Minds & Machines (2009) 19:345–359
DOI 10.1007/s11023-009-9158-2

Empathy with Inanimate Objects and the Uncanny Valley

Catrin Misselhorn

Received: 5 August 2008 / Accepted: 9 June 2009 / Published online: 8 July 2009

Abstract The term "uncanny valley" goes back to an article of the Japanese roboticist Masahiro Mori (Mori 1970, 2005). He put forward the hypothesis that humanlike objects like certain kinds of robots elicit emotional responses similar to real humans proportionate to their degree of human likeness. Yet, if a certain degree of similarity is reached emotional responses become all of a sudden very repulsive. The corresponding recess in the supposed function is called the uncanny valley. The present paper wants to propose a philosophical explanation why we feel empathy with inanimate objects in the first place, and why the uncanny valley occurs when these objects become very humanlike. The core of this explanation—which is informed by the recently developing empirical research on the matter—will be a form of empathy involving a kind of imaginative perception. However, as will be shown, imaginative perception fails in cases of very humanlike objects.

Keywords Empathy · Androids · Uncanny valley · Imaginative perception

If we see somebody beating a child, we feel different than if we see somebody beating a computer: in the case of beating a child we empathize and feel compassion, whereas if we see somebody beating a computer, we do not. However, if an inanimate object looks in certain ways like a human being, we feel ill at ease when someone is beating it. Impressive examples for this phenomenon are provided by the Milgram-style experiments with avatars (Slater et al. 2006). The results of these experiments 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

C. Misselhorn (✉)
Department of Philosophy, University of Tübingen, Bursagasse 1, 72070 Tübingen, Germany
e-mail: catrin.misselhorn@uni-tuebingen.de

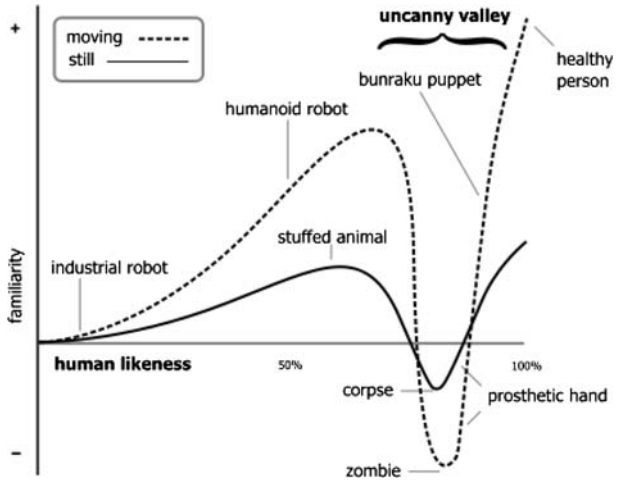

Fig. 1 Simplified version of Mori's original graph (MacDorman MacDorman 2005a)

the avatar tended to respond to the situation at the subjective, behavioral and physiological levels as if confronted with a real person. 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 2005).

One would suppose that this empathetic reaction corresponds to a linear function: the more a thing looks and behaves like a human being the more we will empathize. Yet, there exists evidence that it is not evolving as constantly as one might expect. Masahiro Mori, a Japanese roboticist, formulated a seminal hypothesis concerning the emotional responses of humans to robots and other non-human entities (Mori 1970). He stated that the more human-like a robot or another object is made, the more positive and empathetic emotional responses from human beings it will elicit. However, when a certain degree of likeness is reached, this function is interrupted brusquely, and responses, all of a sudden, become very repulsive. The function only begins to rise again when the object in question becomes almost indistinguishable from real humans. By then, the responses of the subjects approach empathy to real human beings. The emerging gap in the graph (Fig. 1) is called the "uncanny valley." The term "uncanny" is used to express that the relevant objects do not just fail to elicit empathy, they even produce a sensation of eeriness.[1] This effect amplifies if movement is added to the picture.

