Beyond the Uncanny Valley: A Theory of Eeriness for Android Science Research

By

Joshua David Bruce Redstone

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Arts

in

Philosophy

Carleton University
Ottawa, Ontario

© 2013, Joshua David Bruce Redstone
Abstract

The “Uncanny Valley” (UV) theory predicts that highly realistic human-like artifacts, e.g. robots and animated characters, sometimes elicit eeriness in human subjects. But in spite of a rapidly growing body of interdisciplinary research on the UV, a lack of consensus concerning the cause(s) of this phenomenon persists. Therefore in what follows, I undertake a conceptual “overhaul” of the UV theory in order to facilitate an account of the UV phenomenon. Drawing from philosophical and empirical research, I demonstrate that: (1) eeriness is best understood as anxiety caused by uncertainty concerning the ontological nature of the artifact; and (2), that the misfiring of cognitive and affective empathic abilities – viz. an inconsistency between a subject’s perception of the artifact’s apparent animacy and mentation, and her knowledge that artifacts ought not possess such attributes – is the primary causal mechanism of the UV phenomenon. Finally, I revise the UV theory accordingly.
Acknowledgements

I am deeply indebted to my professors and colleagues at Carleton University for their support throughout the completion of this thesis. First and foremost, I am grateful for the guidance and direction of my supervisor, Dr. Heidi L. Maibom. Her expertise, critical commentary and stylistic advice helped me to take this project far beyond what I had initially envisioned. I am likewise grateful for the comments I received on an early conception of this project from Dr. David Matheson and my fellow students in Carleton University’s philosophy graduate research seminar. This early feedback has continually proven to be essential. The ongoing assistance of our departmental administrator, Sandy Kirkpatrick, also facilitated the completion this project. I also wish to express thanks to my defense committee chair, Dr. Christine Koggel, and to my external examiner, Dr. Mark MacLeod.

I’d like to extend further gratitude to a number of current and former colleagues who have helped to inspire and/or direct me. My thanks once again go out to David Matheson, as well as to Matthew Scarfone and Christopher Bannon, for helping me to decide upon the quotations which commence Sections 2 and 4. Thanks as well to Kia Dunn, who brought the existence of a new – and properly naturalized – translation of Masahiro Mori’s essay, “The Uncanny Valley” to my attention. I am also grateful to Jonathan Reardon and Vikram Panchmatia of the Carleton University Philosophy Society (CUPS) for providing me with the opportunity to test-run some of my early ideas about this project at one of CUPS’ weekly philosophy events.
Of course, I cannot forget about my friends and family. It is no exaggeration to say that without the support of my parents, my sister and brother-in-law, and my grandparents, I could never have come this far. I’m also grateful to Linda Piper for showing me that university was the right place for me, and to Dr. Wayne Borody, Dr. Toivo Koivukoski, and the late Georgia Irwin, who encouraged me to continue my academic journey at graduate school. Last – but certainly not least – thank you Sara for your moral support, for your keen eye for typos, and for putting up with my obsession with both the material discussed in this project, and with my seeing it through to completion.

Carleton University

August 2013
# Table of Contents

Abstract .......................................................................................................................... i

Acknowledgements ........................................................................................................ ii

Table of Contents ............................................................................................................ iv

Introduction ...................................................................................................................... 1

1.0 – On Android Science and the Uncanny Valley ....................................................... 9
    1.1: The Uncanny Valley Theory .................................................................................. 10
    1.2: Research Approaches in Android Science ............................................................ 12
    1.3: Does the Uncanny Valley Phenomenon Exist? ..................................................... 17
    1.4: Putting “the Uncanny” in the Uncanny Valley ...................................................... 23
    1.5: Empathy and the Uncanny Valley Phenomenon ................................................... 26
    1.6: The “Imaginative Perception” of Emotion ............................................................. 32
    1.7: Empathy and the Uncanny Valley Theory ............................................................. 36

2.0 – Naturalizing the Uncanny ..................................................................................... 40
    2.1: What is “the Uncanny”? ....................................................................................... 42
    2.2: Jentsch and Freud on the Uncanny ....................................................................... 45
    2.3: Psychological Uncertainty or the Return of the Repressed? ................................. 55
    2.4: Toward a New Psychology of the Uncanny .......................................................... 60

3.0 – On the Emotion of Eeriness .................................................................................. 62
    3.1: What are the Emotions? ....................................................................................... 63
    3.2: Eeriness and Fear ................................................................................................ 65
    3.3: Feeling Eerie ......................................................................................................... 67
    3.4: Angst, Mortality Salience, and the Uncanny Valley Phenomenon ......................... 71
    3.5: What Eeriness Is .................................................................................................. 77

4.0 – Categorical Anomalies and the Uncanny Valley Phenomenon .............................. 82
4.1: What is a “Category”? ................................................................. 84
4.2: Categorization Difficulty and the Uncanny Valley Phenomenon ........................................ 88
4.3: Cognitive Dissonance, Alief, and the Uncanny Valley Phenomenon .................................. 98
4.4: Psychological Uncertainty Reconsidered ............................................................................ 108

5.0 – Of Minds and Machines ......................................................................................... 112
5.1: Perceiving Artifacts with Minds .................................................................................. 114
5.2: Knowing Other Minds ............................................................................................... 119
5.3: “Re-imagining” the Imaginative Perception Hypothesis ............................................... 129
5.4: Completing the Map of the Uncanny Valley ............................................................... 133

Conclusion ..................................................................................................................... 140
References ....................................................................................................................... 145
Introduction

Oh! No mortal could support the horror of that countenance. A mummy again endued with animation could not be so hideous as that wretch. I had gazed on him while unfinished; he was ugly then; but when those muscles and joints were rendered capable of motion, it became a thing such as even Dante could not have conceived.

– Victor Frankenstein, in Mary Shelly’s _Frankenstein; Or, the Modern Prometheus_ (1818)

The so-called “Uncanny Valley” theory – proposed by roboticist Masahiro Mori in 1970\(^1\) – predicts that a subject experiences a positive emotional response (\(y\)), _viz._ “familiarity”/“affinity” (*i.e.* _shinwakan_)\(^2\) toward a human-like artifact such as a robot, an automaton, an animated character, a mannequin, a doll, a puppet, _etc_. This positive response is a linear function of the artifact’s level of human-likeness (\(x\)). However, this relationship is only maintained until the artifact’s appearance nears complete human similitude. At this point, a subject’s emotional response quickly turns negative, and she experiences a sense of eeriness (*i.e.* _bukimi_). Mori also maintains that this sense of eeriness is intensified when the artifact is in motion. Upon encountering such an artifact, a subject might feel somewhat like Victor Frankenstein, who was horrified when he brought his creation to life using a bolt of electricity. The end result is a non-linear relationship between \(x\) and \(y\), resulting in an “uncanny valley” in the function \(y\) as \(x\) approaches complete anthropomorphism. Thus the “uncanny valley phenomenon” occurs.

Whether or not this phenomenon really takes place, and if so, why and how it comes to

---

\(^1\) Cf. Mori (1970/2005; 1970/2012). I primarily reference the latter translation, for it is properly naturalized, _i.e._ made to sound “natural” to an English speaking reader rather than something that was simply transcribed directly from Japanese, unlike the former translation. Gratitude goes out to my colleague Kia Dunn for bringing the existence of this improved translation to my attention.

\(^2\) Note that I focus on the original Japanese terms because their meaning, and their possible translations, shall become important later. I list all terms in Roman rather than in Japanese characters.
occur, are interesting empirical questions. However, I think that putting the results of empirical research on the uncanny valley to use at gaining an understanding of human emotional responses toward robots and other such artifacts requires a more comprehensive approach than experimentation alone.

The approach taken toward ascertaining the cause(s) of the uncanny valley phenomenon in this investigation is philosophically oriented. But the subject matter I am concerned with is very unusual for philosophy, so I had better explain myself just a little bit further. Indeed, apart from say, questions concerning artificial intelligence, roboethics, or the fun and imaginative thought-experiments and “what-if?” scenarios found in science fiction,\(^3\) philosophers usually don’t pay much attention to the field of robotics. But my approach toward this subject matter is “philosophical” insofar as I return to the proverbial drawing board in order to determine whether or not the concepts Mori (1970/2012) uses in his uncanny valley theory – viz. the level of anthropomorphic appearance and movement, “familiarity” or “affinity,” and finally eeriness – actually represent the contributing factors and the emotional phenomena involved in the uncanny valley phenomenon. By giving the uncanny valley theory this “conceptual overhaul,” I hope to make it easier for empirical researchers to explain the uncanny valley phenomenon, and perhaps even to remove some of the mystery behind its cause(s) along the way.

\(^3\) For example, in films such as Ridley Scott’s Blade Runner (1982); or, in television programs such as the many different Star Trek series.
Of the aforementioned concepts, eeriness and familiarity (*bukimi* and *shinwakan*, respectively)⁴ are of particular interest here, although anthropomorphic appearance and movement are to be discussed as well. For these first two concepts describe the emotional responses that human-like artifacts elicit according to Mori’s (1970/2012) theoretical framework. Philosopher Catrin Misselhorn (2009), however, uses “empathy” – not “familiarity” or “affinity” – to describe the positive emotional response associated with human-like artifacts. She is also the first to articulate an explanation for the uncanny valley using a philosophical-empirical – rather than a strictly empirical – explanatory approach. That is, she pays close attention to the core concepts that empirical researchers employ in their attempts to understand the uncanny valley phenomenon, as well as to the data obtained from their research studies. In this investigation I adopt a similar approach, with a focus on capacities for empathy instead of familiarity, or other similar translations of *shinwakan*.⁵ Here, “empathy” refers to both the ability to know what other minds think, believe, desire, *etc.*, and the ability to experience emotions consonant with someone else’s. These are known as “cognitive empathy” and “affective empathy,” respectively.⁶ Furthermore, Misselhorn thinks that to explain the eerie sensation elicited by human-like artifacts, one has to account for the way that a subject’s positive emotional response is related to her negative response, *i.e.* the emotion of eeriness. This is another aspect of her approach that I adopt here. However as the reader shall see, Misselhorn and

---

⁴ Note again that Mori’s paper was published in Japanese originally.
⁵ Interestingly, that empathy plays some role in the uncanny valley phenomenon has also been suggested by MacDorman, Green, Ho & Koch (2009) (henceforth: MacDorman *et al* 2009b), who argue this is evidenced by fMRI neuroimaging studies of perception and human-like artifacts (*e.g.* Gazzola *et al* 2007). Since their approach toward the uncanny valley phenomenon is empirical and not philosophical, I discuss it alongside other empirical literature I review.
⁶ *Cf.* Maibom (2012) for further discussion on these two empathic abilities.
I differ on our explanations for the uncanny valley phenomenon, and how empathy with human-like artifacts misfires.

I shall also provide something which is largely absent from Mori’s (1970/2012) and Misselhorn’s (2009) respective theories of the uncanny valley, viz. an account of the emotion of eeriness. For without knowing what kind of emotion eeriness is, researchers will not be able to explain why human-like artifacts – or other things which are described as “eerie” – elicit this particular emotional response. The problem here is that little has been said about what kind of emotion eeriness is outside of the empirical research that concerns the uncanny valley phenomenon. Therefore, this investigation has two objectives: (1) to determine what kind of emotion eeriness is; and (2), to ascertain why/how human-like artifacts elicit this response. I shall argue – taking my cue from Misselhorn (2009) – that a subject’s empathic abilities “misfire” during interaction with human-like artifacts, and that this can explain the change in a subject’s emotional state from positive to negative. However, my theory differs in several key respects from hers: while we both draw upon some of the same examples from empirical research, I take care to include other empirical results from a more diverse range of disciplines; moreover, the theoretical model I finally arrive at, as I intend to show, offers some conceptual advantages over Misselhorn’s. This shall be made clear in the final section of this investigation.

So much for what I hope to accomplish in this investigation; now I shall say a few words about how I intend to accomplish these objectives. In Section 1, I begin by discussing research on the uncanny valley from android science. Android science is a field that consists of complimentary research approaches in robotics and cognitive
science, wherein Mori’s (1970/2012) theory is of particular interest. I elaborate on the research concerns of android science, and show that there are good reasons to suppose that the uncanny valley phenomenon does occur with some human-like artifacts. I then lay the groundwork for demonstrating that empathy – not “familiarity” (shinwakan) – explains the change in a subject’s emotional response from positive to negative. Since my account of the uncanny valley is developed in parallel, but also in contrast to, Misselhorn’s (2009) theory, I introduce her theory in the latter-half of Section 1. I’ll return to it later in Section 5 after I complete a number of other important objectives.

For instance, one cannot explain why human-like artifacts elicit a sense of eeriness if one does not know what kind of emotion eeriness is. I address this conceptual issue throughout Sections 2 and 3. Specifically, in Section 2 I examine literature on the uncanny from historical psychology; in particular, I perform an exegesis of the work of Ernst Jentsch (1906/2008) and Sigmund Freud (1919/2003). Since there is not much empirical data available on either eeriness or the psychology of the uncanny, my aim here is to provide some general sense of what the “psychology of the uncanny” – especially as it pertains to the uncanny valley phenomenon – is like. I then turn directly toward the question “what kind of emotion is eeriness?” in Section 3, in order to see how eeriness is related to the uncanny. I examine what the emotions are, and consider what some of the results of preliminary empirical research on the uncanny valley phenomenon reveal about the emotion of eeriness. I argue that eeriness is a type of anxiety experienced when a subject is uncertain about the ontological nature of something she encounters in the

---

7 The “Psychology of the Uncanny,” is the title of Ernst Jentsch’s (1906/2008) essay. I borrow the term because it provides a sense of the final goal of this investigation, viz. a psychological account of the uncanny valley phenomenon.
world. In the case of the uncanny valley phenomenon, the object of a subject’s uncertainty is whether or not a person, artifact, or feature of her environment is (1) animate, and (2) possesses a mind. This idea is found in much of the research I consider throughout Sections 2, 4 and 5, e.g. in Jentsch (1906/2008), Gray & Wegner (2012), Misselhorn (2009) and Ishiguro (2006; 2007), amongst others.

Having explained Mori’s (1970/2012) uncanny valley theory, and having obtained some idea of what both the psychology of the uncanny and the emotion of eeriness are like, I examine further possible causes of the uncanny valley phenomenon in Section 4. Here, I focus on the results of recent research results from cognitive psychology, which suggest that the sense of eeriness elicited by human-like artifacts results in some way or another from a subject’s difficulty at placing the uncanny stimulus into a discrete ontological category. I consider the possibility that the psychological state constitutive of this difficulty is analogous to cognitive dissonance, which is a suggestion of Burleigh, Schoenherr & Lacroix (2013) and MacDorman, Green, Ho & Koch (2009). I show that dissonant psychological processes do occur in a subject’s mind when the uncanny valley phenomenon takes place. However, I think that these are better understood as dissonant beliefs or articles of knowledge and what Tamar Gendler (2008) calls “aliefs” – i.e. tendencies to respond to an apparent stimulus in a particular way, which can be either innate or habitual – rather than as dissonant “cognitions,” according to how “cognition” is typically understood on cognitive dissonance theory (cf. Festinger 1957; Wicklund & Brehm 1976). This kind of psychological dissonance is compatible with the definition of eeriness that I provide in Section 3.
In the above-outlined sections, I focus primarily on one saliency that contributes to the uncanny valley phenomenon, viz. apparent animacy. In Section 5, however, I emphasize the other aforementioned saliency, i.e. the appearance of a mind in a human-like artifact. I examine mind perception (Gray & Wegner 2012; Gray, Gray & Wegner 2007) and empathic abilities (Misselhorn 2009; also cf. Maibom 2007a; 2007b; 2012) to determine how these factor into the uncanny valley phenomenon, focusing on both the psychological and neurological levels of cognitive empathy or “folk psychology.” Here, I show that folk-psychological reasoning misfires in cases of the uncanny valley phenomenon such that the aliefs a subject ought to have only when she interacts with conspecifics are also elicited by human-like artifacts. These aliefs then become dissonant with her beliefs and/or judgements concerning the ontological nature of said artifact, viz., that it is not a human, that it is an artifact, that it does not possess animacy or mentation, etc.

In other words during interaction with a human-like artifact, a subject believes/judges/knows that she is in the presence of an automaton. At the same time, however, she alieves that she is interacting with a conspecific. On this model, the uncanny valley is an example of the occurrence of what Gendler (2008) calls “norm-discordant alief,” a concept to be fully explained in Section 5. This psychological dissonance, I argue, accounts for the eerie sensation in the uncanny valley phenomenon. By articulating this alternative cognitive dissonance-type framework for the uncanny valley theory, I will have completed the aforementioned conceptual overhaul of the uncanny valley theory, replacing “familiarity” (shinwakan) with “empathy.”
As this investigation proceeds, I’ll consider several mainstream explanations for the phenomenon, some of which I have gestured at above. They are: “Mortality Salience” (e.g. MacDorman 2005), “Categorization Difficulty” (Ramey 2005; Kang 2011; Burleigh et al 2013; Yamada et al 2013) and “Imaginative Perception” (Misselhorn 2009). All of these are introduced in Section 1. However, since I draw quite heavily upon some of Misselhorn’s ideas in order to explain the uncanny valley phenomenon, I shall devote more attention to her explanation in the first section than the other two explanations. Instead I’ll consider mortality salience and categorization difficulty in Sections 3 and 4, respectively. Some elements of all of these explanations shall be retained and modified; others shall be rejected. In any case, as the reader can see, there is a lot of ground to cover in the following pages. The aforementioned material is drawn from a variety of different disciplines ranging from robotics to philosophy. Since empirical research on the uncanny valley phenomenon is still in the early stages, I shall be covering some uncharted territory, and perhaps even some undiscovered country as well. So let us begin this task with a close examination of Mori’s (1970/2012) theory – and the phenomenon it purports to explain – in order to ensure that everything which follows is set upon a firm foundation.
1.0 – On Android Science and the Uncanny Valley

NATURE (the Art whereby God hath made and governes the World) is by the Art of Man, as in many other things, so in this also imitated, that it can make an Artificial Animal. For seeing life is but a motion of Limbs, the beginning whereof is in some principall part within; why may we not say that all Automata (Engines that move themselves by springs and wheeles as doth a watch) have an artificiall life? For what is the Heart, but a Spring; and the Nerves, but so many Strings; and the Joynts, but so many Wheeles, giving motion to the whole Body, such as was intended by the Artificer?

