1919). Das Unheimliche. Gesammelte Werke Bd. 12, 6th ed., Frankfurt/M.1986, 229-68.

Gallese, V. (2001). The 'Shared Manifold' Hypothesis: From Mirror Neurons to Empathy. Journal of Consciousness Studies 8, 33-50.

Gallese, V./Lakoff, G. (2005). The Brain's Concepts: The Role of the Sensory-Motor System in Reason and Language. Cognitive Neuropsychology 22, 455-479.

Goldman, A.I.

(2006). Imagination and Simulation in Audience Responses to Fiction. The Architecture of the Imagination. New Essays on Pretence, Possibility, and Fiction ed. by Shawn Nichols, 41-56.

(2008). Mirroring, Mindreading, and Simulation. To appear in: Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition, ed. by J. Pineda, Totowa/NJ.

Goldman, A.I./Sripada, C.S.S. (2005). Simulationist Models of Face-Based Emotion Recognition. Cognition 94, 193-213.

Ho, C.-C./MacDorman, K. F./Pramono, Z. A. D. (2008). Human emotion and the uncanny valley: A GLM, MDS, and ISOMAP analysis of robot video ratings. Proceedings of the Third ACM/IEEE International Conference on Human-Robot Interaction, Amsterdam.

Horgan, T./Tienson, J. (2002). The Intentionality of Phenomenology and the Phenomenology of Intentionality. Philosophy of Mind: Classical and Contemporary Readings, ed. by D. Chalmers, Oxford, 520-33.

Ickes, W. (1993). Empathic Accuracy. Journal of Personality 61, 587-610.

Jentsch, E. (1906). Zur Psychologie des Unheimlichen. Psychiatrisch-Neurologische Wochenschrift 22, 203-205.

Leslie, A. M.

(1987). Pretense and representation: the origins of 'theory of mind'. Psychological Review 94, 412–426.

--- / Friedman, O. (2007). The conceptual underpinnings of pretense: Pretending is not 'behaving-as-if', Cognition 105, 103-124.

Lessing, G.E. (1767/1973). Hamburgische Dramaturgie. Werke Vol. IV: Dramaturgische Schriften, München.

Lipps, Th. (1914/1920). Ästhetik, 2 Vol., Leipzig and Hamburg, 2nd ed.

MacDorman, K. F.

--- (2005a). Androids as Experimental Apparatus: Why is there an Uncanny Valley and Can we Exploit it? Paper presented at the CogSci-2005 Workshop: Toward Social Mechanisms of Android Science, Stresa, Italy. (Includes the translation of Mori's paper from 2005 by MacDorman/Minato).

--- (2005b). Mortality Salience and the Uncanny Valley, International Conference on Humanoid Robots, Tsukuba, Japan.

--- (2006). Subjective Ratings of Robot Video Clips for Human Likeness, Familiarity, and Eeriness: An Exploration of the Uncanny Valley. Paper presented at the ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science, Vancouver, Canada.

--- /et al. (2005). Assessing Human Likeness by Eye Contact in an Android Testbed. Proceedings of the XXVII Annual Meeting of the Cognitive Science Society, Stresa, Italy.

--- /Ishiguro, H. (2006). The Uncanny Advantage of Using Androids in Social and Cognitive Science Resarch. Interaction Studies 7(3), 297-337.

Martin, A./Chao, L.L. (2001). Semantic Memory and the Brain: Structure and Processes. Current Opinion in Neurobiology 11, 194-201.

Meltzoff, A.N./Moore, M.K. (1983). Newborn infants imitate adult facial gestures. Child Development 54, 702–709.

Minato, T. et al. (2005). Does Gaze Reveal the Human Likeness of an Android? Development and Learning, [Online] Available: doi:10.1109/DEVLRN.2005.1490953 (accessed 15 October 2008).

Mori, M.

--- (1970). ‚Bukimi no tani', Energy 7(4), 33-35, translated into English by K.F. MacDorman and T. Minato (2005). Proceedings of the Humanoids-2005 workshop: Views of the Uncanny Valley, Tsukuba, Japan. . Available: http://www.androidscience.com/theuncannyvalley/proceedings2005/uncannyvalley.html [22 Mar 2009]

--- (2005). On the Uncanny Valley. Proceedings of the Humanoids-2005 workshop: Views of the Uncanny Valley, Tsukuba, Japan.

Nichols, S./Stich, S. (2000). A Cognitive Theory of Pretense, Cognition 74, 115-47.

O'Shaughnessy, B. (2002). Consciousness and the World, Oxford: Oxford UP.

Pecher, D./Zwaan, R. (Eds.) (2005). Grounding Cognition. The Role of Perception and Action in Memory, Language, and Thinking, Cambridge.

Perpeet, W. (1966). Historisches und Systematisches zur Einfühlungsästhetik. Zeitschrift für Ästhetik und allgemeine Kunstwissenschaft 11, 193-216.´

Prinz, J.J. (2005). The Return of Concept Empiricism. Handbook of Categorization in Cognitive Science, ed. by H. Cohen/C. Lefebvre, Oxford, 679-99.

Ramey, C. (2005). The Uncanny Valley of Similarities Concerning Abortion, Baldness, Heaps of Sand, and Humanlike Robots. Proceedings of the IEEE-RAS

International Conference on Humanoid Robots: Views of the Uncanny Valley, Tsukuba, Japan.

Rizzolatti, G./Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience 27, 169-192.

Rutherford, M. et al. (2006). The Perception of Animacy in Young Children with Autism. Journal of Autism and Development Disorders 36, 983-92.

Slater, M. et al. (2006). A Virtual Reprise of the Stanley Milgram Obedience Experiments. PLoS ONE 1(1): e39, doi:10.1371/journal.pone.0000039.

Scholl, B./Tremoulet, P. (2000). Perceptual Causality and Animacy. Trends in Cognitive Science 4, 299-309.

Sober, E./Wilson, D.S. (1998). Unto Others, Cambridge: Harvard UP.

Soldati, G. (2005). Begriffliche Qualia. Zur Phänomenologie Der Bedeutung.: Anatomie Der Subjektivität. Bewusstsein, Selbstbewusstsein und Selbstgefühl, ed. by Th. Grundmann et al., Frankfurt/M., 140-68.

Stark, W. et al. (2001). Representation of Human Vision in the Brain: How Does Human Perception Recognize Images. Journal of Electronic Imaging 10, 123-151.

Stueber, K. (2006). Rediscovering Empathy, Cambridge/Mass.

0Depiction, Perception, and Imagination: Responses to Richard Wollheim. Journal of Aesthetics and Art Criticism 60, 27-35.

Wischler, L. (2002). Why is this Man Smiling? Digital Animators are Closing in on the Complex System that Makes Faces come Alive. Wired 10.06. [Online]. Available from: http://www.wired.com/wired/archive/10.06/face_pr.html (accessed 12/10/2008).

Wollheim, R. (1998). On Pictorial Representation. The Journal of Aesthetics and Art Criticism 56, 217-233.