As a matter of fact, the uncanny valley has not just been discussed in robotics, but also with respect to motion pictures. This holds in particular for animated movies where very elaborate techniques are used to produce life-like movements and facial expressions with the aim of eliciting as much empathy as with human characters in the spectator. Despite these elaborate technologies, some recent films which were entirely created with their help were accused of falling prey to the uncanny valley,

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

e.g., "Polar Express" (US/2004) from Robert Zemeckis. In the case of the movie "Shrek" the team 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.".[2]

Nevertheless, Mori did not test his hypothesis empirically, and empirical research is just in the early stages of development.[3] This has probably to do with the fact that Mori's article originally appeared in Japanese, and was only translated into English in 2005. Philosophically, there has been so far no serious attempt to explain the phenomenon. I want to make a foray into this lacuna with the intention of showing that philosophical concepts and approaches allow for a clearer grasp of the problem, and provide the tools for a framework for a solution which can integrate the results of the empirical studies. However, this framework is not situated entirely over and above empirical research, but it allows for a better formulation of empirically testable hypotheses. In the first part of the paper, I will review some major strands of empirical research, and show that they cannot provide a satisfying answer to all dimensions of the problem, although they do touch important points. The second part is devoted to developing the philosophical framework. *First*, it has to be elaborated an account of empathy apt for inanimate objects. In the *second* step the particular interplay of perception and imagination in empathy with inanimate objects shall be elucidated. *Finally*, the results will be brought to bear on the explanation of the "uncanny valley" with the aim of integrating the empirical findings.

The Main Strands of Empirical Research

Let me start the survey of the empirical research by first formulating three questions a satisfying account of the "uncanny valley" has to answer:

(1) Why do we feel empathy towards non-human entities with some human-like characteristics at all?
(2) Why do we stop to feel empathy when the entities become very human-like?
(3) Why do we not just stop to feel empathy, but start to respond with eeriness?

Most of the research concentrated on the questions (2) and (3), and more specifically on what kind of factors have an influence on the arousal of the feeling of eeriness. Karl MacDorman, who also translated Mori's article, is one of the pioneers in the field (see MacDorman 2005a, b, 2006; MacDorman et al. 2005; MacDorman and Ishiguro 2006). His most outstanding idea is that uncanny humanoid robots elicit an innate fear of death and produce culturally supported defenses for coping with death's inevitability, i.e., they function as a reminder of mortality (MacDorman 2005b). He tried to test this hypothesis borrowing some methods of terror management research. Terror management theory was inspired by the cultural

[2] Lucia Modesto from PDI/Dreamworks quoted in Wischler (2002).

[3] Landmarks were the *Humanoid Robots and Humanoids* workshop 2005 in Tsukuba/Japan organized by the *Institute of Electrical and Electronics Engineers* (IEEE-RAS) and the *Toward Social Mechanisms in Android Sciences* workshop by the *Cognitive Science Society* 2006 in Vancouver/Canada.

anthropologist Ernest Becker who pointed out that psychological enquiry will not succeed without reference to belief systems. In his book *The Denial of Death* (Becker 1973) he argued that they play a particular role in coping with mortality. More recently, psychologists have tried to experimentally investigate how human beings manage their fear of personal extinction. These so-called terror management studies have mainly correlated subliminal reminders of mortality with a wide range of attitude changes. The basic idea is that having a worldview guard's people from anxiety about the inevitability of death (by giving a literal or symbolic explanation of how death is transcended).

For this reason, the so-called mortality salience hypothesis predicts that people who have been subliminally reminded of death will react more favorably to information that supports their worldview and less favorably to information that undermines it. As support for this hypothesis was counted, for instance, that people confronted with subliminal reminders of death more strongly prefer statements that praise their country to those that criticize it, that they prefer charismatic political candidates to those that are relationship oriented, and tend to judge moral transgressors more harshly. As MacDorman claims to have shown, people confronted with uncanny android robots show this kind of terror management, as well: "On average the group exposed to an image of an uncanny robot consistently preferred information sources that supported their worldview relative to the control group." (MacDorman 2005b, p. 404) He concludes that the feeling of eeriness we have when confronted with android robots is due to the fact that they remind us of death.

Although I am not entirely convinced by the methods of terror management research, I don't want to go into detail here. I want to focus my critique on a gap that opens up between MacDorman's claims and his experimental results. Maybe the picture of an android robot he presented to the subjects of the experiment really reminded people (unconsciously) of death. But this is not a good explanation for the *feeling* of eeriness.[4] Being a reminder of death might, at most, be a necessary condition for something to be perceived as eerie. But it is certainly not a sufficient condition. There are a lot of still lives in the history of art that remind the informed spectator of death, but don't cause eeriness.