– Thomas Hobbes, in Leviathan: Or, the Matter, Forme and Power of a Common-Wealth Ecclesiasticall and Civil (1651)

Before I ascertain what kind of emotion eeriness is, and how human-like artifacts such as robots, automata and animated characters are capable of eliciting it, it would be wise to examine the uncanny valley theory in detail. This is one of the objectives of the present section. I’ll also focus on how this theory has become an object of interest for android scientists, and reveal some of the problems that they face when it comes to explaining the uncanny valley phenomenon. For that matter, I shall also discuss what android science is, for it is an emerging discipline with which the average reader may well not be familiar.

As I mentioned at the outset of this investigation, there is a lack of consensus on the cause(s) of the uncanny valley phenomenon; there is also some controversy concerning whether it even exists at all. However, I shall demonstrate below that there are good reasons to suppose that the uncanny valley – or a phenomenon uncannily similar to the one that Mori (1970/2012) describes – does exist.8 Afterwards, I briefly introduce some of the principles of Misselhorn’s (2009) explanation of the uncanny valley phenomenon, so as to provide a few preliminary reasons for replacing “familiarity” or “affinity” with “empathy” in the uncanny valley theory. I’ll also sketch out a few of my

8 I present further evidence for this claim in Sections 2 and 4.
own reasons for making this conceptual change. For now, let us see what else Mori (1970/2012) has to say concerning the sense of eeriness that is sometimes elicited by highly realistic human-like artifacts.

1.1: The Uncanny Valley Theory

Mori (1970/2012) conceived of the uncanny valley theory by extrapolating from his observations of people’s responses to the types of prosthetic limbs that were available during the late 1960s – early 1970s (99-100). He thinks, since these early prosthetics were quite eerie, that a robot that looks and moves a lot like a human would seem just as eerie, or perhaps even more so. He calls this phenomenon **bukimi no tani**, which translates to “[the] eerie valley” or “[the] valley of eeriness.” In English translations this term is always rendered as “the uncanny valley.”

The phenomenon that Mori’s theory describes seems counter-intuitive. For why should something that resembles a human being very closely cause someone to experience eeriness? After all, one usually feels affinity and familiarity for one’s conspecifics. Other people are often quite literally familiar – e.g. friends and family – and people typically experience a sense of compassion, sympathy, or camaraderie for one another, and even for humanity in general, i.e. “fellow-feeling.” Considering this, it’s

---

9 I shall return to Misselhorn’s theory in full detail in Section 5.
10 Note that Borody (2011) suggests that Mori (1970/2012) may also have been inspired by Buddhist philosophy. See his paper for discussion on this matter, which I unfortunately do not have the space to consider here.
11 Note that Japanese, unlike English, has no articles (“the,” “a” or “an”). “The uncanny valley” is the popular translation, and even though I think Mori’s term is best translated this way, the uncanny valley may not have a single cause, but a set of related causes. That is, it may be the case that, as Karl F. MacDorman (quoted in Sofge 2010) suggests, there is more than one uncanny valley.
reasonable to think that a very realistic human-like artifact ought to elicit a positive emotional response just like real humans, or even other human-like artifacts such as great paintings and sculptures which depict the human form, often do.\textsuperscript{12} However, Mori (1970/2012) proposes that this does not happen with some human-like artifacts, \textit{viz.} those which are very close in appearance and movement to a real human.\textsuperscript{13}

It comes as no surprise, then, that there is such a growing interest in Mori’s (1970/2012) concept of the uncanny valley. For example, a Google search for “uncanny valley” currently turns up about 3.3 million results. Google Scholar alone returns nearly 34,000 hits as of this paper’s publication.\textsuperscript{14} Of course, this bloom of interest is also attributable to advances in robotics that allow for the construction of highly realistic humanoid machines. This engineering achievement has only become possible within recent years, for as historian Minsoo Kang (2011) points out, in 1970 “no robot existed whose resemblance to a human being was so perfect that one could verify whether it did pull itself out of the uncanny valley” (47).\textsuperscript{15} The uncanny valley theory has also gained interest within the field of computer graphics (CG) where, according to Bartneck, Kanda, Ishiguro & Hagita (2009, 269), it was discussed before it became a prominent idea within robotics.\textsuperscript{16} Those working in CG to design characters in films and video games want the

\footnotesize
\par
\textsuperscript{12} Note that these kinds of human-like artifacts are rather different from the kind that I am concerned with.
\textsuperscript{13} Mori (1970/2012) certainly is not suggesting that all human-like artifacts elicit eeriness. For example, for Mori a \textit{bunraku} puppet – used for performances in traditional Japanese theatre – or a toy robot, are both “familiar” and elicit a positive emotional response. Some of the artificial limbs he mentions are not made with silicon skin, but crafted out of wood. These hands are not eerie at all, but are actually quite beautiful.
\textsuperscript{14} This is a staggering increase from just 364 entries as of August 2009, as reported by Bartneck \textit{et al} (2009, 269).
\textsuperscript{15} Note, however, that Kang (2011) discusses automata – predecessors of robots – which were so life-like that they could appear to eat, or breathe (\textit{cf.} Chs. 4-5 especially).
\textsuperscript{16} Mori’s paper languished in obscurity until it was translated into English in 2005 by MacDorman and Takeshi Minato. MacDorman and his colleagues’ initial research also helped spark a growing interest in the
audience to respond to them positively. Specifically, they don’t want the audience to be
distracted from the intended emotional experience of the narrative by the strange
movements and appearance of the characters.17 As I discuss below, the uncanny valley –
both the theory and the phenomenon – concerns robot engineers and android scientists in
a similar way.

1.2: Research Approaches in Android Science

When it comes to creating human-like artifacts such as robots and androids, a
typical engineering goal is to design machines with which humans can interact
comfortably. Determining how to do this is an especially important concern in Human-
Robot Interaction (HRI) and Android Science research. Android science is, in the words
of self-described android scientist Hiroshi Ishiguro (2006; 2007), a “cross-
interdisciplinary framework” that incorporates complementary research approaches from
robotics and cognitive science. On Ishiguro’s (2007) conception of android science, “the
approach from robotics tries to build very humanlike robots based on knowledge from
cognitive science [and] the approach from cognitive science uses the robot to verify
hypotheses for understanding humans” (2007, 105).

Some of the results of android science research may prove to be essential for
designing social robots, i.e. robots that interact with humans in social situations. These

---

17 The most often cited examples of this are the characters in Robert Zemeckis’ films Beowulf (2007) and
The Polar Express (2004). Cf. Bartneck et al (2009); Misselhorn (2009); Looser & Wheatley (2010); and
Burleigh et al (2013) for discussion.
types of machines – which psychologist Shelly Turkle (2004) refers to as “relational artifacts”18 – can provide companionship and cognitive stimulation in senior care centres, work as assistants in classrooms, and even take on roles in the household. A contemporary example is Paro, a companion robot which resembles a baby Harp Seal, but which interacts with humans using human-like facial expressions and vocal cues.19 Of course, another research concern of android science is: does the uncanny valley phenomenon really exist? If it does exist, what are its causes? What constitutes an instance of the uncanny valley phenomenon? How is the uncanny valley phenomenon similar to, or different from, other experiences of eeriness? Investigating these sorts of questions is one of the tasks for which android scientists can use robots to learn more about the human mind.20

The research concerns discussed above – determining how to design human-like artifacts so that humans interact with them comfortably, and ascertaining the cause(s) of the uncanny valley phenomenon – constitute two different sorts of problems that android science research contends with. The approaches toward these problems are schematized below so as to provide a clear picture of the goals of android science research. The first of these problems is what I call the “problem of design.”

The problem of design prompts researchers and designers to ask: how do we create human-like artifacts that do not elicit the sense of eeriness that Mori thinks is

18 I shall keep to using “human-like artifact(s)” in what follows. Many of the examples I look at count as relational artifacts, i.e. robots and androids. Other examples I mention include photographs, mannequins, automata, etc. Whether these are “relational artifacts” in the sense Turkle (2004) has in mind is less clear, and is a matter for a separate investigation.
19 Cf. Taggart, Turkle & Kidd (2005) for remarks on the uses of Paro in senior care facilities.
20 This is the approach that is taken, for instance, by researchers such as Ishiguro (2006; 2007), and MacDorman & Ishiguro (2006).
attributable to their appearance and movement? Not only that, how do we design these artifacts so that they don’t just avoid a negative response from a subject, but instead elicit a positive one? One answer is that the uncanny valley phenomenon can be avoided by designing robots to appear less human-like and more robotic. This is the “escape by design,” a piece of advice Mori (1970/2012) offers robot engineers in his seminal paper (100). Others suggest the results of android science and related research can be used to design artifacts that are highly anthropomorphic in appearance and movement, but which don’t elicit an uncanny response. In fact some researchers suggest, contra Mori, that the uncanny valley should not hold back designers from creating highly realistic human-like artifacts; this applies to robots and androids as well as to animated or CG characters. For example, cf. Bartneck et al (2009, 275) and Burleigh et al (2013, 770-71).

The problem of design is not the main focus here, although answering questions related to the second problem might have implications for the problem of design. As such, the problem I am about to discuss has a sort of priority over the problem of design. The second problem is what I call the “problem of eeriness.” It prompts researchers to ask questions like those mentioned earlier: why should human-like artifacts elicit a sense of eeriness in the first place? In other words, if there is an uncanny valley phenomenon, then what causes it, and how is it similar to or different from other instances when a subject experiences a sense of eeriness?

There has been no shortage of answers to questions concerning the problem of eeriness. Those that I consider in this investigation, which I mentioned briefly in the introductory section, are: the “Mortality Salience hypothesis,” i.e. human-like artifacts such as robots elicit eeriness because they are reminders of death (cf. MacDorman 2005);
what I call the “Categorical Anomaly hypothesis,” *i.e.* human-like artifacts violate psychological categories that we use to understand the world, resulting in a psychological tension which may explain a subject’s sense of eeriness (cf. Kang 2011; Burleigh *et al* 2013; Yamada, Kawabe & Ihaya 2013); and the “Imaginative Perception hypothesis,” *i.e.* subjects empathize with inanimate objects by “imaginatively perceiving” characteristics of human minds such the experience of emotions, and this misfires in cases of uncanny artifacts like robots (cf. Misselhorn 2009). This last idea is similar in some ways to Gray’s & Wegner’s (2012) hypothesis that mind perception – especially the perception of the capacity to experience – plays a role in the uncanny valley phenomenon. Yet another similar theory appears in historical psychology, where Ernst Jentsch (1906/2008) and Sigmund Freud (1919/2003) considered the possibility that the perception of agency and animacy – wherein one or both of these is somehow out of place given a subject’s typical experience with such attributes – can elicit eeriness. The reader ought to these similarities in mind, for they shall prove to be important later.

Before moving on, I must point out that not everyone who studies human emotional responses toward robots identifies him/herself as an android scientist. Some researchers approach questions in HRI research without much concern for explaining the

---

21 Gray & Wegner (2012) concern themselves with two such attributions: agency – the ability to do things; and experience – the capacity to feel things, sense things, *etc.* (cf. Gray *et al* 2007). As I discuss later (see Section 5), Gray & Wegner (2012) believe attributions of experience play a greater role in the uncanny valley phenomenon than attributions of agency. Making attributions of mind is also implied by Misselhorn (2009), who thinks the perception of imaginary emotions, mental states, *etc.* might help to explain the Uncanny Valley.

22 I take care not to misrepresent Freud. He does not discuss perceptual agency or animacy *per se*, but some of the cases of uncanniness that he discusses involve what I argue – upon careful exegesis – are what modern psychologists would call misattributions of agency/animacy. In fact, Freud is only concerned with these insofar as he is responding to Jentsch (1906/2008), for whom perceptions of agency and animacy seem quite important to consider. For as I explain in Section 2, Freud’s theory of the uncanny amounts to a psychoanalytic re-structuring of Jentsch’s theory.
uncanny valley phenomenon either. Instead, they focus on what emotions besides eeriness that human-like artifacts elicit, as well as the reasons why (e.g. Turkle 2004; Taggart et al 2005; Bartneck, Rosalia, Menges & Deckers 2005; Rosenthal-von der Pütten, Krämer, Hoffman, Sobieraj & Eimler 2013; Rosenthal-von der Pütten, Schulte, Eimler, Hoffman, Sobieraj, Maderwald, Krämer & Brand 2013; Von der Pütten, Kramer, Gratch & Kang 2010). Misselhorn (2009) combines elements from both of the above-described approaches when she asks why human-like artifacts seem eerie in addition to how a subject’s positive emotional responses – i.e. on her view, empathy; or alternatively, “familiarity” or “affinity” (Mori 1970/2005) and “likeability” (Bartneck et al 2009; Yamada, Kawabe & Ihaya 2013) – are related to negative responses.

Even though not all of these researchers identify as android scientists, they still employ the research approach that Ishiguro (2006; 2007) has in mind for android science. Namely, they use an interdisciplinary approach that incorporates elements from robotics with elements from cognitive science such as psychology, neurology, computer science – and sometimes philosophy – with the aim of understanding more about both the human mind and creating machines that facilitate comfortable HRI. Insofar as the research I discuss in this investigation takes the above-described approach, and insofar as this research addresses questions related to either the problem of design or the problem of eeriness, it can be thought of as android science research, even though some of these researchers do not employ the term as Ishiguro does. Android scientists usually do not make a distinction between the problems of eeriness and of design as I have done either. However, this schematic is a simple way to think of the general research aims of android

23 Henceforth Rosenthal-von der Pütten et al 2013a & 2013b, respectively.
science. I should note further that there is sometimes an overlap between the problem of eeriness and the problem of design, as there is, for example, in Misselhorn’s (2009) research. From here onward, I’ll be primarily concerned with the problem of eeriness.

1.3: Does the Uncanny Valley Phenomenon Exist?

Earlier I said that addressing the problem of eeriness entails ascertaining whether the uncanny valley is a real phenomenon. It is better to answer this question before moving on, for it provides an opportunity to sketch out some of the conceptual issues that research on the uncanny valley phenomenon faces. Owing to the fact that this research only began recently, whether the uncanny valley really exists continues to be a subject of debate. Some observers, such as science journalist Erik Sofge (2010), have been quick to decry the uncanny valley as a “theory in disarray.” For despite its popularity, there is not much consensus on whether Mori’s theoretical model is accurate or even whether it describes anything real. Consider Sofge’s remarks:

Despite its fame, or because of it, the uncanny valley is one of the most misunderstood and untested theories in robotics. […] [W]e expected to spend weeks sifting through an exhaustive supply of data related to the uncanny valley – data that anchors the pervasive, but only loosely quantified sense of dread associated with robots. Instead, we found a theory in disarray. The uncanny valley is both surprisingly complex and, as a shorthand for anything related to robots, nearly useless (2010, n. pg.)

Sofge’s assessment is a little premature, as is clear from some of the research results I examine further on. However, if by “useless” he means “incorrect,” and if he is right in claiming that Mori’s (1970/2012) theory is indeed incorrect, then perhaps it should be

---

24 Mori’s article was first translated into English in 2005.
discarded. Indeed, some empirical evidence suggests Mori’s theoretical model is questionable. A good example of this comes from the results of a study conducted by Bartneck et al (2009), who used a human-like robot made in the likeness of Hiroshi Ishiguro, viz. the Geminoid HI-1, and compared subjects’ responses toward the robot to their responses toward Ishiguro himself, who was used as a control.

Specifically, Bartneck et al (2009) tested three predictions of the uncanny valley theory: (1) robots which are distinguishable from humans are less likeable \(^{25}\) than humans; (2) a moving robot will be liked differently \(i.e.\) less so than a still robot; and (3), various degrees of anthropomorphism result is varying degrees of likeability. \(^{26}\) The 32 subjects – male and female university students in Japan – interacted with either the robot or the human for a few minutes each, and were then asked to complete questionnaires to assess the robot/human depending on which condition they were assigned to. The results, Bartneck et al (2009) conclude, do not confirm the three hypotheses that their study tested. The subjects were able to distinguish the robot from Ishiguro, but the robot was not rated significantly less likeable than he was. The authors also failed to find a significant difference in the likability of the robot when it was in motion compared to when it was still. Finally, ratings of human-likeness and likeability were not significantly affected by the robot’s visor. Bartneck et al (2009) further conclude that the uncanny

\(^{25}\) As I discuss later, Bartneck et al (2009) chose to translate shinwakan as “likeability” rather than “familiarity.” How to translate this word is one of the conceptual issues research on the uncanny valley faces.

\(^{26}\) Since only one robot was used, they manipulated anthropomorphism by covering the robot’s eyes with a metallic visor that featured blinking LED lights. The eyes and the upper part of the face are integral for non-verbal communication cues. Moreover, Tinwell, Grimshaw, Nabi & Williams (2011) found that the emotional expressions of CG faces with limited motion in the eyes and upper part of the face were perceived as more eerie than those with full motion, like a real human. Other studies have also shown that the eyes are most salient when it comes to perceiving that something is animate (Looser & Wheatley 2010). So, covering the eyes obscures important anthropomorphic features.
valley phenomenon is probably too complex to explain by appealing to the factors of human-like appearance and movement.

The above study used one of the most human-like robots currently in existence. This may explain why the robot was not liked significantly less than Ishiguro, *i.e.* the Geminoid HI-1 was (rather conveniently) able to pull itself out of the uncanny valley. Or, perhaps the robot was perceived by the subjects as more of a curiosity than something that is creepy, which is consistent with Kang’s (2011) observation that automata – the predecessors of robots and androids – have been perceived as both eerie contraptions and wonderful curiosities throughout modern history. Whatever the case, the conclusions of this particular study do not show that eeriness doesn’t occur during HRI, as is clear from some of the studies I examine later in this investigation, nor was demonstrating such the objective of the aforementioned study. But Bartneck *et al* (2009) are probably right to conclude that Mori’s (1970/2012) graph does not represent the uncanny valley phenomenon exactly. This is not surprising since Mori had no empirical data, only anecdotal observations he had made during his career as a roboticist, when he conceived of the uncanny valley theory.

Moving on, the results of other empirical studies support the existence of the uncanny valley phenomenon. In light of these results I think that, rather than being discarded, which is what Sofge (2010) seems to imply, Mori’s (1970/2012) theory can be revised. For example, Gray & Wegner (2012) ran a study consisting of several sets of trials, the first of which examined whether or not the uncanny valley phenomenon exists, as well as what kinds of mind perception – *i.e.* perceived capacities for experience and
agency (cf. Gray et al 2007) – might play a role in it. In this experiment the subjects watched videos of Kaspar, a robot used for Autism research and therapy purposes. 125 subjects viewed videos of Kaspar moving about in one of two conditions: the human-like condition videos depict Kaspar from the front, while the mechanical condition videos depict it from behind, revealing its wires and circuits. The subjects were asked to rate what extent Kaspar made them feel “uneasy,” “unnerved,” and “creeped out” on 5-point Likert scales. Subjects rated the mechanical videos as more unnerving than the human-like ones. Their data do not fit exactly with Mori’s (1970/2012) graph of the uncanny valley; however, the subjects did report that they felt more unnerved, uneasy, etc. when they observed the human condition videos. Interestingly, they also made greater attributions of experience to the robot in the human condition. Gray & Wegner (2012) conclude from these results that an uncanny valley-like phenomenon exists. They also suggest that the apparent capacity for experience, which the subjects perceived Kaspar as having, may play some role in the uncanny valley phenomenon. I discuss this further in Section 5.

As the results of these studies show, it is difficult to prove conclusively with just one series of trials where or not the uncanny valley phenomenon is real. However, I think it is the case that a phenomenon like the one that Mori (1970/2012) describes does exist. As I discuss in Section 2 where I examine some of Jentsch’s (1906/2008) and Freud’s (1919/2003) observations about the uncanny, it’s clear that the predecessors of robots, viz. automata, can elicit eeriness. In fact both of these scholars – but especially Jentsch –

27 I examine the results of their other trials in Section 5, where I shall also discuss mind perception in further detail.
28 Cf. Scassellati, Admoni & Matarić (2012) for more on this topic.
pay very close attention to the uncanniness of human-like automata in their essays. So, it’s not unreasonable to claim that robots have been perceived as eerie since before they were commonly called robots. Moreover, the materials I review in the rest of this investigation clearly show that an uncanny valley-like phenomenon does exist, and that empirical research has come quite close to explaining it. Current research simply lacks an adequate theoretical framework for this purpose.

The problem of determining the cause(s) of the uncanny valley phenomenon is compounded by the different measures used in empirical research. Bartneck et al (2009) chose to measure the robot’s “likeability,” but MacDorman & Ishiguro (2006) measured “eeriness” in one of their empirical studies. They found what Kang (2011) describes as an “eerie peak.” For instance, in one of MacDorman’s & Ishiguro’s (2006) experiments, subjects were shown images of a robot “morphed” into either a human female or a human male. Here, “morphing” refers to a technique which gradually changes one static image into another. The result is a series of images, each a few degrees of human-likeness apart from one another on a continuum of anthropomorphism, from fully robotic to fully human. The subjects’ reported levels of eeriness increased when the images they were shown were close to, but not fully, human-like. So, whether one finds an “uncanny

---

29 Automata, however, are not always eerie; it’s also apparent that cultural and historical background can influence how human-like artifacts are perceived. For instance, Kang (2011) provides illuminating examples of the variety of different attitudes people have shown toward automata throughout history, from “wonders” in antiquity, to metaphors for the human being and the universe as a divinely crafted machine during the enlightenment. Kang shows how many of these attitudes are often either the result of the historical-cultural context, or an expression of them, yet he nonetheless acknowledges the power of the automaton – and now, the robot – to elicit a sense of both fright and eeriness (cf. Ch. 1-2 and 5-7 especially).

30 The word “Robot” began to come into popular usage thanks to Karel Čapek’s play R.U.R. (Rossum’s Universal Robots) (1921), in which the word was first introduced.

31 Morphing is particularly useful since many laboratories cannot afford the type of robots necessary to explore the uncanny valley phenomenon.
valley” or an “eerie peak” in their data depends upon what emotional responses those data are measures of. The list of possible measures for the uncanny valley phenomenon also include the aforementioned “familiarity” (Mori 1970/2005), “affinity” (Mori 1970/2012), “rapport” (Kang 2011), “likeability” (Bartneck et al 2009; Yamada et al 2013), “unnervingness” (Gray & Wegner 2012), and “pleasantness” (Gray & Wegner 2012; Burleigh et al 2013; note that Burleigh et al also measure “eeriness”), et cetera.

Of course, even when what is being measured remains consistent across a variety of studies, some research results still seem to speak against the existence of the uncanny valley phenomenon. This data comes from studies concerning how an artifact’s level of human-likeness is related to a subject’s sense of eeriness. Recall that Mori’s (1970/2012) theory predicts that the relationship between eeriness and anthropomorphism is non-linear. When MacDorman et al (2009b) used CG faces to determine the relation between a subject’s emotional response to an artifact and the artifact’s level of human-likeness – manipulating the level of realism and the saliency of atypical features – they found a linear relationship between positive and negative affective responses and eeriness. Burleigh et al (2013) obtained similar results in several of their trials. They suggest that the conditions necessary for the uncanny valley phenomenon to occur don’t simply include where an artifact lies on a continuum of human-likeness. Rather, they suggest that the uncanny valley occurs when an artifact such as a robot or a CG character possesses categorically anomalous features, i.e. a human-like artifact can possess features/attributes which belong to distinct ontological categories, e.g. an inanimate object that appears to possess animacy. I discuss this hypothesis in Section 4. For now, I think it’s reasonable to claim that highly realistic human-like artifacts do sometimes elicit
a sense of eeriness in certain individuals, and this is precisely the phenomenon that Mori (1970/2012) describes in his paper. In other words, it is reasonable to say that some sort of uncanny valley phenomenon does exist. What is necessary is an improved theoretical framework – *i.e.* a “conceptually overhauled” uncanny valley theory – that can explain the phenomenon by integrating the various research results obtained from android science on the uncanny valley.

1.4: Putting “the Uncanny” in the Uncanny Valley

Misselhorn (2009) suggests a possible solution to the lack of consensus on the existence and cause(s) of the uncanny valley phenomenon, which was briefly mentioned in the introductory section. Specifically, she suggests that researchers ought to adopt a more philosophical rather than a strictly empirical approach, toward the end of re-examining the conceptual foundations of the uncanny valley theory. She is the first to articulate such a view, writing that “philosophical concepts and approaches allow for a clearer grasp of the problem [of the uncanny valley], and provide the tools for a framework for a solution which can integrate the results of the empirical studies” (347). I agree, and I even take some of these suggestions a step further than Misselhorn does. What I mean by “philosophical” will become fully clear as this investigation unfolds, but I shall provide a sense of the sort of philosophically grounded approach that I have in mind presently.

The cause(s) of the uncanny valley phenomenon are a mystery to be sure; however, the problem of determining what they are is, according to Misselhorn (2009) compounded by further problems which might be addressed by re-evaluating the
concepts and research methods used in uncanny valley research. She thinks one of these problems, which I also briefly mentioned earlier, is as follows: to explain the uncanny valley phenomenon, researchers must not only account for why human-like artifacts are sometimes perceived to be eerie, but also why a subject’s emotional response changes from positive to negative by examining how the positive and negative emotional responses are related to one another (347-49); other than Misselhorn, no one has made this effort, and empirical studies on the phenomenon are generally not this broad in scope. Misselhorn’s contribution to addressing this problem is her use of “empathy” in place of “familiarity” or “affinity,” the most often encountered translations of shinwakan, the word Mori (1970/2005; 1970/2012) uses. She trades “empathy” with “familiarity” because her explanation for the uncanny valley phenomenon turns on the “imaginative perception” of emotions of human-like artifacts. I’ll discuss this idea further on.

For now, there is a second problem in uncanny valley research: android scientists lack conceptions of both “the uncanny” and “eeriness” that are amenable to empirical research, and researchers cannot explain why human-like artifacts elicit eeriness if they do not know what kind of emotion it is. As MacDorman & Ishiguro (2006) note, it’s difficult to use “uncanny” as a measure in empirical trails because it is not an often-used word, and its meaning may be somewhat ambiguous to experimental subjects. This is why some researchers measure other things like “likeability” (e.g. Bartneck et al 2009; Yamada et al 2013), or “eeriness” (e.g. MacDorman & Ishiguro 2006; Ho, MacDorman & Pramono 2008; Burleigh et al 2013). To explain the uncanny valley phenomenon, I shall have to explain what type of emotion eeriness is, as well as how it relates to other
similar emotions. However, eeriness is not a well-researched emotion. This is an additional reason so many measures besides eeriness are used in experimental studies on the uncanny valley phenomenon. In fact, what recent research there is on eeriness comes from research related to the uncanny valley phenomenon. Some researchers, however, have taken steps toward addressing the above-mentioned problem (cf. Ho et al 2008; Burleigh et al 2013), and their results show that eeriness is related to fear. I think that eeriness is a type of anxiety, which is a species of fear characterized by uncertainty or ambiguity about things which affect one’s self-identity, or even one’s existence (cf. Lazarus 1991; Lazarus & Lazarus 1994). I’ll argue that eeriness is best understood in this way throughout Sections 2 and 3.

There is a third problem that highlights the need for a more philosophically oriented approach, viz. the uncanny valley phenomenon may not have a single cause. Rather, as Ho et al (2008) conclude, it could be “a nexus of phenomena with disparate causes” (175). Similar conclusions have been reached by others as well (e.g. the aforementioned conclusions of Bartneck et al 2009, 274-76; also cf. MacDorman, quoted in Sofge 2010). If the uncanny valley phenomenon turns out to have a set of distinct but related causes, then the theoretical framework that explains it must also provide a way of discussing the causes and constituents that different cases of eeriness have in common, instead of providing a single explanation for a single phenomenon. Of course it should also account for any isomorphism or extreme variation in the phenomenology of eeriness in these various cases. In other words, it must be able to identify what Ludwig
Wittgenstein (1953/1963) calls “family resemblance.”\textsuperscript{32} Providing a theory of the uncanny valley that also explains the so-called “family resemblance” of these different uncanny valley phenomena should, for that matter, explain other instances of eeriness too. For these other cases just present the problem of eeriness once again with a subject matter other than human-like artifacts; however, this is a matter for another investigation, one which can take place after I’ve articulated a theory of eeriness for android science research.

1.5: Empathy and the Uncanny Valley Phenomenon

I shall deal with the first of the aforementioned three problems below. Recall that Misselhorn (2009) replaces “familiarity” with “empathy” in order to explain how a subject’s positive emotional response toward human-like artifacts turns to eeriness. As I show below, she is primarily concerned with affective empathy, \textit{i.e.} the capacity to experience emotions consonant with another person. These empathic phenomena include sympathy, compassion, emotional contagion, \textit{etc.}\textsuperscript{33} The ability to elicit an empathic response may even prove to be important for the success of social robots. For instance, according to Lee (2006) the capacity to show emotion, and to respond empathically to the emotions of others, shall be an important design feature for the successful social machines of the future. These characteristics can ensure that people develop positive relationships with their robotic helpers, and that they remain unperturbed by them. Similarly, Scassellati (2002) suggests that robot engineers provide autonomous machines

\textsuperscript{32} Wittgenstein’s (1953/1963) concept of “family resemblance” applies to what I am concerned with: members of the family of uncanny experiences. I’ll therefore borrow the term “family resemblance.”

\textsuperscript{33} \textit{Cf.} Maibom (2012) for an in-depth discussion of these empathic abilities.
– such as the social robots or relational artifacts discussed in Section 1.2 – with a robotic version of “theory of mind,” i.e. the ability to understand the thoughts, beliefs and desires of other people. He thinks that providing social robots with a rudimentary theory of mind – sometimes called “cognitive empathy” or “folk psychology” – will prove to be essential for comfortable HRI, and will allow such social robots to perform their duties smoothly and effectively.

MacDorman et al (2009b), as I noted in the introductory section, suggest that cognitive empathy may play some role in the uncanny valley phenomenon; they do not say much concerning affective empathy. Actually, they are not this specific either; however, it is obvious from their discussion that cognitive empathy – i.e. the ability to understand the thoughts, beliefs, desires and goals of other minds – is what they have in mind (696-97). I shall further discuss MacDorman et al’s (2009b) speculations concerning cognitive empathy’s role in the phenomenon in Section 5. For now, let us resume the discussion of Misselhorn (2009), who as I mentioned is concerned with affective empathy. She draws attention to some interesting examples of empathy with human-like artifacts, viz. the compassion that subjects show toward human-like artifacts when they see them being mistreated. Here Misselhorn (2009) has in mind the results of Milgram-style experiments using human-like artifacts; specifically, those obtained by Bartneck et al (2005) and Slater et al (2006).
In the studies mentioned above, Milgram-style experimental designs were used to determine whether subjects would act any more or less cruel\textsuperscript{34} toward robots and digital characters, respectively, than toward other humans. Bartneck \textit{et al} (2005) found that all 20 of their participants administered the maximum electric shock to the robot in their experiment (450 volts), whereas only 40% of participants in Milgram’s original experiments continued on at the instructor’s prompting to deliver (what they believed was) a lethal shock. However, subjects still reported that they felt compassion for the robot that they were required to shock. Slater \textit{et al} (2006) had 11 participants interact with/shock a CG\textsuperscript{35} character through a text-only interface, and 23 participants who could see and hear the character while shocking it. The study found that “the participants who saw and heard [the artifact] tended to respond to the situation at the subjective, behavioural and physiological levels as if it were real” (n. pg.), even though they knew the character couldn’t be harmed, albeit their responses were not as pronounced as they would have been if a human replaced the CG character.

Misselhorn (2009) thinks the existence of some capacity for affective empathy with human-like artifacts is evidenced by these empirical results. For in each experiment the subjects expressed compassion for the artifact, despite the fact that they knew it couldn’t really come to any harm. This reaction – even if it is less pronounced than we typically show toward conspecifics – is different than what we typically show toward non-anthropomorphic inanimate artifacts. Take Misselhorn’s (2009) example of this

\textsuperscript{34} I mean “cruel” in the sense that if the artifact really were capable of experiencing pain, discomfort, etc., then we would say any behaviour toward the artifact which causes these things is cruel. Of course, these artifacts can’t really experience pain or discomfort.

\textsuperscript{35} Recall: “computer graphics” (see 1.1). Henceforth, CG refers to “computer generated” characters.
difference: if somebody sees someone else beat a computer, she feels differently than if she were to see someone beating a child (345). I want to add that if someone were to beat a computer – or better yet, my computer – or another piece of property, I could certainly show some emotion, e.g. anger at my property being damaged, or grief over the loss of something I cherish, that I have become attached to, or that I require for work etc. However in cases like these, I show no sympathy for the computer, as I would for e.g. a child who is being beaten. It is conceivable that someone would show empathy or sympathy for me over the destruction of my computer. But no one feels anything for the artifact.

The subjects in the above experiments, on the other hand, had no specific attachments to the artificial agent that they were required to shock, i.e. no one was asked to “harm” their own robot or digital character, say by breaking it or deleting it, respectively. In other words the subjects weren’t experiencing distress because their own or someone else’s property was being harmed. Instead, the emotional reactions that the subjects showed are probably due to the anthropomorphism of the artificial agents; these agents apparently convinced or fooled the subjects in some way or another that they were capable of being harmed even though the subjects knew this isn’t so.36

36 I should note that human beings demonstrate varying levels of compassion for other humans, never mind artifacts. One is probably able to empathize or sympathize better with one’s relatives, for instance, just as one can be expected to feel bad if someone smashes their favorite domestic robot, but not if they are required to ‘hurt’ or ‘kill’ the antagonist of their favorite video game. Also note that even if humans have, on the whole, a greater ability to show compassion for other humans, we have nevertheless demonstrated for ourselves throughout recorded history just how easy it is for us to hurt and kill each other. The variance between the level of compassion shown for a human-like artifact versus the level shown for a real human is therefore no reason to think that these particular instances of compassion aren’t really compassion, or that they are insignificant.
In other words the subjects discussed above demonstrated compassion; they sympathized with the artifacts they had to “shock.” As I have mentioned previously, sympathy – like empathy – entails experiencing emotions that are consonant with another person’s; moreover, sympathy typically involves concern for the welfare of the other person. For example, when I sympathize with someone, I feel bad because they feel bad; my feeling bad is for the other person, and because her situation has an impact on her welfare (cf. Maibom 2012, 354-56). In the above examples, the subjects showed compassion, ostensibly because they were somehow concerned about the welfare of the robot/digital character, even though no such concerns were warranted.