So maybe the relevant object must not just be a reminder of death, but appear as a dead body. Perhaps android robots share some visual anomalies with corpses, and, therefore, elicit the same alarm and repulsion. But the sight of a corpse does not always produce feelings of eeriness. Mori himself pointed out in a more recent statement that sometimes, "the face of a dead person gives us a more comfortable expression than the one given by a living person's face, dead persons are free from the troubles of life, […] and this is the reason why their faces look so calm and peaceful." (Mori 2005) Moreover, the imagination and depiction of—particularly female—corpses, like Ophelia, has been a subject that fascinated artist and their audience. Edgar Allan Poe (Poe 1846) even said that the death of a beautiful woman is the most poetical topic of all. This seems to suggest that not death, but certain

[4] In a later article (MacDorman and Ishiguro 2006) MacDorman himself expresses some doubts on the this matter.

aesthetic properties may be relevant to the problem of the uncanny valley. Accordingly, skeptics about Mori's hypothesis tried to show that the feeling of eeriness is not so much correlated with human likeness, but with physical attractiveness or beauty (Hanson 2006). Hanson points out that, universally, clear skin, well-groomed hair and large expressive features are considered attractive. Other characteristics are universally regarded as ugly or disturbing, for instance, sickly eyes, bad skin, extreme asymmetry, and poor grooming. Those features are considered to be signs of illness, neurological conditions or mental dysfunction, and, therefore, supposed to lead to repulsive reactions. This is backed up by an evolutionary explanation. It is very useful to have a capacity to detect signs of genetic disorders or lack of genetic fitness which prevents one from trying to reproduce with those members of the species. This is according to this approach the task of the feeling of eeriness towards deviant organisms. If something appears very humanlike—as the objects in the uncanny valley—it is seen as part of the human species, and emotionally evaluated by the same standards, although it is known that it is not a human being.

Human aesthetic preferences also transfer to non-human objects and beings. The explanation why we react with aversion against the android robots along these lines is that they do not match our aesthetic standards, respectively, that they show anomalies which make us react as we would to deviant human individuals showing the "ugly" signs of poor health or bad genes.[5] Nevertheless, Hanson admits that "more realistic faces trigger more demanding expectations for anthropomorphic depictions" (Hanson 2006), i.e., realism has an impact on the phenomenon, but, from Hanson's point of view, the aesthetic features carry the main burden. He tried to test this hypothesis with the help of morphing. This is a technique used in motion pictures and animations which changes one image into another through a seamless transition. The morphing started with a picture of an abstract robot, and went on via several transitory stages to an image of a humanoid robot and finally to the human model on which the robot was based.

The crucial thing is that *two* morphings were carried out. Whereas the first one just fused the images in question, the second one was "tuned" to make the transitory stages of the series appear more attractive. This effect was achieved by introducing the image of a doll which looked like Barbie's partner "Ken" the idea being obviously that this is a stereotype of male human attractiveness. Hanson presented the subjects with a series of eleven stills from each morphing, and made them evaluate the pictures along the dimensions of human likeness, familiarity, eeriness, and degree of appeal. With respect to the first, unturned morphing Hanson reproduced a result by MacDorman in a similar experiment which confirms the hypothesis of the uncanny valley. However, with respect to the second morphing which was designed to be more appealing, this was not the case. The subjects made rather consistent evaluations of low eeriness and high appeal. Hanson, therefore, comes to the conclusion: "If the illusion of life can be created and maintained, the

[5] Rozin and Fallon 1987 claim that the relevant feeling is disgust. Yet, disgust does not seem to play a role for the uncanny valley, at least it was not among the feelings subjects reported after having been confronted with creatures judged to be uncanny (see MacDorman and Ishiguro 2006, p. 312).

uncanny effects may be mitigated. It may be that any level of realism can be socially engaging if one designs aesthetics well." (Hanson 2006)

What can we say about this? *First* of all, Mori's claim does not have to be read as strong as Hanson's critique demands. He does not have to say that any very humanlike object not being a real human necessarily falls into the uncanny valley. It is enough to say that a lot of things do, granting that the uncanny valley can be overcome sometimes. *Secondly*, and more directly concerning the experimental apparatus, one can criticize that the selection of pictures in the "tuned" series directly jumps from a doll to something very humanlike. So it simply avoids the critical stage, since Mori did not think that dolls fall in the uncanny valley.[6] Moreover, it is debatable whether the kind of repulsion caused by ugly conspecifics is of the same kind as the feeling of eeriness caused by objects in the "uncanny valley". Ishiguro modeled android robots that many people find eerie after his own and his daughter's physical appearance, and they are both rather on the nice than on the ugly side of the divide. Despite my critique of Hanson, I am prepared to accept that different factors might influence the occurrence of uncannyness, among them aesthetic ones. So maybe Mori did not get the shape of the graph correctly, and the function is not steady, as he believed.

Another study (Mac Dorman 2006) also points in this direction. MacDorman made the participants of his experiment watch 14 video clips showing a wide range of mainly humanoid and android robots performing different actions in different contexts, sometimes with speech accompaniment, sometimes not. With this experimental setup he tried to overcome the restriction to stills in the morphing experiments. The subjects had to evaluate the videos according to two types of criteria: mechanical vs. humanlike and familiar vs. strange. The results did not indicate a single uncanny valley for a particular range of human likeness. Therefore, MacDorman concludes: "Rather, they suggest that human likeness is only one of perhaps many factors influencing the extent to which a robot is perceived as being strange, familiar, or eerie." (Mac Dorman 2006) However, there is also another way to interpret the results of the experiment. Maybe the category "human likeness" is not sufficiently fine-grained, and we have to study more subtle characteristics with respect to human likeness.[7] Which ones these are is up to further empirical research (see, for instance, Ramey 2006). I think all these accounts do touch some important aspects of the problem. What is lacking, however, is an explanation at a more abstract level of the cognitive and emotional mechanisms responsible for the emergence of the uncanny valley. Therefore, I am now going to outline a more philosophical approach which can provide such a unifying explanation. I will begin by addressing question (1), why we feel empathy with inanimate objects at all, which has not received as much attention as the other two. The first step will consist in the development of a concept of empathy which is apt for inanimate objects.

[6] For some more general reservations on morphing as a means to investigate the uncanny valley, see MacDorman/Ishiguro 2006, 308.

[7] It has been argued that the distinction between appearance and behavior is of particular importance (Götz et al. 2003).

Empathy With Inanimate Objects

In ordinary language, *empathy* is often characterized as the ability to "put oneself into another's shoes" where this is meant to involve feeling the other's emotion within oneself. *Sympathy*, in contrast, is an emotional reaction to the situation of others which does not involve experiencing the emotions they feel. Empathy demands that the emotional reaction is more appropriate to the other person's situation, and that you feel it, because the other person feels it. Let us take the definition of empathy proposed by Elliott Sober and David Wilson as a starting point:

(1) "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." (Sober and Wilson 1998, p. 234f.)

Obviously, (1) gives a too restricted definition of empathy for our purposes. The demand that there be a *belief* that O feels E gets in conflict with desideratum (1) of the sought explanation: how we can feel empathy to inanimate objects at all. I take it for granted that we do not have to *believe* that these things really have feelings. The definition of Sober and Wilson, therefore, conceptually excludes empathy with inanimate objects.

However, if we do not want to beg the question, we should also look for independent evidence against the claim that we have to believe that O feels E in order to empathize. As a matter of fact, we sometimes react immediately with empathy when perceiving another's emotional expression. We feel sad when seeing someone in tears or happy when we see someone smiling. It is seeing their facial expression or hearing the tone of their voice that produces the feeling. The belief that a person is happy or sad emerges only simultaneously or even subsequently to our emotional reaction. This is supported by the fact that children can emotionally respond to other's feelings before they are able to form beliefs about them.[8] Moreover, adopting a facial expression can produce the relevant emotion in us without invoking beliefs. At least sometimes smiling makes us happy (this was pointed out already by James 1884, see also Johnson-Laird and Oately 2004) There is experimental evidence that this plays are role for face-based emotion recognition and empathy, as well. We can recognize other's emotions by perceiving and simulating their facial expressions, and, thereby, producing the same feelings in us.[9] As this suggests, we come to feel like others at least sometimes just by seeing how they feel. I, therefore, propose to complement the definition of Sober and Wilson with a second form of empathy, taking up a proposal of Maibom:

(2) "*S* empathizes with *O*'s experience of emotion *E* if S perceives O's T-ing and this perception causes S to feel E for O." (Maibom 2007, p. 168)

[8] This ability seems to be related to the propensity of neonates to mimic facial expressions (Meltzoff and Moore 1983).

[9] Evidence for this is provided, for instance, by studies finding a reliable correlation between deficits in face-based emotion recognition of some emotions with deficits in producing the relevant emotions. For an instructive discussion in the context of simulationist models of emotion recognition, see Goldman and Sripada 2005. That this is, however, an intricate matter is shown by the fact that Goldman and Sripada also mention some evidence to the contrary.