More recent research has yielded results similar to those of Bartneck et al (2005) and Slater et al (2006). For example, Rosenthal-von der Pütten et al (2013a) ran a behavioural study that examined subjects’ self-reported responses and physiological reactions to videos depicting toy robots, humans, and inanimate objects being mistreated. In these studies, however, the subjects weren’t asked to mistreat the artifacts themselves; rather, they viewed videos of robots, inanimate objects and humans being mistreated by another human. As in Bartneck et al’s (2005) and Slater et al’s (2006) studies, Rosenthal-von der Pütten et al (2013a) found that subjects showed empathic concern – in other words, sympathy – for the artifact being “tortured,” albeit at lower levels than subjects typically demonstrate for human beings.

37 Note that this is not the sort of welfare usually associated with inanimate objects, e.g. to ensure the welfare of proper by making sure it goes un-damaged or un-stolen.

38 Empathic concern is one of the alternate formulations of sympathy. (Cf. Maibom 2012, 254). The term originates with C. Daniel Batson. On his view, empathic concern is an emotional response roughly consonant to that of another who is in distress. The concern is of the welfare of the other. So what Batson calls “empathic concern,” Maibom and other researchers call “sympathy.” Rosenthal von der Pütten et al (2013a; 2013b) employ Batson’s terminology.
Rosenthal-von der Pütten et al (2013b) followed up on the results of their previous study by searching for the neural correlates of empathy with robots. They used fMRI imaging to examine the brains of 14 subjects who viewed short videos of a robot, a human, and an inanimate object (viz. a box) being treated in either an affectionate or violent manner. They found activity in areas of the brain associated with empathic reactions at lower levels for the robot videos than for the human.\(^{39}\) Overall, these results appear to be consistent with those of their earlier study (2013a) as well as with Bartneck et al’s (2005) and Slater et al’s (2006) results. I think all of these research results make it clear that affective empathy with human-like artifacts is possible. The question is: how does this explain the uncanny valley phenomenon?

If I were to attempt to explain the uncanny valley phenomenon with recourse to empathy now, my efforts would be premature and ill-spent. For have not yet worked out what kind of emotion eeriness is. But since the account of the uncanny valley I present later – like Misselhorn’s (2009) – turns on the facticity of empathy with human-like artifacts, I have at least provided some examples from empirical research which show that affective empathy with human-like artifacts is possible. As for how empathy misfires and causes a subject to experience eeriness, Misselhorn thinks the cause is a failure of what she calls “imaginative perception,” which I explain below. I shall limit myself to explaining her theory for the present time; I’ll present my theory of how empathy with human-like artifacts misfires in Section 5, after I have both explored the emotion of eeriness in greater depth and examined additional empirical research results.

\(^{39}\) For a detailed overview of these neural correlates, cf. Iacoboni (2009).
1.6: The “Imaginative Perception” of Emotion

Misselhorn (2009) explains the uncanny valley phenomenon by taking Maibom’s formulation of affective empathy as her starting point. Recall that, according to Maibom (2012, 254-55), the ability to perceive the emotional states of others – and to experience emotions consonant with them – is commonly known within the psychological and philosophical literature as affective empathy. Although Misselhorn (2009) does not specify this in her account of the uncanny valley phenomenon, it is clear that she is primarily concerned with this particular capacity for empathy. I shall explain this below.

On Maibom’s (2007a) theoretical framework, a subject (S) empathizes with another person (O) in circumstance (C) if S feels an emotion (E) for O as a result of S’s belief that O feels E, or as a result of S’s perception of O’s T-ing of E, or because S imagines herself being in C. Since Misselhorn (2009) is concerned with emotions that people perceive these artifacts to have, she alters one of Maibom’s (2007a, 168) conditions for empathy. Since one cannot actually perceive an emotion if there is not any real emotion to perceive – as is the case with human-like artifacts – Misselhorn (2009) ends up with the following formulation:

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 (352).

How does the “imaginative perception of emotion” work? First off, Misselhorn (2009) stipulates that “the imagining […] cannot just be propositional” (352). Here Misselhorn

40 That is, O’s expression of an emotion E (by acting a certain why, or by uttering how she feels, etc.)
means that a subject is not “make-imagining” that an inanimate object’s T-ing of E is a human T-ing of E. In other words, “imaginative perception” doesn’t really involve a subject actively imagining that the human-like artifact has emotions. The reason is that this kind of imagining does not depend on perception in the way that Misselhorn wants to say the imaginative perception of emotion does. She seems to have in mind something that this is reflexive and non-conscious, in the way that a speaker of her native language can recognize linguistic utterances reflexively, without conscious effort (see the discussion in Section 5.2).41 In other words when S “imaginatively perceives” an artifact’s emotions, S is not make-imagining or actively “imagining that” an artifact really has emotions, but S perceives these emotions nonetheless.

The result of imaginative perception is that perceiving one thing – in this case a human-like artifact – can take on the phenomenal characteristics of perceiving another, i.e. a real human being. Misselhorn (2009) explains that “[b]ecause of certain salient similarities” between objects (F) and (G), perceiving an F can trigger the activation of concepts associated with a G. But these concepts cannot affect the representational content of the perception of F. They can, however, have some effect on the “phenomenal feel” – i.e. what it feels like to perceive a thing – of perceiving an F, causing the perception of an F to feel phenomenally similar to perceiving a G. The more the G-related concepts are triggered, the more the “phenomenal feel” of perceiving an F changes to that of perceiving a G. The number of similar features between F and G – as well as their salience and typicality – is what, according to Misselhorn, determines the

---

41 Also cf. Fodor’s (1983) theory of modular input systems.
extent to which the perception of F activates concept G (354). When applied to human-like artifacts, Misselhorn gets the following:

The human-like features M of an inanimate object trigger the concept of a human N, [and] 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, [and] for that reason, seeing the smile of a doll feels like seeing the smile of a human being (354).

This is ultimately how the “imaginative perception of emotion” comes to occur.

As I’ve mentioned numerous times, the failure of imaginative perception is what, on Misselhorn’s (2009) view, can explain the uncanny valley phenomenon. Here, a subject stops empathizing and even feels a sense of eeriness because the salient features of an uncanny artifact (M) triggers the concepts associated with a human (N). But if the human-like features M are not sufficiently human-like to trigger a full application of these N-related concepts, the result is a sense of eeriness:

[B]ecause 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 (2009, 356-57).

A “gestalt switch” occurs when the apparent perception of one thing switches to the perception of another. Think here of the duck-rabbit illusion (cf. Wittgenstein 1953/1963), which flips back and forth from a duck to a rabbit when an observer looks at it for an extended length of time. The same thing happens when one looks at a Necker cube. Its orientation in three dimensions appears to change even though it is a static, two-
dimensional image. Thus the “baseline explanation” for the UV phenomenon on Misselhorn’s (2009, 356) view, is that the number and typicality of saliencies associated with N – which are similar to those of M – cause the “mental oscillation” mentioned above, which is like a very rapid, continuously repeating gestalt switch-like phenomenon. According Misselhorn, since perceiving a human has a distinctive “phenomenal feel” (356-57), then when imaginative perception fails and the above described gestalt switch-like oscillation occurs, the phenomenology of perceiving the human-like artifact becomes very eerie.  

Unlike Misselhorn (2009), I don’t limit myself to affective empathy when explaining the uncanny valley phenomenon. The account that I present in Section 5 includes cognitive empathy as well, following the suggestion of MacDorman et al (2009), the research of Gray & Wegner (2012), and similar ideas found in the work of Jentsch (1906/2008). However, I think Misselhorn (2009) is certainly correct that affective empathy plays some role in eeriness with human-like artifacts, even if it is not through the failure of imaginative perception.

For instance, it seems reasonable to hypothesize that a subject’s emotional responses to a human-like artifact could be mediated by the emotions – and other mental states – that she perceives a robot to have because of its level of realism. If the features the artifact uses to express these apparent emotions are somehow faulty, or if they are

---

42 Here is one of the only points at which Misselhorn (2009) notes that the distinctive phenomenal feel of perceiving another human is attributable – in part – to perceptions of animacy. As I discuss in Sections 2 and 3, the perception of animacy probably plays a very large role in the uncanny valley phenomenon. But Misselhorn seems to avoid investigating this possibility in any substantial detail. For instance, recall that she talks about “empathy with inanimate objects” even though she is obviously concerned with human-like inanimate artifacts specifically, which can appear to move as a human does if they are switched on.
mismatched (cf. MacDorman, quoted in Sofge 2010), then the result might be a sense of eeriness. There is empirical support for this idea. For example, in a study on the uncanny valley phenomenon, Tinwell, Grimshaw, Nabi & Williams (2011) used videos of CG and human faces which displayed the six facial emotional displays common to all human cultures: fear, anger, disgust, surprise, sadness and happiness, along with a neutral face.\textsuperscript{43} The subjects’ responses to the real faces, CG faces with full motion, and CG faces with limited motion in the upper face were measured. Not only did Tinwell \textit{et al}’s (2011) subjects indicate that they preferred the appearance of the real faces compared to the virtual faces, but in addition, they preferred the limited-motion faces even less than virtual faces with full facial expressions.\textsuperscript{44} Moreover, subjects were more likely to confuse one emotional expression with another (\textit{e.g.} anger with surprise, or vice versa) where movement of the upper part of the CG face was limited. This suggests that some particular features of these faces – in this case expressions in the upper part of the face which humans use in the nonverbal communication of emotional states – may have an effect on whether humans perceive them as eerie or uncanny (745-48).\textsuperscript{45}

\textbf{1.7: Empathy and the Uncanny Valley Theory}

The above empirical data and philosophical theorizing seem to make a good case for the proposal that empathy is in some way a factor in the uncanny valley phenomenon.

\textsuperscript{43} The subjects were 129 university students (all male) studying video games design, software engineering, and other related fields.
\textsuperscript{44} Note that motion in the upper part of the face carries a lot of information as to the emotional expression a subject is communicating. This has been noted since Darwin (1872/1955).
\textsuperscript{45} Of course there may be other salient features which cause us to feel ill at ease when interacting with human-like robots or digital characters besides simply the expressions they are able (or indeed, unable) to make with their faces. Likewise if the study had included women or focused exclusively on women, different results may have been had.
But there is also a good *prima facie* reason to consider replacing “familiarity” or “affinity”\(^{46}\) with “empathy” in the conceptual framework of the uncanny valley theory: empathy may simply be a more useful way of describing the subject’s positive emotional response in the uncanny valley phenomenon than, say, “familiarity.” As Bartneck *et al* (2009) maintain, the word that Mori (1970/2012) uses which is often translated as “familiarity” or “affinity” – *shinwakan* – is not a commonly used word in Japanese and is difficult to translate properly. Consider Bartneck’s and his colleagues’ summary of this problem:

In his original paper, Mori plots human-likeness against [*shinwakan*] which has previously been translated as “familiarity” [Mori 1970/2005] Familiarity depends on previous experiences and is therefore likely to change over time […] We questioned whether Mori’s shinwa-kan concept might have been “lost in translation” and in consultation with several Japanese linguists, we discovered that “shinwa-kan” is not a commonly used word. It does not appear in any dictionaries and hence it does not have a direct equivalent in English. The best approach is to look at its components “shinwa” and “kan” separately. The *Daijirin Dictionary* (second edition) defines shinwa as “mutually be frien” or “having similar mind.” Kan is being translated as “the sense of.” Given the different structure of Japanese and English, not perfect translation is possible [sic], but “familiarity” appears to be the less suitable translation compared to “affinity” and in particular to “likeability.” After an extensive discussion with native English and Japanese speakers we therefore decided to translate “shinwa-kan” as “likeability” (Bartneck *et al* 2009, 270).

In addition to likeability, “rapport” has been proposed as translation of *shinwakan*, however as Kang (2011, 47-51) points out, this can be just as ambiguous as “familiarity”; moreover, he further points out that what is unfamiliar is not always eerie. Bartneck *et al* (2009) reach the same conclusion.

\(^{46}\) Note that in the first translation of Mori’s (1970/2005) paper, “familiarity” was used. In a subsequent – and properly naturalized – translation, “affinity” rather than “familiarity” is the translation of *shinwakan*. 
One of the reasons I think Misselhorn (2009) is wise to trade these translations for “empathy” – understood in the broad sense of “putting oneself in another’s shoes” – is that “empathy” seems to characterize the meaning of shinwakan well, even though shinwakan is not the Japanese word for empathy (i.e. kyōkan). That is, “empathy” is much more precise than “familiarity” for the purposes of constructing a theoretical model to explain the uncanny valley phenomenon. For instance, the sense of “having similar mind” is arguably present in the origins of the word for empathy (Einfühlung, literally “feeling-into”). Of course when a subject empathizes with another person, either affectively – i.e. experiencing emotions consonant with another’s – or cognitively – i.e. representing the thoughts, beliefs and desires of someone else in her own mind – that subject has, in a certain sense, a “similar mind” to the other person. Moreover, to have the sense of being “mutually friendly” with someone is strikingly similar to what is expressed in “fellow-feeling,” which refers to the sympathy and/or compassion we feel for our fellow humans, as I discussed in Section 1.1. By the end of this investigation, I will have shown conclusively that “empathy” is a more helpful concept to employ than either “affinity,” “familiarity,” “rapport,” etc. in the uncanny valley theory.

By now, the reader will have become familiar with the uncanny valley and android science. It is clear from the above discussion that the uncanny valley phenomenon exists, but that explaining it has been complicated by contradictory research results and different research approaches. A possible solution, however, is to focus how empathic abilities misfire with human-like artifacts. These empathic abilities include both affective empathy (Misselhorn 2009) and cognitive empathy (MacDorman et al 2009b).

47 Cf. “Einfühlung,” n., OED Online.
The ways in which these abilities factor into the uncanny valley remains to be determined, but I have provided some additional reasons why “empathy” represents the emotional phenomena characteristic of what Mori (1970/2012) describes with shinwakan, even though this word does not exactly mean “empathy.”

I shall not say anything further about empathy with human-like artifacts as it pertains to understanding the uncanny valley phenomenon for the moment. Otherwise, I would be putting the cart before the horse, so to speak. I’ll return to empathy’s role in the uncanny valley phenomenon in Section 5. For now, a better understanding of the emotion of eeriness is necessary in order to explain the uncanny valley phenomenon. Once I obtain this, I shall be in a much better position to demonstrate how the misfiring of empathy with human-like artifacts can explain the uncanny valley phenomenon. Therefore, over Sections 2 and 3, I shall show that eeriness is an emotion associated with encounters with what I call the “negative uncanny.” I shall also show that eeriness is a type of anxiety. This way the emotions associated with “the uncanny” are made more amenable to the sort of empirical study that is necessary to explain the uncanny valley phenomenon. So let us now venture further into uncanny country, and examine some literature on the uncanny from historical psychology in order to gain a sense of what the psychology of the uncanny is like.
2.0 – Naturalizing the Uncanny

The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents. We live on a placid island of ignorance in the midst of black seas of infinity, and it was not meant that we should voyage far…

– Francis Wayland Thurston, in H.P. Lovecraft’s “The Call of Cthulhu” (1928)

In this Section, I examine the concept of “the uncanny,” in order to make a start at determining how eeriness and the uncanny are related to one another. My objective is to demonstrate that eeriness is elicited during encounters with what I call the “negative uncanny,” i.e. uncanny objects/situations which elicit a negatively valenced emotional response.\(^{48}\) In Section 3, I want to show that eeriness is best understood as a type of anxiety. The present section is where I establish a basis for this proposal.\(^{49}\) I do so by conducting an exegesis of Ernst Jentsch’s (1906/2008) and Sigmund Freud’s (1919/2003) essays concerning the “psychology of the uncanny.”\(^{50}\)

To show that eeriness is best understood as a type of anxiety, a good place to begin might be to consider: (1) what eeriness is like from a subjective standpoint; and (2), what a look at the empirical data can reveal about the emotion of eeriness from an objective standpoint. Arguably most people have had an encounter with something that

\(^{48}\) The authors I review make it seem as though all encounters with the uncanny are negative. However, as I discuss below, this is not the case; some experiences of the uncanny can be positive. Moreover, what follows should not be read as a suggestion that eeriness is the only emotion elicited by encounters with the uncanny. There may be others, but I am choosing to focus on eeriness because that is the emotion which appears in Mori’s (1970/2012) paper (bukimi).

\(^{49}\) Anxiety is often thought of as a type of fear, but it is distinct from many other kinds of fear such as fright, shock or horror in that the hallmark characteristics of anxiety are uncertainty or ambiguity. Fear – or more accurately “fright” – often has a concrete object, whereas anxiety does not. Cf. Lazarus (1991) and Lazarus & Lazarus (1994). See Section 3.

\(^{50}\) As I mentioned in the introductory section, his phrase is Jentsch’s (1906/2008), but I think it also captures the gist of what Freud (1919/2003) sets out to accomplish in his own essay. It also captures the sense of what I am formulating herein, i.e. a psychological theory of eeriness.
they would describe as “uncanny,” and we have all at one time or another experienced a sense of eeriness, so addressing (1) is fairly easy. However as I discussed in Section 1.4, empirical research into these states of mind is not particularly abundant; what data there are is found largely within android science research on the uncanny valley phenomenon. Examining the conclusions in Jentsch’s (1906/2008) and Freud’s (1919/2003) essays concerning the uncanny might therefore provide a starting point for addressing (2).

Connections are sometimes drawn between Mori’s (1970/2012) uncanny valley theory and Jentsch’s (1906/2008) and Freud’s (1919/2003) essays on the uncanny (e.g. in Kang 2011, 22-27; 38; 50-52; also cf. Borody 2011, 6-7). However, elsewhere these connections are not discussed in any great depth, and sometimes authors represent Freud’s and Jentsch’s ideas incorrectly (e.g. Tinwell et al 2011, 741). This provides additional reason to consider what these two have to say about the uncanny; but I shall approach both of their essays with a fresh perspective. This way, I can highlight some similarities between the results of their work and Mori’s (1970/2012), and also, draw out some helpful insights for determining what kind of emotion eeriness is.

---

51 Kang’s (2011) discussion of Jentsch (1906/2008) and Freud (1919/2003) is actually quite well done. Likewise is Borody’s (2011). Others, however, miss important details, or get them completely wrong. For example Tinwell et al (2011) note that Jentsch characterizes the uncanny as a feeling of mental uncertainty, and that Freud “characterized the uncanny as a feeling caused when one cannot detect if an object is animate or inanimate upon encountering objects such as ‘waxwork figures, ingeniously constructed dolls and automata’”(741). As I show in Section 2, this latter idea really belongs to Jentsch (1906/2008), not Freud (1919/2003). This doesn’t really have an effect on Tinwell et al’s (2011) results (discussed in Section 1.3), however since I do not wish to overlook any important insights, it’s better to take to be more precise than Tinwell et al (2011) are.
2.1: What is “the Uncanny”?

For many people, encounters with the uncanny occur when enjoying film, literature and art. Here, uncanniness predominantly occurs in the genres of horror, suspense, thriller, and gothic/neo-gothic. In these genres, feelings of suspense, disgust, fright, horror, and eeriness are elicited in the audience. Such emotions are aroused in one’s everyday experience too. For example, the odd stares of a stranger, or cases when something seems to “go bump in the night,” are often described as “eerie,” “spooky” or “creepy,” and make one feel uneasy. A credulous person might appeal to a paranormal or supernatural explanation for some of these instances, e.g. a ghost or a spirit, or the so-called “evil eye.” Of course not all instances of the so-called “everyday uncanny” are negative (cf. Wittgenstein 1980, in Thornton 2013), as I discuss in Section 2.4. But since I am interested in questions concerning the problem of eeriness, I focus primarily on the negative uncanny, and how eeriness may be related to encounters with it.

One problem with examining the uncanny is contending with the number of ways there are to describe uncanny things. As noted, words like “eerie,” “creepy,” “spooky,” “weird,” etc., all seem connected in one way or another to what can also be described as “uncanny,” in the negative sense. Moreover, some of these terms can be used to describe attributes of an uncanny object/situation, while others can be used to describe how the object/situation makes a subject feel. So, perhaps by focusing on what is talked about

52 Note that Freud (1919/2003) draws attention to the association between the uncanny and the supernatural (148). The word Mori (1970/2012) uses – *bukimi* ("eerie"/"eeriness") – carries a similar meaning.

53 While I say affective responses associated with these are negatively valenced, I do know that people still sometimes enjoy experiencing them. For example, watching a horror/thriller film with friends is quite enjoyable. Others take pride in their affection for weird things and oddities, and these can be a source of interest and curiosity as well as a source of eeriness. Whether one is disturbed by the negative uncanny, or enjoys it, is probably a matter of personal taste.
when we use “uncanny,” “eerie,” “creepy,” and so forth, I can determine what exactly “the uncanny” is, and what affective responses are elicited during encounters with it. An approach like this can be found in early psychology literature that will serve as an ideal starting point.

Jentsch (1906/2008) remarks that the word “uncanny” is noteworthy because it offers a particularly “psychologically correct” description of what it refers to. He continues:

Without a doubt, this word appears to express that someone to whom something ‘uncanny’ happens is not quite ‘at home’ or ‘at ease’ in the situation concerned, that the thing is or at least seems to be foreign to him. In brief, the word suggests that a lack of orientation is bound up with the uncanniness of a thing or incident (217).

Without closer examination, the reasoning behind Jentsch’s assessment risks getting lost in translation. The English word “uncanny” expresses something quite similar to the German unheimlich, the word that Jentsch uses in his seminal essay, “On the Psychology of the Uncanny.” Das Unheimliche translates to “the un-homely,” and later authors like Freud (1919/2003) contrast the un-homely with the homely, familiar, domestic or well-known (heimlich, heimisch, or alternatively, vertraut) (124-25). In any case, Jentsch’s (1906/2008) suggestion here is that the reason why some objects/circumstances are uncanny is because they are in some way “not at home” within the usual experience of

---

54 Emphasis is the author’s.
56 Also cf. translator’s notes (159). Freud, like Jentsch, also pays quite close attention to the etymology of unheimlich, in order to reveal the nature of uncanny phenomena Cf. Part I of “The Uncanny” (1919/2003).
the subject, *i.e.* she is confronted with a sort of “uncertainty” or “unfamiliarity” about some aspect(s) of the uncanny object or situation. This suggestion stems from Jentsch’s analysis of the German usage of *unheimlich*, and this usage corresponds fairly well to its English equivalent. “Uncanny” comes from the Scots **canny,** derived from “*can,*” meaning “knowledge,” “know-how” or “skill.” Since “uncanny” is often associated with what is “creepy,” “eerie,” “spooky,” *etc.*, it can be likened to *unheimlich* insofar as both words express the unfamiliarity or foreignness of those instances and things which lie outside of the subject’s usual experience.

I shall pause to note that I am using the word “eeriness” to refer to the emotional response elicited by encounters with the uncanny. As noted, “eerie” could also be used to refer to the characteristics of the object/situation in question. Likewise, “uncanny” can refer to an object/situation, or to the feeling that the object/situations elicits in a subject. But here, “uncanny” shall be predominantly used to refer to the objects or situations which cause a subject to feel eeriness. This distinction is simply to ensure the investigation proceeds clearly. The reader must bear in mind, however, that I am referring to the negative uncanny. I use “eeriness” and “uncanny” in this way, because “eerie” and “uncanny” are closest in semantic content to *unheimlich* (*cf.* Freud 1919/2003, 124).

Of course, to supplement what I have discussed concerning how “uncanny” and *unheimlich* are used in their respective languages, I must turn to a more detailed body of empirical literature on the topic. If only there was such a body of literature. Following

---

57 I use “Scots” rather than “Scottish” to avoid confusion. Lowland Scots is the Germanic language from which the word “uncanny” comes. Scottish Gaelic – so named to distinguish it from Irish Gaelic – is a Celtic language that is quite distinct from Scots.

Sigmund Freud’s essay, “The Uncanny” (1919/2003) – in part a response to Jentsch’s essay – the study of the uncanny has been confined largely to the humanities. Indeed, it is Freud’s work on the uncanny rather than Jentsch’s that has had the most enduring influence within academia, an influence that extends into areas such as literature, poetry, film, feminism, queer theory, cultural studies, and architecture (for example, cf. the detailed studies on the uncanny collected in Collins & Jervis 2008; also cf. Royle 2008 and Vilder 1992). Unfortunately, there is not much research dedicated specifically to understanding the psychology of the uncanny taking place within the sciences, besides research on the uncanny valley phenomenon (see Section 3.2). So before dealing with this problem, let us see how Jentsch’s and Freud’s essays on the uncanny compare to one another.

2.2: Jentsch and Freud on the Uncanny

As I just mentioned, Sigmund Freud is the scholar to whom interest in the uncanny within the humanities is primarily owed. However Freud’s (1919/2003) essay, along with Jentsch’s (1906/2008) – as Jentsch’s essay’s title suggests – is concerned with understanding the psychology of the uncanny, which falls under the purview of the social and behavioural sciences. Freud is unsatisfied with Jentsch’s suggestion that uncanny experiences result from of “psychical uncertainty.” Instead, he offers a psychoanalytic account of the psychology of the uncanny. He emphasizes that what is uncanny seems unfamiliar and unsettling because uncanny things are otherwise familiar things that have been made unfamiliar by repression. But before explaining the specifics of Freud’s psychoanalytic approach toward understanding the uncanny in full detail, let us examine
Jentsch’s argument that the uncanny emerges as a result of a psychological “lack of orientation” (1906/2008, 217) a bit further.

Jentsch (1906/2008) does not attempt to define the essential nature of the uncanny, for he thinks a stimulus may not necessarily elicit the same experience of the uncanny in every subject. It’s also conceivable to Jentsch that a subject may not find the same stimulus uncanny each time she is exposed to it, and that the same stimulus may not seem uncanny in the same way during each and every different exposure. For these reasons, he attempts to understand how one’s “psychical conditions must be constituted so that the ‘uncanny’ sensation emerges”\(^59\) (217-18), \emph{i.e.} he looks for the essential conditions for an uncanny experience to occur, rather than for the essence of the uncanny.

The “psychical conditions” Jentsch (1906/2008) identifies are, once again, “uncertainty,” “undecidability,” and “unfamiliarity,” which a subject experiences when she encounters something that is “not quite ‘at home’ or ‘at ease’ in the situation concerned” (217). He names a number of examples of things which can cause one to feel uncertain in this way, but his discussion of the uncanniness of human-like artifacts is of particular interest here. For instance, he notes that wax figures of the human form can elicit an “uncanny sensation” in a subject (222-223), as can mistaking lifeless things “as part of an organic creature”\(^60\) (224). Additionally, epileptic seizures can leave “the unschooled observer” with the impression that “mechanical processes” have taken over the otherwise organic-seeming human being, yet again resulting in an uncanny sensation,

\(^{59}\) Note that “psychical conditions” can be contemporized as “psychological conditions.” I use Jentsch’s term when I quote him of course, but use “psychological conditions” elsewhere.

\(^{60}\) E.g. “a rafter covered with nails [for] the jaw of a fabulous animal, a lonely lake [for] the gigantic eye of a monster and the outline of a cloud or shadow [for] a threatening Satanic face.”
i.e. eeriness (226). But out of all the examples Jentsch provides, the most outstanding are automata. He adds that the “feeling of uncertainty” characteristic of encounters with the uncanny is very pronounced when imitations of human beings “appear to be united with certain bodily or mental functions”:

For many sensitive souls, [a life-size wax or similar figure] also has the ability to retain its unpleasantness after the individual has taken a decision as to whether it is animate or not. Here it is probably a matter of semi-conscious secondary doubts which are repeatedly and automatically aroused anew when one looks again and perceives the finer details; or perhaps it is also a mere matter of the lively recollection of the first awkward impression lingering in one’s mind […]

He continues:

This is where the impression easily produced by the automatic figures belongs that is so awkward for many people […] A doll which closes and opens its eyes by itself, or a small automatic toy, will cause no notable sensation of this kind, while on the other hand, for example, the life-size machines that perform complicated tasks, blow trumpets, dance and so forth, very easily give one a feeling of unease. The finer the mechanism and the truer to nature the formal reproduction, the more strongly will this special effect make its appearance” (222-23).61

According to Jentsch, what all of these examples have in common is that they cause a subject to experience a powerful uncertainty as to whether inanimate things are not in fact animate, and vice versa. That is, the perception of apparent animacy in a seemingly inanimate object, such as an automaton – or alternatively, not perceiving life in something which is expected to exhibit animacy such as a corpse – creates the feeling of “uncertainty” or “unfamiliarity” because the attribute of animacy is in some way out of

61 Note here the similarities between Jentsch’s (1906/2008) remarks about the uncanny, and Mori’s (1970/2012) uncanny valley theory. On both of these theories, the more human-like an artifact is the greater the likelihood that it will produce a sense of eeriness in a subject.
place given a subject’s usual experience with, or expectations about, that particular uncanny stimulus.

Jentsch’s (1906/2008) idea that uncertainty about whether something is alive\(^\text{62}\) elicits an uncanny response foreshadows some of the more recent empirical work on perceptual animacy and the uncanny valley phenomenon (e.g. Saygin et al 2013; Looser & Wheatley 2010), but is overlooked by others (Misselhorn 2009).\(^\text{63}\) Before getting to that, however, let us further examine what Freud (1919/2003) has to say about the psychology of the uncanny. Freud’s argument turns on the veracity of his suggestion “that the uncanny is that species of the frightening that goes back to what was once well known and had long been familiar” (124; also cf. 152), \(i.e.\) the homely and familiar (\textit{heimlich, vertraut}) becomes unfamiliar, “un-homely” (\textit{unheimlich}) and therefore frightening.\(^\text{64}\) How does this occur? According to Freud, the answer has to do with repression; or, more accurately, with the return of the repressed, an idea that features prominently throughout Freud’s essay and much of his other work. On this view, things which are otherwise familiar – \(e.g.\) things we might have fears and anxieties about – become unfamiliar as they are repressed and kept hidden in the unconscious mind. When

\(^{62}\) However, as Freud (1919/2003) points out, uncertainty and unfamiliarity don’t provide a complete explanation for the uncanny (125; also cf. Kang 2011, 50). I answer this criticism below.

\(^{63}\) An astute reader will notice that the examples mentioned above, \(e.g.\) automata, are animate insofar as they move. Indeed, “automata” means “self-mover.” They are not animate in the same way that you or I are animate, \(i.e.\) they aren’t alive, so they are called inanimate objects. Yet they also do not move in the same way that other inanimate things subject to natural forces move, \(e.g.\) the universe as a whole conceived of by early modern philosophers as a clockwork device, or, what astronomer Carl Sagan (1980) was fond of calling the “awesome machinery of nature.” Misselhorn (2009) refers to human-like artifacts as “inanimate objects,” but it’s clear that she knows they aren’t one’s typical inanimate objects. For further discussion, see Section 5.

\(^{64}\) This idea is foreshadowed by Freud’s (rather lengthy) analysis of the etymology of \textit{unheimlich}. Cf. “The Uncanny,” (1919/2003), 123-34.
they return to consciousness, symbolized in dreams for instance, they make a subject feel uneasy.

The most outstanding example of uncanniness discussed by Freud (1919/2003) is E.T.A. Hoffman’s (1815) short story “Der Sandman.” Freud apparently seizes upon this particular example because Jentsch (1906/2008) takes note of Hoffman’s use of uncertainty in order to elicit an uncanny impression in his readers, and because of the presence of automata in this particular story (224). In “Der Sandman,” the reader meets Olympia, after whom the young Nathaniel begins to lust. It is eventually revealed that she is actually an automaton, but the reader is left uncertain about Olympia’s true nature until well into the narrative, when the revelation of this fact to Nathaniel nearly drives him insane. Freud rejects the idea that this story creates an uncanny impression in the reader because he or she is uncertain about whether Olympia is alive or simply some kind of clockwork device. Instead, he maintains that the “more potent” (138-39) instance of the uncanny in “Der Sandman” is the titular antagonist himself.

Nathaniel learns of the Sandman from the old woman who helps to care for his little sister, who tells him the Sandman throws sand into the eyes of children who do not go to bed when they are told to. The eyes then fall out of the children’s heads, and the Sandman takes them back to his home on the moon. There, he feeds the eyes to his children, who are horrible-looking creatures with hooked beaks in the place of mouths. The Sandman is eventually revealed to be Coppelius, a friend of Nathaniel’s father who

---

65 According to Kang (2011), the theme of the automaton appears in some of Hoffman’s other works, e.g. “Die Automate” (1814) which features a male automaton in contrast to the character Olympia that appears in “Der Sandman.”

66 Who is also the spyglass salesman Coppolla, who makes the eyes of the automaton Olympia.
threatens to throw hot coals\textsuperscript{67} into Nathaniel’s eyes when he spies Coppélius and his father after hours in the study, when he should have been in bed sleeping like his mother instructed him (\textit{cf.} Hoffman 1815).

I’ll have to leave many of the other details of the narrative of “Der Sandman” aside. What matters, for my purposes, is that Freud (1919/2003) thinks uncanniness in this story is attributable to the castration complex (139-140). That is to say that the fear of damaging or losing one’s eyes symbolizes the repressed fear of the loss of the male member. Nathaniel’s fear of the Sandman is, according to Freud, ultimately a case of castration anxiety, or \textit{kastrationangst}. This fear – which had been repressed – returns to consciousness as the fear of damaging another part of the body like the eyes.\textsuperscript{68} Thus, the castration complex is one example of the return of the repressed that I mentioned earlier.

I think that Freud (1919/2003) places far too much emphasis on Nathaniel’s own fears and anxieties in “Der Sandman,” rather than what the audience is intended to feel, which is what Jentsch (1906/2008) is concerned with when he discusses Hoffman’s talent for eliciting a sense of eeriness in his readers. Moreover the castration complex – toward which Freud (1919/2003) devotes a lot of attention in his essay – may not be the best explanation for eeriness with human-like artifacts. Women also experience eeriness and can encounter “the uncanny,” so the castration complex as an instance of the return of a

\textsuperscript{67} The coals take the place of the sand from the version of the tale told to Nathaniel by his mother.
\textsuperscript{68} Freud also thinks that severed body parts – especially when they move by themselves – are uncanny for this reason. On his view, they elicit the same sort of castration anxiety that fear of damaging the eyes does (1919/2003, 150).
repressed fear seems to be an untenable explanation. However, the castration complex may also be symbolic or metaphorical (cf. Lazarus 1991, 234). That is, one doesn’t literally fear the loss of one’s member, but rather the power which the male member is sometimes said to symbolize, e.g. in the Freudian tradition of thought.