Yet, (2) is still not apt for our purposes, because we neither want to say that inanimate objects really have emotions nor that they literally show emotional expressions, facial or otherwise. And if they do not literally express their emotions, we cannot literally perceive them. Therefore, I am suggesting that the imagination must somehow be involved in the emergence of empathy with inanimate objects. This idea is captured by the following modification of (2):

(3) S empathizes with an inanimate object's imagined experience of emotion E if S imaginatively perceives the inanimate object's T-ing and this imaginative perception causes S to feel E for the inanimate object.

However, the big question is, of course: How is what I call "imaginative perception" supposed to work?

Imaginative Perception

In order to clarify the concept of an imaginative perception, I want to draw on the resources provided by the debate about pictorial representation where the notion has been most prominently discussed. The discussion about pictorial representation focuses on the peculiarities of pictorial representations in contrast to linguistic ones. It is guided by the intuitions that pictorial representations seem to be somehow less conventional than linguistic ones, that they somehow seem to be more similar to the depicted objects, and allow us to recognize them more easily without any special kind of semantic knowledge. An important account of pictorial representation put forward in different varieties says 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 (see, for instance, Walton 1990, 2002; Wollheim 1998; O'Shaughnessy 2002). I think we can gain some valuable insights from this approach concerning the interplay of perception and imagination in empathy with inanimate objects without having to endorse it as an adequate explanation of pictorial representation. Reference to the debate about depiction is supposed to help us to answer two questions: (1) how is imagining involved in the kind of imaginative perception relevant to empathy with inanimate objects? And (2) in which way are the perception and the imagining linked.

Let me start with question (1). The first thing to notice is that the imagining involved cannot just be propositional. Just *supposing* that an inanimate object's T-ing is a human T-ing does not lead to the right results, since there is no connection with perception, and we are dealing with perception-based empathy. Another straightforward suggestion can be discarded for similar reasons. Can we not simply say that we imagine the inanimate object's T-ing being a human T-ing? Yet, this does not get us far enough in the direction of perception, either. Let me explain why with the help of an example. We can imagine of 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 phenomenon we are after. As we conceived of the situation, it is the perception of an inanimate object that produces empathy in us, and not the pretending of its being human. This holds independently of how

exactly one thinks of pretense (for different accounts, see Leslie 1987, Currie 1995, Nichols and Stich 2000).

So let me finally introduce a more promising proposal made by Kendall Walton with respect to pictorial representation (Walton 1990). He claims that somebody who is looking at a picture, for instance, Meindert Hobbema's *Water Mill with the Great Red Roof* imagines of his looking that its object is a mill. In other words: in seeing the canvas he imaginatively sees a mill. As I said, I don't want to quarrel whether this is an adequate explanation of pictorial representation. However, the idea of imagining of a perceptual experience of one thing that it is a perceptual experience of something else extends quite naturally to non-pictorial cases. In a later article (Walton 2002) Walton himself gives an example when he is taking up a suggestion by Patrick Maynard in order to refute an objection against his account. He refers to Hitchcock's movie *Vertigo* where Scottie dresses up 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 T-ing of an inanimate object we imagine perceiving a human T-ing.

Let us now come to question (2): How are the two episodes of seeing something and thereby imagining perceiving something else related to each other? One might suggest as a minimal condition that the two occur *simultaneously*. However, this does not seem to be a sufficient condition, since we can entertain an imagining while at the same time having quite different perceptual experiences, e.g., when I am imagining going out tonight while looking out of the window of my office. Yet, we have already established a stronger link than mere coincidence, since the two episodes are intentionally related: my perception is supposed to be the object of my imagining. Yet, what does this mean exactly? The desideratum which has to be met by an adequate account is that imagining of one experience to be another must be something more experiential than supposing or pretending that one experience is the other (Wollheim 1998). A straightforward suggestion could be that the imagined perceiving consists in visualizing: seeing a picture in this view amounts to a fusion of two things, the experience of seeing the surface of the picture and the experience of visualizing the scene depicted. This account has some initial plausibility with respect to the distinctive phenomenology of pictorial experience (Budd 1992)