Here it seems that the castration complex might explain fears or anxieties about robots in certain cases. Consider, for example, how in science fiction stories robots often revolt against their human masters, compromising the power of the human over its creation. This theme goes back to the very first work of fiction in which the word “robot” appears, viz. Karel Čapek’s play *R.U.R. (Rossum’s Universal Robots)* (1921). It is also found in Mary Shelly’s *Frankenstein* (1818), and in the tale of the Golem of Prague from Jewish folklore. In both of these stories, an artificial life-form turns on its master. Of course, the theme of losing power over one’s creation, symbolized by castration, is found in ancient myth as well. Think here of the Titan Cronus who castrated his father Uranus, which is the act after which the castration complex is named, or of Zeus who, together with the Olympians, later supplanted the Titans by overthrowing Cronus and casting him down into Tartarus.

However, in reality robots have not yet proliferated enough to conclusively answer whether or not people perceive them as threatening to their sense of power and/or control. If this is a human tendency when it comes to artificial life, then it may change

---

69 Although a staunch Freudian might reply that this fear in women is attributable in some way to penis envy – perhaps one is afraid the loss of a member one unconsciously wishes one had – I don’t think this is enough to save Freud’s theory.

70 “Robot” would eventually replace “automaton” as the word used to describe machines that are fashioned after the human form. It comes from the Czech word *robota*, which refers to a serf or a slave and the drudgery which they must perform.

71 Note that I am aware this latter example is not a science fiction story.
with the widespread use of social robots. Further research is required, although such questions have been addressed by some researchers. MacDorman, Vasudevan & Ho (2009), for example, found that Japanese and Western subjects exhibit similar implicit biases against robots, casting doubt on the occasionally encountered cultural stereotype than the Japanese are more comfortable with robots than North Americans or Europeans. In any case, as I discuss in Section 2.3, Freud’s (1919/2003) theory of the uncanny works quite well when it is applied to works of art, literature and film. It is uncertain whether it applies equally well to real human-like artifacts.

What else might be repressed besides fear of castration? The castration complex is but one example of the return of the repressed, and it’s possible that some other repressed fear might explain the psychology of the uncanny. A summary of the other examples of the uncanny that Freud (1919/2003) examines includes, for instance, the idea of the doppelgänger or double (141-143). He also lists coincidences which elicit a sense of deja vu, such as: when one becomes lost yet repeatedly finds oneself back in the same place where one just was; when one notices the same names or numbers in different circumstances and places within a short space of time. To these he adds other strange coincidences, such as: when one is thinking of a person one has not seen in years and then suddenly encounters him or her; and cursing someone to be struck dead only to learn that this person dies shortly after the “curse” is uttered, as examples of uncanny occurrences (143-46).

To explain uncanniness in cases like those listed above, Freud (1919/2003) appeals to what he calls “the omnipotence of thought.” The omnipotence of thought is the belief that our thoughts can affect the world (147). Freud thinks that this belief stems from an animistic view of the universe, which many of us outgrow as we leave childhood and repress when we adopt more rational ways of understanding the world around us. These repressed elements of a more primitive worldview then make their appearance when we are confronted with something that causes our animistic intuitions to run amok. Thus once again what is unfamiliar or foreign, i.e. un-homely (unheimlich), turns out to be “what was long familiar” (heimlich), returning from repression in the unconscious mind. He continues:

> The analysis of cases of the uncanny has led us back to the old animistic view of the universe, a view characterized by the idea that the world was peopled with human spirits, by the narcissistic overrating of one’s one mental processes, by the omnipotence of thoughts and the technique of magic that relied on it, by the attribution of carefully graded magical powers (mana) to alien persons and things, and by all the inventions with which the unbounded narcissism of that period of development sought to defend itself against the unmistakable sanctions of reality (147).

Freud’s description of the animistic worldview is in many ways just re-statement of a similar proposal by Jentsch (1906/2003). Recall that Jentsch thinks uncertainty about whether something is animate or inanimate is a prime example of an encounter with the uncanny. On Jentsch’s view, this can apply not only to human-like artifacts, but to the whole world:

---

73 Emphasis is the author’s. Also cf. 155-57.
[An] important factor in the uncanny is the natural tendency of man to infer, in a kind of naive analogy with his own animate state, that things in the external world are also animate or, perhaps more correctly, are animate in the same way. It is all the more impossible to resist this psychical urge, the more primitive the individual’s level of intellectual development is.

Note that here, “primitive” is refers to the ancient or pre-historic. It’s clear that these primitive “psychical urges” manifest themselves despite our cultural advancement. For “even in highly cultivated Greece,” Jentsch continues, “a dryad still lived in every tree.” Therefore, he concludes that:

It is […] not astonishing if that which man himself semi-consciously projected into things from his own being now begins to terrify him in those very things, or that he is not always capable of exorcizing the spirits which were created out of his own head from that very head. This inability thus easily produces the feeling of being threatened by something unknown and incomprehensible that is just as enigmatic to the individual as his own psyche usually is as well (225).

Consider further Jentsch’s (1906/2008) proposal that “intellectual certainty provides psychical shelter in the struggle for existence [and] the lack of such certainty is equivalent to lack of cover in that never-ending war of the human and organic world […]” (227). Thus Jentsch’s explanation for the uncanny, as Borody (2011) points out, is really an evolutionary explanation. In other words, a contemporary interpretation of Jentsch’s (1906/2008) essay is that the “uncanny sensation” is a consequence of the evolutionary history of the mind. This insight is important, because when it comes to the uncanny valley phenomenon, Mori (1970/2012) also thinks that the eerie response elicited by human-like artifacts may have an evolutionary explanation, remarking that it is probably “an integral part of our instinct for self-preservation” (100).
Freud’s (1919/2003) version of the animistic worldview also retains significance when looked at from an evolutionary perspective. Indeed, he discusses “primitive man” in much the same way that Jentsch (1906/2008) does. But Freud’s (1919/2003) explanation for the uncanny is ultimately psychoanalytic. I think that his response to Jentsch (1906/2008) amounts to a re-statement of most of his original ideas with a psychoanalytic veneer, and I don’t think there is need to appeal to a psychoanalytic explanation for the negative uncanny when one encounters it in everyday life, or with human-like artifacts. This is because the animistic worldview is not really “repressed” in the Freudian sense. Instead, these tendencies are probably mediated by mental processes which operate on an unconscious and automatic level, so as to leave higher cognitive abilities free to deal with novel situations (cf. Bargh & Chartrand 1999), an important adaptive ability for the evolutionary development of *homo sapiens*. That is, these abilities are simply primitive in the evolutionary sense. Let us now see whose research – Freud’s or Jentsch’s – can prove to be the most helpful when it comes to explaining the psychology of the uncanny, as well as what kind of emotion eeriness is.

2.3: Psychological Uncertainty or the Return of the Repressed?

I think that, while Freud (1919/2003) is correct that the animistic worldview plays a role in experiences of the uncanny, on Jentsch’s (1906/2008) theory there is no need to appeal to the return of the repressed to explain the psychology of the uncanny. Jentsch’s view is also, as noted, consistent with Mori’s (1970/2012) suggestion – as well as the
suggestion of other researchers (e.g. MacDorman & Ishiguro 2006, 310-12; Yamada et al 2013, 21; 28-30) – that the uncanny valley phenomenon might have an evolutionary explanation. But if Jentsch’s (1906/2008) theory is better suited for explaining eeriness with human-like artifacts, this does not mean that there is nothing valuable in Freud (1919/2003) when it comes to understanding the uncanny. Certainly, his work is useful for interpreting works of literature or even films. Consider, for example, Slavoj Žizek’s psychoanalytic interpretation of film in The Pervert’s Guide to Cinema (2006). Alternatively, consider the earlier-mentioned rebellious robots from science fiction films and television series, and the concept of the castration complex which seems to be employed thematically within them (see Section 2.2).

After all, one way to understand the uncanny is to re-create it in art, or with art. That is to say, the artist can have her characters encounter the uncanny, or the audience encounter the uncanny, or both. Here is where I think Freud’s (1919/2003) ideas concerning the psychology of the uncanny are the most helpful. This also explains why his work enjoys such attention within the humanities. Examining these experiences of the uncanny can also help to answer (1), viz. what is it like – from a subjective standpoint – to encounter the uncanny? However caution is required when taking this approach toward answering (1). For as Freud points out, “we should distinguish between the uncanny one knows about from experience from the uncanny one only fancies or reads about” (153-54). Feeling eeriness when watching a horror film, for example, might feel different than

74 Here MacDorman & Ishiguro have in mind evolutionary aesthetics, i.e. physical traits which may indicate the presence or absence of pathogens or genetic defects in potential mates (cf. Etcoff 1999). I think that the uncanny valley phenomenon probably has an evolutionary explanation, but not the one that MacDorman and Ishiguro discuss. I discuss this idea further in Section 3.4.
during an encounter with a human-like artifact; it also may be attributable to different causal mechanisms. The eerie sensation in these cases also co-occurs with feelings of shock, disgust and suspense. I’m concerned with eeriness that one knows from experience, so I shall not say any more about the matter of Freudian approaches toward the uncanny in the humanities.

Jentsch’s (1906/2008) contributions are just as helpful as the above is for contending with (1), but what he extrapolates about the psychology of the uncanny from his examination encounters with uncanny objects and/or situations is, as I shall show throughout the rest of this investigation, more helpful for answering (2), viz. what does an objective view of the uncanny reveal about the emotion of eeriness? However, even though I think Jentsch’s (1906/2008) theory of the uncanny is superior to Freud’s when it comes to eeriness with human-like artifacts, Jentsch’s account is not without its problems. I shall have to address these before moving on to determine what kind of emotion eeriness is. Consider Freud’s (1919/2003) criticism of Jentsch’s argument:

On the whole Jentsch does not go beyond relating the uncanny to the novel and the unfamiliar. For him the essential condition for the emergence of a sense of the uncanny is intellectual uncertainty. One would suppose, then, that the uncanny would always be an area in which a person was unsure of his way around: the better oriented he was in the world around him, the less likely he would be to find the objects and occurrences in uncanny (125).  

It is true that Jentsch’s (1906/2008) “essential condition” for experiencing an “uncanny sensation” – what I’m calling “eeriness” – is uncertainty, and Freud’s account shows that

75 Also cf. Kang (2011). He makes a similar observation about “familiarity” (shinwakan). That is, what is unfamiliar is not always eerie (50).
sometimes even familiar and well-known things can elicit an uncanny response. But recall that Jentsch also notes that differences exist in people’s sensitivity to uncanny sensations. This might explain why not everyone is perturbed by the unfamiliar in every case. Admittedly, however, this doesn’t explain why familiar things, e.g. artifacts such as automata which have a human and therefore familiar form, are perceived as eerie. This is why Freud tries to account for the connection between the familiar and the unfamiliar using the psychoanalytic concept of the return of the repressed.

I think that Freud (1919/2003) is correct to conclude that uncertainty or unfamiliarity is not sufficient to explain why encounters with the uncanny feel the way they do. However his criticism of Jentsch (1906/2008) may be a bit uncharitable, for Jentsch does in fact go beyond simply relating what is uncanny to what is “un-homely” and unfamiliar. Recall that the most outstanding cases of uncanniness in Jentsch’s essay occur when a subject is “psychically” or psychologically uncertain about whether something is animate or inanimate. Indeed, uncertainty concerning whether or not something in the world is animate or not is architectonic in Jentsch’s theory of the uncanny; this kind of uncertainty is present in nearly every one of the examples of encounters with the uncanny which he lists. Freud (1919/2003) strays quite far from the discussion concerning automata, but he devotes a lot of attention to the earlier mentioned omnipotence of thought and the animistic worldview as they pertain to experiences of the uncanny (147).

76 See also my remarks on the counter intuitive nature of the uncanny valley phenomenon in Section 1.1.
If there is one thing to take away from both Freud (1919/2003) and Jentsch (1906/2008), it is that this animistic worldview – which is arguably reducible to the misperception of animacy and agency – may explain why things seem uncanny. Take human-like artifacts for instance: they move, and appear to be animate, but they can also appear to act with a certain goal in mind, for example, when they are programmed to carry out a specific task; they can appear to possess “bodily and mental functions” (Jentsch, 1906/2008, 223). In other words, things which appear to be animate such as human-like artifacts also appear to possess agency, one of the attributes of mind identified by Gray et al (2007) (see Section 1.3). This might somehow clash with one’s knowledge that robots and other artifacts don’t possess animacy and/or agency like humans do, and this uncertainty might factor into a subject’s sense of eeriness. Indeed, the conceptually overhauled uncanny valley theory that I am working toward describes something generally along the lines of what I’ve just outlined above.

The above-described uncertainty might also apply to other cases wherein one experiences a sense of eeriness. For instance, it’s conceivable that someone who thinks they’ve encountered something “spooky” has somehow misperceived animacy and/or agency where neither of these things is present. For example, a subject perceives that, apparently, a ghost or spirit utters a word. This is in fact a misrecognition: the subject may have thought she heard something when only the background noise of the environment was present; or, she may have perceived something moving (apparently without a cause) and attributed the motion to some unseen force, when in fact it is just the house shifting, or the a draft causing something to move, etc. In the examples mentioned above a subject misperceives agency and forms the belief that there is some mind or
conscious force at work within the world, perhaps spirits or ghosts, or in more primitive societies, gods and demons. Making such an attribution might even be a way for a subject to remove some of her “psychical uncertainty” in that situation, which otherwise makes the subject feel very uncomfortable.

2.4: Toward a New Psychology of the Uncanny

In the above I have kept largely to examples of the negative uncanny; however, at this point I must acknowledge that the uncanny does not always elicit a negative response like eeriness. Thornton (2012, 195-96) for example highlights Wittgenstein’s (1980) discussion of the “everyday uncanny,” which can, in modern psychological terminology, cause either a positively or negatively valenced emotional response. Wittgenstein (1980) uses the example of a play in which the actors portray activities that all of us do in our day-to-day lives, but which are transported out of the world and onto a stage, making them seem uncanny, albeit but not in a negative way. Rather, he thinks that this example is both “uncanny” and “wonderful.”

Consider another example of the uncanny from Jean-Luc Goddard’s film *Vivre sa Vie* (1962), during the 11th scene when Nana (portrayed by Anna Karina) has a discussion with a *Philosophe* in a café. Here I draw attention to the moment that Nana breaks eye contact with her interlocutor and stares directly into the camera. This is uncanny insofar as an actor staring into the camera is generally avoided in film. An audience normally expects the characters of a film to engage with one another; looking at the audience breaks the so-called “fourth wall,” and violates the usual expectations an
audience has about what watching a film is supposed to be like. This is certainly uncanny; it may even be a bit surprising or shocking, but it is not particularly eerie (at least not to this author).  

Therefore, what is uncanny is not always eerie or frightening. However, following the results of Freud’s (1919/2003), and especially Jentsch’s (1906/2008), investigations, it is clear that human-like artifacts provide sufficiently regular instances of uncanniness in order to study the psychology of the negative uncanny more closely. Although I have yet to review further empirical research on the uncanny valley phenomenon, I think that the preceding exegesis of Jentsch and Freud also shows that a phenomenon something like the uncanny valley exists, further supporting the results of my discussion in Section 1.3. That is, things which appear both animate and very human-like can be perturbing when these apparent features are sufficiently human-like. Now, it is time to determine conclusively what kind of emotion eeriness is. This is the task that I shall take up in the following section.

---

77 Of course, the “mocumentary” genre – comedies presented as if they were documentaries with commentary by the characters occurring between scenes as in Ricky Gervais’ and Stephen Merchant’s The Office (2001-2003) – provides another example of the positive uncanny. The “interviews” with the characters and the situations they find themselves in are filmed as if they were real. These interviews also break the “fourth wall,” for the characters are speaking into the camera, and thus to the audience. In this sense, “mocumentaries” are uncanny, but they are not frightening. Instead, they often aim for a comedic effect.
3.0 – On the Emotion of Eeriness

How can I describe my emotions at this catastrophe, or how delineate the wretch whom with such infinite pains and care I had endeavoured to form?

– Victor Frankenstein, in Mary Shelly’s *Frankenstein; Or, the Modern Prometheus* (1818)

In the previous section, I showed that automata – the predecessors of modern robots – have been regarded as uncanny since long before Mori (1970/2012) proposed the theory of the uncanny valley. However, I have yet to show that eeriness is a type of anxiety, and why the results of Jentsch’s (1906/2008) and Freud’s (1919/2003) studies on the uncanny lend support to this proposal. These are the tasks of the present section. I begin with a brief discussion of what emotions are, and then discuss the results of two empirical trials, one of the goals of which was to investigate the connection between eeriness and fear. The results of the empirical studies discussed below lend support to Freud’s (1919/2008) suggestion that eeriness is related to fear (123-25; 152-53), which I discussed in Sections 2.2-2.3.

Following this, I turn toward more recent work on the topic of fear and anxiety. Specifically, I examine Lazarus’ (1991) proposal that “fear” is an ambiguous term, i.e. there are many different kinds of fear. So, Freud’s (1919/2003) suggestion that the uncanny is a “species of the frightening” (124; 152) may need to be refined. I shall show that upon on close examination, eeriness appears to be a type of anxiety. Next, I explore the possibility that death anxiety and mortality salience – following MacDorman’s (2005) Mortality Salience hypothesis – play a role uncanny valley phenomenon. I ultimately reject this suggestion, instead proposing that eeriness is a type of anxiety that is elicited specifically because of psychological uncertainty or ambiguity concerning the ontological
nature of, in this case, the human-like artifact. I’ve already mentioned of the attributes that can become the objects of this uncertainty in my discussion of Mori (1970/2012), Jentsch (1906/2008) and Freud (1919/2003), viz. apparent animacy and human-like appearance (see Sections 1 and 2), and briefly mentioned another, viz. mind (cf. Jentsch 1906/2008; Gray & Wegner 2012; Misselhorn 2009). I’ll address the question of precisely how these salient attributes elicit a sense of eeriness in the final two sections of this investigation, after I’ve determined what kind of emotion eeriness is.