However, at this point we can leave the analogy with pictorial representation behind for two reasons: *first,* I don't think that perception-based empathy with inanimate objects shares its distinctive phenomenology with pictorial experience. *Secondly,* it does certainly not involve visualization. We do not visualize a human T-ing when we imaginatively perceive the T-ing of an inanimate object as a human T-ing. In order to come to terms with this problem, I want to have a closer look at what perceptual experience is. Of course the debate about perceptual experience is vast, but I won't go into detail, because I hope to get along without any deep theoretical commitments in this area. Perceptual experience, I take it very generally, has two aspects: it has content, and it has some phenomenal character (in the broadest sense of the term which does not commit one to any particularly strong

position in the qualia debate—or so I hope…) Setting the problem of perceptual content aside let's just care about phenomenology. The phenomenal feel of a perceptual experience is what it is like for a subject to have it. To be sure, perceptual experience does not just include vision, but the other sense modalities, as well, and each of them has a particular phenomenology. Sometimes the boundaries between them may be hard to draw, for instance, in the case of certain olfactory and gustatory phenomena, but that is not to say that this is always like this. Moreover, there is a phenomenal difference in perceptual experiences within different sensory modalities, e.g., between seeing red and seeing blue.

Now, what I want to suggest is that imagining of a perceptual experience of an (or the) object F that it is a perceptual experience of an (or the) object G involves the following: Because of certain salient similarities between the perceived object F with another object (or kind of objects) G the concept of a (or the) G is triggered, but it is not applied to the perception, but entertained "off-line" to use a common metaphor in the theory of the imagination. For this reason it is a kind of imagining, although not actively done. Despite 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.[10] Yet, the concept doesn't change the content of the perception—I am still seeing an F and not a G—but it does influence its phenomenal feel, such that 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 up to full application. The more strongly it is triggered, the more vivid is the imagining, and the more feels the perception of F 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.

To make this a little less abstract let's come back to the Vertigo-example: Because of the dressing up of Judy as Madeleine the perception of Judy triggers the concept of Madeleine, and that has the effect that perceiving Judy feels somehow like perceiving Madeleine. The more features both women share, and the more typical and salient they are the more does the perception of Judy feel like a perception of Madeleine. The perception of Judy that feels somehow like a perception of Madeleine then produces emotions (like hate or affection) the protagonist would have towards the real Madeleine, but not to the real Judy. Applied to empathy towards inanimate objects, that is to say: The human-like features M of an inanimate object trigger the concept of a human N, for that reason, seeing the T-ing of an inanimate object feels (to some extent) like seeing a human T-ing, e.g., the human-like features of a doll's face trigger the concept of a human face, for that reason, seeing the smile of a doll feels like seeing the smile of a human being. Given that perceiving facial expressions can cause the same emotions in us, as argued above, we have by now arrived at a the core of my explanation of how empathy with inanimate objects is possible.

[10] This might involve something like a blending mechanism (Fauconnier and Turner 1998, 2002).

Even if one grants some initial plausibility to this explanation one might nevertheless be troubled by several worries, since the position I am trying to defend seems to deviate in some important aspects from philosophical mainstream. The *first* question which arises is whether this account commits me to an implausibly strong view concerning qualia. I tend to think that this is not the case, and that my position is acceptable to anybody except eliminativists *tout court*. The reason for this optimism is that my 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 the 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* (Horgan and Tienson 2002). It says that mental states of the sort commonly cited as paradigmatically intentional have phenomenal character that is inseparable from their intentional content. I am, however, not committed to hold the reverse claim labeled the *Intentionality of Phenomenology* saying that mental states of the sort commonly cited as paradigmatically phenomenal (e.g., sensory-experiential states such as color-experiences) have intentional content that is inseparable from their phenomenal character. The second thesis is just not relevant for the issue at stake. For that reason I neither have to accept nor to reject it. Of these two the first claim seems to me rather innocuous whereas the second is not at all trivial with respect to the qualia debate.

However, one might still challenge my position by pointing out that it is not at all evident to assume that concepts have phenomenal character, at all, only perceptions or sensations do. This objection is fed by a certain view of concepts which is widely held in philosophy. It analyses concepts with the help of abstract models which are detached from bodily aspects. One popular version of this view (see, for instance, Fodor 1998) assumes that concepts are abstract, amodal and arbitrary representations in a "language of thought" made up of symbols. However, this approach has been challenged by evidence from clinical neuropsychology and cognitive neuroscience (Damasio 1989, Martin and Chao 2001, Caramazza and Martin 2003, Rumiati and Caramazza 2005, Pecher and Zwaan 2005; from a more philosophical point of view, Prinz 2005). The evidence is supposed to show that distributed networks of discrete brain regions are active during object processing, that the distribution of these categories varies as a function of semantic category, and that 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 a view of concepts as "embodied" was proposed which is claiming that (at least concrete perceptual) concepts are neural representations located in sensory-motor areas in the brain (Gallese and Lakoff 2005, who even want to expand this view to all kinds of concepts). That is, 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 also have a qualitative aspect.