3.1: What are the Emotions?

This question cannot be answered definitively herein, for there remains an ongoing debate amongst philosophers and psychologists concerning what the correct answers to it are. However having at least some idea of what emotions are is integral to specifying what type of emotion eeriness is. A good place to begin, then, is with the work of Richard Lazarus. On his view, emotions indicate that a significant event, e.g. something that is either “harmful, threatening, or beneficial” (Lazarus & Lazarus 1994, 4), is happening to the person experiencing the emotion. Elsewhere (Lazarus 1991, 38-39) explains this idea in terms of “relational themes,” i.e. how an individual judges the significant event in her environment to affect her. On Lazarus’ view each emotion involves personal meaning which turns on what is important to the individual, and comes with its own distinctive “dramatic plot” that defines the significance of what has happened to a person who experiences the associated emotion (1994, 4-5). Emotions, in other words, help a subject to tell what things and/or events are important, and why.
Furthermore, on Lazarus’ (1991) view emotions are highly cognitive. We make a number of “appraisals” – *viz.* judgements about the aforementioned relational themes – every time we experience a particular emotion. This is why each emotion has its own set of relational themes, and it is also how each emotion comes to have its own distinct “dramatic plot.” One makes an appraisal about how an event affects oneself, and that appraisal determines the emotion that is experienced. However, others posit that emotions are perceptual rather than cognitive. For example, Jesse Prinz (2004) – following in the footsteps of William James and Carl Lange, who thought that bodily sensations were the primary components of the emotions – thinks that emotions represent body states, which in turn represent the relation of important features in the subject’s environment to the subject. Thus, Prinz retains the relational themes from Lazarus’ (1991) theory. But on Prinz’s (2004) theory, we don’t form judgements the way Lazarus (1991) describes, rather, these relational themes are perceived. They are what the emotions ultimately represent (body states and features of the environment).

I shall use elements from Lazarus’ (1991) theory of emotion to determine what kind of emotion eeriness is. However, I must point out that some of the results of the empirical studies I discuss herein show that there is a significant perceptual component to eeriness, whatever type of emotion it is. So ultimately, the theory of eeriness I present

---

78 They are: goal relevance (is the object or situation (X) relevant to the subject’s (S) goals?); goal congruence (is X congruent with S’s goals, or does it frustrate S’s attempt to accomplish them?); ego-involvement (how are S’s goals, well-being, ethical concerns, *etc.* effected by X?); “blame or credit?” (S assesses X to determine how X facilitates goal congruence and ego-involvement); coping potential (how does S deal with the situation?); and, future expectancy (how X effects S’s future goals) (*cf.* Lazarus 1991; Prinz 2004, 14-7; Ch. 2). The aforementioned relational themes are mediated by these appraisals, so how a subject appraises X determines what emotion(s) she experiences.

79 At least, in many cases we don’t. But Prinz (2004) thinks some emotions may be more complex, built up out of simpler emotions. We may make judgements in cases like this.
may be more compatible with Prinz’s (2004) theory of emotion than Lazarus’ (1991). This question falls outside of the scope of this investigation, although it is certainly an important topic for future research. In any case, neither Prinz (2004) nor Lazarus (1991) investigates what kind of emotion eeriness is, so let us move on to examine what the research on the uncanny valley reveals about the possible relation between eeriness and fear. Along the way, I shall say a bit more about the emotion of fear itself.

3.2: Eeriness and Fear

Answering the question of how to describe the emotions elicited by modern human-like artifacts, such as CG characters and robots, is undertaken by Ho et al. (2008). In an empirical study on the uncanny valley phenomenon, Ho and his colleagues measured eye movements and other physiological responses while the participants viewed video clips of dozens of robots, some very human-like and others more mechanical (a video of a human female was used as a control). Questionnaires which asked the subjects to describe the emotions they experienced while viewing the videos were used to collect additional data. The results of the questionnaires and the

---

80 Since the uncanny stimuli I am focusing on – viz. human-like artifacts – have aesthetic features such as human-like appearance and motion, it may be worth considering whether eeriness is what Lazarus & Lazarus (1994) call an “aesthetic emotion,” i.e. an emotion aroused by an aesthetic experience (116). However, a problem immediately emerges by pursuing this option: if the only criterion for qualifying as an aesthetic emotion is being elicited by an aesthetic experience, e.g. an event in a play, film, painting, opera, concert, etc., then many emotions count as aesthetic ones. Fear can be elicited this way. Think, for example, of watching a horror or a thriller film. Evolutionary aesthetics (cf. Etcoff 1999) have been suggested to play a role in the uncanny valley phenomenon. For example, it might be similar to a disgust response (MacDorman 2005; MacDorman & Ishiguro 2006). However, it seems that eeriness is probably a distinct emotion from disgust, although the two can arguably co-occur.

81 Sometimes called “avatars.” Note that avatar properly refers to a CG character controlled by a human being. E.g., a character in Blizzard’s World of Warcraft, a popular multiplayer online game, is an avatar. A character from a CG film is not an avatar (except perhaps the character Jake Sully’s avatar, from James Cameron’s film Avatar (2009)).
physiological responses reveal, according to Ho and colleagues, that the subjects associated the words “eerie” and “creepy” with the robot videos that they observed, and that the subjects related these words most closely with “fear,” “shock,” “disgust” and “nervousness” (169; 172-76). More recently, Burleigh et al (2013) achieved results similar to those of Ho et al (2008). Burleigh et al (2013) used “fear,” “disgust,” and “attractiveness” as proxies for the emotion of eeriness, and found that measures of eeriness were positively correlated with measures of fear and of disgust, but negatively correlated with measures of attractiveness. They conclude that their results are compatible with those of Ho’s and his colleagues’ previous study (765-66). Together, these results suggest that eeriness is strongly associated with fear, at least when it comes to human-like artifacts.

These results also facilitate answering questions (1) and (2) which I posed at the beginning of Section 2. That is, they help to ascertain: (1) what eeriness is like from a subjective standpoint; and (2) what can be said of eeriness from an objective standpoint, respectively. However, Ho et al (2008) do not engage the question “what is eeriness?” directly; they simply determined that “eeriness” best describes the emotion subjects feel around uncanny artifacts, and that this emotion is associated with fear. Likewise, while Burleigh et al (2013, 766) think the results of their study – and Ho et al’s (2008) study – help to clarify the “phenomenology of eeriness,” they do not say much about what it is like to experience eeriness from a subjective standpoint beyond that it is associated by

82 Note that Burleigh et al’s (2013) decision to compare measures of attractiveness with these has implications for their other trials, which examine whether human-like artifacts are eerie because they possess categorically ambiguous or atypical attributes. I think this hypothesis shares some similarities with Jentsch’s (1906/2008) conclusions about the psychology of the uncanny, which is one reason why I return to it in Section 4.
their subjects with fear and disgust. That the subjects in these studies related “eeriness” to “fear” might be an indication that feeling eerie is a bit like feeling frightened. So overall, these results are helpful for showing that Freud (1919/2003) may be correct when he argues that the sensations associated with the uncanny belongs – generally – within the emotional category of “fear.” But as I show below, there are many types of fear. What kind of “fear” did the subjects in the above trials associated with eeriness? It is clear that the question “what type of emotion is eeriness?” still requires a firm answer.

3.3: Feeling Eerie

Almost everyone knows what it is like to experience fear, even though fear comes in many varieties, just as we have all probably had an encounter with something that can be described as “uncanny” or “eerie,” (see Section 2.1). Aside from certain exceptions – such as psychopaths (cf. Hare 1993/1999) – fear generally results in an increase in heart rate, respiration, and perspiration; one’s hairs can stand up on end, one’s eyes widen, and one feels a burst of adrenaline that prepares one to fight or flee from a possible threat. Darwin (1872/1955) noted that the physiological and behavioural characteristics of fear confer important adaptive benefits. For instance, the widening of the eyes may allow one to better perceive, and deal with, a present danger or threat. He notes further that the fact that our hair stands on end when we are frightened may be an evolutionary artifact, left over from the days when humans were hairier. Puffing up our coat of hair may have
made us look bigger, in turn making a potential predator/hostile conspecific question whether to attack. Many animals with fur, feathers or frills do something similar.\textsuperscript{83}

However, on closer examination it is easy to see that some experiences of fear are quite different from others. If, for instance, I am startled by something, the emotion I experience will be different than if I fear the loss of someone or something important to me. Indeed, panic feels much different than shock, and horror feels much different than anxiety, and so on. All of these different emotions cause different changes in the body and the brain, and motivate people to act in different ways. In short, the emotions grouped under the label “fear” come in many forms. This is another reason why, even if subjects associate eeriness with fear, a more precise definition of eeriness is required.

On the face of it, it is difficult to see how feeling eerie is all that similar to other cases of feeling afraid, like fear of the loss of something important, or for one’s own safety, or in more extreme cases phobias of the dentist, of heights, of pathogens or contaminants, of public places, \textit{etc.} (dentophobia, mysophobia, acrophobia and agoraphobia, respectively). This may be why Freud (1919/2003) devotes such effort to showing that uncanny sensations are a result of repressed fears, as discussed in Section 2. Yet Freud (1919/2003) is not exactly clear about what he means by “frightening.” In his discussion of the uncanny, he mentions horror and dread. But he also focuses extensively on anxiety during his discussion of the castration complex (\textit{kastrationangst}). He even switches from fear to anxiety when he discusses Nathaniel from “Der Sandman,” who fears the Sandman, but who also has castration anxiety on Freud’s (1919/2003) reading of

\textsuperscript{83} Note that such features can also be used for communication, or as displays to attract mates. The best example is, of course, the peacock.
the story. Both of these are types of “fear,” but which one is more closely related to the emotion of eeriness?

To answer this question, I’ll return to Lazarus (1991), who refers to these kinds of fear-emotions collectively as “fright-anxiety” (234-35), or alternatively, “anxiety-fright” (Lazarus & Lazarus 1994, 41-2). Anxiety-fright encompasses the related, but distinct, emotions which are often grouped together under the label “fear,” viz. shock, dread, fright, panic, horror, anxiety, nervousness, etc. I should note further that, on Lazarus’ view, the emotional category “fight-anxiety” belongs in the larger category of “existential emotions,” so-named because “the threats on which [these emotions] are based have to do with meanings and ideas about who we are, our place in the world, life and death, and the quality of our existence” (Lazarus & Lazarus 1994, 41-42).

Fright, on Lazarus view, refers to the types of fear is associated with “concrete and sudden” threats to our well-being, which could be physical, social, etc. This is closer to what it seems like Freud (1919/2003) has in mind when discusses Nathaniel’s fear of the Sandman coming to take his eyes, for instance (see Section 2.2). One’s fear of being physically threatened or harmed, of committing a social blunder, or of finding oneself in the midst of some sort of natural disaster, would count as fright on this view, for in each instance the subject’s fear has a concrete object, or an easily identifiable source.

Anxiety is a much more ambiguous kind of fear than fright on Lazarus’ (1991) theory of emotion. In fact, uncertainty and ambiguity are two of the defining characteristic of anxiety for Lazarus. Fright, as I just noted, is intentional: in cases when a subject is frightened, she has a fear of something that is concrete and direct. Conversely,
on Lazarus’s theory anxiety often has no specific object; or, someone who feels anxiety may not be able to articulate exactly why she feels anxious because she is uncertain about the reasons for her emotional reaction, or, they may be no apparent reasons for her anxiety at all, etc. Furthermore, according to Lazarus (1991, 234) the psychological hallmarks of anxiety, *i.e.* “ambiguity” and “uncertainty,” produce a “psychological tension” in the subject. So if eeriness is a kind of fear, then it may not really belong in the realm of the frightening, strictly speaking. Rather, I think the psychological tension associated with anxiety may be related to the psychological uncertainty discussed by Jentsch (1906/2008). Following these elements from Freud’s, Jentsch’s and Lazarus’ theories, I think that eeriness is a type of fear; however, it is not a type of fright, but instead a type of anxiety.

I should pause now in order to emphasize the fact that anxiety can sometimes be intentional, like fright. This sounds contradictory to what I discussed above, so let me explain: I mean that the object of anxiety might seem to be immediate, yet less concrete than things which frighten someone, such as a hostile conspecific, a predator, or the dentist. So if eeriness is a type anxiety, then it’s not because eeriness has no object. Rather, it is a form of anxiety because one has trouble articulating why something that is uncanny causes such a strange sensation. In other words if someone finds an object, person or situation “creepy” or “eerie,” she may not be able to tell you why, besides simply saying that it is an eerie and/or creepy object/person/situation. But there is one

---

84 Of course, sometimes a subject is also unable to articulate why she feels, say, anger at someone, or love for someone, but she probably will not feel anxiety as a result of this uncertainty. In cases of anger like this, I should reply that the subject is in an angry mood rather than experiencing the emotion anger. For when we become angry, it is usually because of a perceived slight or something of that sort, *i.e.* anger has a concrete object when it is felt as an emotion. But someone who is a “grouch” may simply have an angry
other salient intentional object of anxiety that I have yet to discuss, which Lazarus (1991, 234) also highlights. Namely, the prospect of one’s own death can be a source of anxiety; it might even be the *ultimate* source of anxiety. Perhaps this species of anxiety can explain why human-like artifacts can disturb a subject better than, say, castration anxiety can. Although Jentsch’s (1906/2008) theory is more promising when applied to the uncanny valley phenomenon, let us follow a suggestion of MacDorman’s (2005) and consider whether or not death-anxiety plays a role in the uncanny valley phenomenon.

**3.4: Angst, Mortality Salience, and the Uncanny Valley Phenomenon**

The prospect of our mortality can be a powerful influence on many of life’s enterprises. For instance, this is a pervasive theme in Freud’s work. More recently, Ernest Becker famously argued in *The Denial of Death* (1973) that most human accomplishments are ultimately a result of the anxiety we feel about our mortality. That is, the prospect that we will someday cease to exist, which is always present but not always consciously acknowledged, has a prolific and undeniable influence on us. Human beings are motivated by the prospect of death to create things which will outlast their finite existence.

Existentialist philosophers also emphasize the prospect of one’s death – or more precisely, the prospect of nothingness/non-existence, *i.e.* of ceasing to “be” (*cf.* Lazarus
1991, 234) – as an important part of coming to terms with the nature of one’s existence in the world. For instance, Martin Heidegger – although he vehemently denied the “existentialist” label – argues the inevitability of one’s Being coming to an end in death is a source of angst, or existential anxiety.\(^{85}\) Angst, typically viewed as a negative emotion, thus presents one with the positive opportunity to come to terms with the finitude of one’s existence (cf. Heidegger 1927/1962; 1929/1977).\(^{86}\) Subsequent existentialists like Albert Camus and Jean Paul Sartre would follow suit, paying similar attention to nothingness and the prospect of death. Now, I am not arguing that angst results in eeriness.\(^{87}\) However, if eeriness is truly a type of anxiety, then any role that the prospect of death might play in eeriness with human-like artifacts cannot be ignored.

As it happens, Mori (1970/2012) discusses the prospect of death toward the end of his essay on the uncanny valley phenomenon. He remarks that when we die, our bodies take on an eerie quality; we move from the highest peak down into the depths of the uncanny valley. He thinks that a zombie, \textit{i.e.} something which is “un-dead,” is even more eerie than a corpse (99-100). As I noted, Mori also thinks this eerie response to a corpse – and, ostensibly, eeriness in general – is an “integral part of our instinct for self-preservation.” Yet he is quite vague about how eeriness or uncanniness plays a role in

\(^{85}\) Note than “\textit{angst}” is simply the German word for anxiety. I use \textit{angst} to refer to existential anxiety or death anxiety, as is common in existentialist philosophy.

\(^{86}\) “Being” (capital B) refers to Being as such, \textit{i.e.} existence. Lower case “being” refers to particular kinds of beings, \textit{e.g.} humans, trees, toasters, \textit{etc.} Heidegger’s terminology is notoriously confusing. In fact, according to many, it approaches incomprehensible nonsense.

\(^{87}\) It is interesting to note that Heidegger pays close attention to the uncanny (\textit{unheimlich}). For Heidegger, the world takes on an uncanny quality when one realizes one’s existence will end. Coming to grips with the knowledge of the finitude of one’s Being, in other words, makes the otherwise familiar world seem somehow “un-homely.” I don’t have the space to consider Heidegger’s remarks on the uncanny in full detail, although it is interesting to note in light of what I’ve considered so far that he makes this connection between uncanniness and death-anxiety. Similar ideas recur in Mori (1970/2012) and MacDorman (2005), which I shall consider shortly. For more on Heidegger’s views concerning Angst and the un-homely, \textit{cf.} Being and Time (1927/1962) and “What is Metaphysics?” (1929/1977).
this instinct, other than that eeriness might be an indication of proximal threats, *e.g.* hostile conspecifics, *etc.* as opposed to distal threats (100, n. 2).  

It’s also not entirely clear what the connection between human-like artifacts and corpses is on Mori’s (1970/2012) theory. He does not discuss this connection – or the uncanny valley phenomenon – very extensively in his other writings (*e.g.* in Mori 1989). So, determining what the connection between a corpse and a human-like artifact is requires further exegesis. That Mori is suggesting that robots are “corpse-like” can be ruled out, however, for he rejects this idea in his subsequent work:

A dead person’s face may indeed be uncanny: it loses color and animation with no blinking. However, according to my experience, sometimes it gives us a more comfortable impression than the one given by a living person's face. Dead persons are free from the troubles of life, and I think this is the reason why their faces look so calm and peaceful (2005, n. pg.).