However, I am not going defend this theory in detail here. My aim in invoking it was just to show that there are accounts of concepts on which the idea that concepts have a qualitative dimension is not that far-off. However, if any proponent of

another theory of concepts is ready to concede that much, too, I am happy with that. On the other hand, one might be tempted to see in my explanation of the uncanny valley a peace of evidence in favour of the account of "embodied" concepts. However, we have not yet reached the full-fledged explanation of the uncanny valley, but stopped half way with the explanation of how empathy with inanimate objects is possible. After having done some trouble-shooting with respect to the invoked account of concepts we can now use our tools to finally explain the uncanny valley.

Explaining the Uncanny Valley

Before I get to the core of the explanation I want to show that the empirical findings discussed earlier fit in quite smoothly with the suggested account. To begin with, we can understand why, as a rule of thumb, the more human-like an object is the more empathy it will cause, and why this is *just* a rule of thumb. An increase of human likeness in terms of number, salience and typicality of features will trigger the relevant human concept more strongly. However, the interplay of the different dimensions does not allow for a clear ranking such that innumerable trade-offs are possible: An object which does not display a great number of human-like features, but very typical and salient ones, might do better in terms of perception-based empathy than an object that shares a great number of human-like, even typical features which are not salient (for instance, because it looks for some reason entirely different).

We can also understand the influence of aesthetic features investigated in the study of Hanson referred to above. As he points out quoting several studies, averaged faces (i.e., faces generated out of a great number of different faces) are considered particularly attractive. Yet, the faces considered most attractive deviate from the average in very specific ways, usually in features associated with neoteny, sexual maturity, or senescence, where each of these exaggerated feature-sets inspires different behavior in humans: neoteny features arouse nurturing, sexual maturity features inspire both sexual attraction and friendship, and senescence features encourage mentoring relationships.[11] These findings suggest two things: *first*, the influence of aesthetic features can be spelled out in terms of typicality and salience, and *secondly*, different features may be relevant for empathy with different kinds of emotions.

Of the three questions I formulated in the beginning, we have by now dealt mainly with the first one: Why do we feel empathy with inanimate objects at all? There remain the other two, which are the really tough ones with respect to the puzzle of the uncanny valley: Why do we stop to feel empathy when an entity becomes very human-like, and why do we not just stop to feel empathy but respond with eeriness? I think the baseline of the explanation has to be that in these cases the triggering of the concept gets so strong that it is about to turn into full-fledged concept application. However, the attempt to apply the concept fails since the object

[11] Hanson refers here to studies by Etcoff 2000, as well as several articles in Rhodes and Zebrowitz 2002.

of the perception is not accepted as an instance of the concept. This is the reason why the process leading to empathy is interrupted. Yet, because of the similarity of the features the concept is triggered again and is repeatedly about to be elicited. This leads to a kind of very fast oscillation between four situations which resembles a gestalt switch: 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 keeping on to perceive the object. This reminds a bit of a radio receiver trying to tune into a transmitter in bad conditions when the reception of one station is always interfered with by another one and sheer noise.

This conception can be underpinned empirically by the study of vision (Stark et al. 2001). Research in the perception of ambiguous or fragmented figures showed that, as the visual impression changes, so does the scan-path of the eye-movements (i.e., the sequence of fixations).[12] 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" and "*a* is not a human being" would take place of which one would expect that it is correlated with incoherent eye-movements. It seems to be more than probable that this is 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 (this position is held by Ramey 2005).

It is a matter of empirical research to gather more evidence with respect to this question, but there is reason to think that there is something special about the uncanny valley. 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.[13] 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 disenchantment in the Weberian sense: something that seemed to be alive and soulful a moment ago now appears cold and dead. The alternation between these two states amounts to the feeling of eeriness. This gets us to the true core of the intuition that the uncanny valley has something to do with the terrors of death.

References

Bartneck, C., et al. (2005). Robot abuse—a limitation of the media equation. *Proceedings of the Interact 2005 Workshop on Agent Abuse*, Rome.

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

[13] 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).

Becker, E. (1973). *The denial of death*. New York: Free Press.
Budd, M. (1992). On the foundations of representational arts. *Mind, 101*, 195–198. doi:10.1093/mind/101.401.195.
Caramazza, A., Martin, A. (Eds.) (2003). *The organisation of conceptual knowledge in the brain. Cognitive Neuropsychology 20 (special issue).*
Currie, G. (1995). Imagination and simulation: Aesthetics meets cognitive science. In A. Stone & M. Davies (Eds.), *Mental simulation: Evaluations and applications* (pp. 151–169). Oxford: Blackwell.
Damasio, A. R. (1989). Time-locked multiregional retro activation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*, 25–62.
Etcoff, N. (2000). *Survival of the prettiest*. New York: Anchor Books.
Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science, 22*, 133–187.
Fauconnier, G., & Turner, M. (2002). *Conceptual blending and the mind's hidden complexities*. New York: Basic Books.
Fodor, J. A. (1998). *Concepts: Where cognitive science when wrong*. Oxford: Oxford UP.
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., & Sripada, C. S. S. (2005). Simulationist models of face-based emotion recognition. *Cognition, 94*, 193–213.
Götz, J. et al. (2003). Matching robot appearance and behaviour to tasks to improve human-robot cooperation. *Robot and Human Interactive Communication*, doi: 10.1109/ROMAN.2003.1251796.
Hanson, D. (2006). Expanding the aesthetic possibilities for humanlike robots. In *Proc. IEEE Humanoid Robotics Conference*, special session on the Uncanny Valley, Tskuba, Japan.
Ho, C.-C. et al. (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. In: D. Chalmers (Ed.). *Philosophy of mind: Classical and contemporary readings*. Oxford: Oxford UP, 520–33.
James, W. (1884). What is an emotion? *Mind, 9*, 188–205.
Johnson-Laird, P., & Oatley, K. (2004). Cognitive and social construction in emotions. In M. Lewis & J. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 458–475). New York: Guilford Press.
Leslie, A. M. (1987). Pretense and representation: the origins of 'theory of mind'. *Psychological Review, 94*, 412–426.
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.
MacDorman, K. F. (2005b). *Mortality salience and the uncanny valley*, international conference on humanoid robots, Tsukuba, Japan, doi: 10.1109/ICHR.2005.1573600.
MacDorman, K. F. (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.
MacDorman, K. F. 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.
MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies, 7*(3), 297–337.
Maibom, H. L. (2007). The presence of others. *Philosophical Studies, 132*, 161–190.
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*, doi: 10.1109/DEVLRN.2005.1490953.
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.
Mori, M (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*, 115–147.

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*. Cambridge: Cambridge UP.
Poe, E. A. (1846). The philosophy of composition. *Graham's Magazine, April, 1846*, 163–167.
Prinz, J. J. (2005). The return of concept empiricism. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 679–699). Oxford: Oxford.
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.
Ramey, C. (2006). *An inventory of reported characteristics for home computers, robots, and human beings: Applications for android science and the uncanny valley*. Proceedings of the ICCS/CogSci-2006 Long Symposium 'Toward Social Mechanisms of Android Science', Vancouver, Canada.
Rhodes, G., & Zebrowitz, L. A. (Eds.). (2002). *Facial attractiveness: Evolutionary cognitive and social perspectives, Westport*. Westport: Greenwood Publishing Group.
Rozin, P., & Fallon, A. E. (1987). A perspective on disgust. *Psychological Review, 94*, 23–41.
Rumiati, R. I., & Caramazza, A. (Eds.). (2005). *The multiple functions of sensory-motor representation. Cognitive Psychology 22 (special issue).*
Rutherford, M., et al. (2006). The perception of Animacy in young children with autism. *Journal of Autism and Development Disorders, 36*, 983–992.
Scholl, B., & Tremoulet, P. (2000). Perceptual causality and animacy. *Trends in Cognitive Science, 4*, 299–309.
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.
Sober, E., & Wilson, D. S. (1998). *Unto others*. Cambridge: Harvard UP.
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.
Walton, K. (1990). *Mimesis as make-belief*. Cambridge: Harvard UP.
Walton, K. (2002). Depiction, 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. http://www.wired.com/wired/archive/10.06/face_pr.html. Cited 21 Mar 2008.
Wollheim, R. (1998). On pictorial representation. *The Journal of Aesthetics and Art Criticism, 56*, 217–233.