Misselhorn (2009) adds to this by pointing out that corpses have been a staple subject in the history of art for centuries, and most of these do not elicit eeriness (348). In fact, Kant (1790/1952) thought that one of the artistic genius’ abilities is to represent things that are otherwise disgusting and grotesque, *e.g.* corpses, without eliciting that impression of disgust, *etc.* in the audience, writing that “the Furies, diseases, devastations of war, and the like, can (as evils) be very beautifully described, nay even represented in pictures”

---

88 Mori (1970/2012) names natural disasters as examples of distal threats, whereas proximal threats refer to threats from within one’s social group. So if eeriness is the emotion that people experience when they encounter an uncanny object or situation, then perhaps it is really a kind of fright which is more concrete and direct than anxiety. After all, the examples of eeriness I am concerned with are the result of a person’s encounters with something concrete and specific, *viz.* uncanny objects like robots and automata. However as I mentioned, the reason eeriness is a kind of anxiety is not because it has no object. Instead it is because there is a psychological uncertainty or tension about that object. I should further add that distal threats can be just as concrete as proximal ones. For instance, one feels frightened if one is caught in, say, a flood, tornado, or severe earthquake. Similarly, there are proximal sources of anxiety, *e.g.* anxiety about certain social situations like parties (or thesis examinations!).
The great artists have known for far longer than roboticists and CG animators how to create human-like artifacts that do not fall into the uncanny valley, as it were. This is consistent with Jentsch’s (1906/2008) observation that “true art, in wise moderation, avoids the absolute and complete imitation of nature and living beings, well knowing that such an imitation can easily produce uneasiness” (223).\(^89\)

Another possibility is that human-like artifacts are eerie because they are either conscious or unconscious reminders of death. This is the aforementioned Mortality Salience hypothesis, proposed by MacDorman (2005). Following in the line of thought found in Becker (1973), MacDorman (2005) investigated whether humans employ “terror management” strategies when interacting with highly realistic robots. In this particular study, the artifact used was an android developed by Kokoro, overseen by Hiroshi Ishiguro who also designed and built the Geminoid HI-1. Subjects watched videos of the robot, and video of a female human was used as a control.

MacDorman (2005) hypothesized that if human-like artifacts such as robots are conscious or unconscious reminders of death, then the subjects should be expected to employ terror management strategies to cope with this \textit{angst} and/or fright.\(^90\) That is to say, when faced with a reminder of death, humans manage their fear by appealing to culturally supported worldviews. For having a worldview often entails having some beliefs about the nature of death, for instance, that it is not the end of one’s existence.

\(^{89}\) Note here the uncanny similarity between this suggestion, and Mori’s (1970/2012) “escape by design” (99-100), discussed in Section 1.1.

\(^{90}\) Here MacDorman (2005) has in mind the coping strategies uncovered by terror management research (\textit{e.g.} Solomon \textit{et al} 1998; Arndt \textit{et al} 1997; Greenberg \textit{et al} 1986). MacDorman (2005, 400) notes further that the support for terror management theory is abundant; over 200 studies, in addition to these three, have been performed which support the theory.
Think here of belief in an afterlife, in re-incarnation, or in the possibility of communication with the dead.

The subjects in MacDorman’s (2005) study demonstrated this kind of terror management. That is, the subjects who viewed the video of the android were more receptive to statements which supported their worldview than those who viewed the video of the human. Some of the subjects also reported, anecdotally, that the android was quite “scary” and “creepy.” However, I don’t think these results have much explanatory power when it comes to the uncanny valley phenomenon. For the appearance of the android used in MacDorman’s (2005) experiment was controlled such that its eyes were rolled back into its head, making it appear very corpse-like indeed. It might be reasonable to expect a corpse-like artifact to elicit terror management strategies – the above remarks on corpses in art notwithstanding – but what about artifacts which are not corpse-like? So I do not think it is conclusive that human-like artifacts are reminders of death. All that can be said for certain is that corpse-like robots can elicit terror management strategies. Needless to say, there are many human-like artifacts which are not corpse-like in the way that the above-mentioned robot is. In fact human-like artifacts – especially robots and CG characters – often give the opposite impression. That is to say that they appear to an observer to be alive, i.e. animate, as I discussed in Section 2.2.

However, there may be a connection between angst, death and the uncanny valley phenomenon, albeit an altogether different one than MacDorman (2005) proposes. Here, consider once again Mori’s (1970/2012) remarks that corpses often tumble down into the uncanny valley (100). I think that this can be explained – to a certain extent – by appealing to psychological uncertainty, as Jentsch (1906/2008) does when he explains the
uncanniness of corpses. He writes that “the horror which a dead body (especially a human one), a death’s head, skeletons and similar things cause can also be explained to a great extent by the fact that thoughts of a latent animate state always lie so close to these things” (227). In other words, people usually expect another human being to exhibit animacy. A corpse – especially a recently deceased one – may still appear in some way as if it could get up and move about again, for example, as if it were only sleeping or in a coma. Note that I am referring specifically to the recently deceased. For instance, Mori’s (2005) recent remarks on corpses aside for the moment, the body in an open-casket funeral can seem quite uncanny for the reasons that Jentsch (1906/2008) provides. Corpses which are very decomposed or disfigured – as on a battlefield (cf. MacDorman & Ishiguro 2006, 313) – can certainly elicit a disgust response. However, robots do not usually look like this, unless effort is made to make them appear somehow corpse-like, which is the case in MacDorman’s (2005) experiment. 91

When corpses seem uncanny, or when they elicit an eerie response from an observer, it may be because one is “psychologically uncertain” as to the animacy – or lack thereof – of the corpse. I think that this is why Mori (1970/2012) places zombies, along with robots, in the depths of the uncanny valley. For it isn’t really the fear of death, but uncertainty about whether something is alive or not which explains the uncanniness of corpses and human-like artifacts. Once again, I have returned to the fundamental idea behind Jentsch’s (1906/2008) account of the psychology of the uncanny: “psychical

91 Note here that feeling disgusted is quite different from feeling eerie, subjectively speaking. However, it’s conceivable that disgust and eeriness can co-occur. Think here of horror movies, many of which strive to elicit both of these responses.
uncertainty” concerning what is apparently alive but shouldn’t be, or what appears dead
or mechanical, but which should be alive, elicits in a subject a strong uncanny sensation.

3.5: What Eeriness Is

The material that I have reviewed in both the preceding and present sections allows a more conclusive answer to (2), viz. what is eeriness like, objectively speaking? I’ve argued – following my review of the aforementioned material – that eeriness is a type of anxiety. Following Jentsch (1906/2008) – and the material discussed herein – it is apparent that the intentional objects of this uncertainty are particular characteristics of an uncanny stimulus, viz. uncertainty as to whether it is alive. As I discuss in Section 5, these kinds of movement can also cause subjects to make attributions of mind; this is an idea that emerges from an empirical study by Gray & Wegner (2012), also to be discussed later. For now, it’s clear that the sort of motion Jentsch (1906/2008) discusses in his essay can also cause a subject to attribute mental functions to an inanimate artifact. I think that his discussion of the uncanniness of automata makes this quite clear.

Recall further that Jentsch’s explanation for the psychology of the uncanny is primarily an evolutionary explanation (cf. Borody 2011; see also Section 2). If, for example, one is certain that there is something alive in the immediate environment, e.g. a predator, or a conspecific, then one is in a better position to negotiate his or her way about in that environment, and to avoid possible dangers/secure benefits within it than if one were not certain, or even worse, oblivious. Indeed, knowing what’s alive and what it thinks about you – whether that approaching shape on the savannah is a predator or a
conspecific, and what it wants from you – would have put a hominid in the Pleistocene-era in a more secure position when it comes to coping with proximal threats (cf. Mori 1970/2012, 100, n. 2). Still today, knowing what others are thinking and feeling is an integral part of life, as it was for our evolutionary ancestors, for we are all highly social animals. Finally, as Misselhorn notes (2009), perceiving a human being has a distinctive phenomenology (357). It usually feels quite different, in other words, to perceive and to interact with a human than it does a duck, or a tree, or a computer, or whatever. Bearing this in mind, it is not difficult to conceive of just how uncertainty concerning what is animate and what has a mind can elicit anxiety, especially if one is uncertain about one or both of those attributes being present in something that otherwise appears so human-like, such as a robot, automaton, CG character, *etc*.

But human-like artifacts apparently cause a subject’s perceptions of animacy and mind to misfire somehow; here, subjects perceive an apparent mind, or apparent animacy, when really neither of these things is present. I think it is the case that, as Jentsch (1906/2008) observed of automata and wax figures, “semi-conscious secondary doubts [concerning “certain bodily or mental functions”] are repeatedly and automatically aroused” (1906/2008, 223) owing to the realism of the artifact (see Section 2.2). That is, such artifacts are sufficiently human-like such that even after one has formed a judgement that the artifact is indeed an artifact, one continues to perceive it in certain ways as if it were a human, or has a mind, or is alive, *etc*.

Following Misselhorn’s (2009, 357) comment above, I think that this particular type of psychological uncertainty – the intentional objects of which are the aforementioned salient attributes of life and mental functions – is what gives eeriness its
distinctive phenomenology. The average reader may not have encountered many robots or automata recently. So to gain a sense of the phenomenology of this kind of psychological uncertainty, think of what it is like when the eyes of someone in a photograph appear to follow you across a room, a common visual illusion most of us will have witnessed. Of course, this is just a trick of the eye. You know it’s a photograph. Moreover, you know that the motion of the eyes is just an illusion; in reality, the eyes are quite fixed, and it’s really your visual system “tricking you” into perceiving animacy in the eyes. But in spite of that, it still feels in some way as if the person in the photograph is watching you. This sort of thing can feel very eerie, especially if a subject walks back and forth across the room, watching the eyes of the photograph “following” her. Now think of what is happening here: it is as if the photograph is somehow alive, or as if there is a mind behind the eyes. Of course, that’s impossible. This is why when the subject continuously perceives these attributes, she feels eeriness.

Recall the discussion of what emotions are, and of the phenomenology of eeriness in Sections 3.1 and 3.3, respectively. As I noted above, neither Prinz (2004) nor Lazarus (1991) discusses this particular emotion. In fact, as I discussed in Sections 2.1-2.2, since the publication of Freud’s (1919/2003) essay, the study of the negative uncanny has been largely limited to the humanities, and most research on eeriness comes from that which is concerned with the uncanny valley phenomenon. In any case, it is difficult to see how eeriness fits within either Lazarus’ (1991) cognitively oriented view or Prinz’s (2004) perceptually oriented view of what emotions are. Yet I have discussed eeriness as an

---

Note that if it were a real human watching you, you could engage with it, figure out what its intentions are, etc. With the photograph – and other human-like artifacts – this is impossible. It’s probably the case that no one would think of doing it anyway, for as I noted, one knows the artifact is not alive.
emotion, and not as an aesthetic property of uncanny objects/circumstances. So is eeriness not really an emotion at all?

I’ll address this potential criticism before moving on. It’s difficult to see, for example, what relational themes characterize emotion eeriness on Lazarus’ (1991) cognitive theory of emotion or on Prinz’s (2004) perceptual theory of emotion. But this may simply be because neither of them – nor any psychologists who are not concerned with the uncanny valley phenomenon\(^{93}\) – examines this particular emotion in much depth. However, it might also be that case that, if emotions are cognitive, a subject who experiences eeriness cannot judge how the uncanny stimulus relates to, for example, her well-being or her goals, because she is uncertain about the ontological nature of the stimulus (does it have a mind? is it alive? \textit{etc}). In fact, this uncertainty may be one of the relational themes characteristic of eeriness. If, on the other hand, emotions are perceptual, then what relational theme does eeriness represent? Is it uncertainty about the uncanny stimulus? I think this is a possibility, however determining whether this is the case falls outside of the scope of this project. I shall have to withhold any firm conclusions about which theory of emotions best explains eeriness for the present time, although it is certainly a topic for future research.

Finally, I want to re-iterate that eeriness may not be the only emotion associated with the “negative uncanny,” \textit{viz.} encounters with the uncanny that elicit a negatively valenced emotional response. Moreover, the negative uncanny might not cause everyone to feel eerie in the same way, and other emotions may co-occur with eeriness.

\(^{93}\) Even those who do conduct research on the uncanny valley phenomenon sometimes steer clear of this question. That is why there is such an abundance of different measures of various emotional responses toward human-like artifacts, which I listed in Section 1.3.
Determining whether or not eeriness truly is a type of anxiety will require additional empirical study, and further subsequent discussion amongst researchers. However over the following two sections, it will become clear that eeriness *qua* anxiety has explanatory power when it comes to the uncanny valley phenomenon.

Nevertheless, the rough picture of eeriness with human-like artifacts that I have sketched out above requires refining. I haven’t fully explained the uncanny valley phenomenon quite yet either. It remains to be demonstrated exactly how eeriness is related to the failure of empathy, the concept that shall replace *shinwakan* following my conceptual overhaul of the uncanny valley theory (see Section 1.3). Moreover, although I’ve ruled out one possible explanation for the uncanny valley phenomenon – MacDorman’s (2005) Mortality Salience hypothesis – I have yet to examine another hypothesis that I am concerned with, *viz.* the Categorical Anomaly hypothesis. This is undertaken in the following section.
4.0 – Categorical Anomalies and the Uncanny Valley Phenomenon

I thought of what I called an “automatic sweetheart,” meaning a soulless body which should be absolutely indistinguishable from a spiritually animated maiden, laughing, talking […] as if a soul were in her. Would anyone regard her as a full equivalent? Certainly not, and why? Because, framed as we are, our egoism craves above all things inward sympathy and recognition, love and admiration. The outward treatment is valued mainly as an expression, as a manifestation. Pragmatically, then, the idea of the automatic sweetheart would not work […]

– William James, in “The Pragmatist Account of Truth and its Misunderstanders” (1908)

So far there has been occasion to discuss a few examples of empirical research on the uncanny valley phenomenon. For instance, in Section 1.7 I suggested that Tinwell et al.’s (2011) research may support Misselhorn’s (2009) hypothesis that the misfiring of affective empathy plays a role in the uncanny valley phenomenon. In Section 3.2 I showed how some of Ho et al.’s (2008) and Burleigh et al.’s (2013) empirical results lend support the conclusion that eeriness is related to fear.94 Finally, in Section 3.4 I considered and rejected MacDorman’s (2005) Mortality Salience hypothesis. Now that I have an improved understanding of the emotion of eeriness to work with, I can return to giving the uncanny valley theory its conceptual overhaul. Toward that end, in this section I examine the results of additional empirical studies. Specifically, the material reviewed herein is concerned with the uncanny valley phenomenon, and with what cognitive psychologists call “categorization,” i.e. the psychological processes that people use to recognize and understand things in the world. My goal in this section is to further develop the account of the psychological state associated human-like artifacts, which I began to develop in the previous two sections.

94 Recall further that this idea is similar to a suggestion from Freud (1919/2003): that the “uncanny sensation” is a type of fear or anxiety (see Section 2).
The explanations for the uncanny valley phenomenon that emerge out of the above-mentioned research areas – to which I refer collectively as the “Categorical Anomaly hypothesis” (CAH) – offer much insight into what Jentsch (1906/2008) calls the “psychology of the uncanny.” The CAH emerges from different lines of research which all arrive at a similar conclusion: human-like artifacts share attributes of more than one ontological category. Subjects in these experiments thus demonstrate difficulty categorizing categorically ambiguous/anomalous human-like artifacts, i.e. determining with certainty what kind of being such an artifact is. This difficulty may result in a psychological tension that contributes to a subject’s sense of eeriness in the uncanny valley phenomenon. Fortunately, the idea that some kind of psychological categorization system – in one sense or another – helps us to make sense of the world has long been a topic of interest for philosophers. I shall briefly review a few classic examples of what can be thought of as “psychological categories,” and then discuss Kang’s (2011) remarks on “binary categories.” Both of these topics shall prove to be somewhat helpful for clarifying what an ontological category is, as well as for understanding just how human-like artifacts violate categorical boundaries.

95 It could be said – as Saygin et al (2011) do – that human-like artifacts violate a subject’s “expectations” about inanimate objects and human beings. I discuss this in Section 5.

96 This may also explain what Kang (2011) calls “the power of the automaton,” i.e. the automaton’s ability to elicit emotions ranging from eeriness to curiosity, wonder and awe. However, wonder and awe are outside of the scope of this investigation. While I agree with Kang’s hypothesis, I shall have to leave the question of wonderful robots aside for now, and focus on the creepy ones instead.
4.1: What is a “Category”?

The idea that the mind organizes objects in/features of the world into discrete categories dates back to antiquity. Plato, for example, thought that the objects and properties that one perceives in the phenomenal world – the world of “appearances,” i.e. what we can see, hear, touch, etc. are “projections” or “after images” of pure Ideas or Forms which exist independently from the phenomenal world. For instance, a honeybee gets its “bee-ness” because it shares certain essential bee-like qualities with every other bee that has ever existed and that ever will exist; these essential qualities belong to the Form of Bee. Something more abstract such as goodness has as its source of being the highest form of all, the Form of the Good. According to Plato, knowledge of the Forms is innate, although to access it one must “recollect” it by practising philosophy (cf. The Meno; also cf. The Phaedo). Thus the Platonic Forms – even if they do not actually exist in some metaphysical “beyond,” i.e. something like what Kant would later call the “noumenal” world – can be thought of as ontological categories in the sense that I am concerned with here. They can even be thought of psychological categories in a certain sense. For the ancient Greeks thought of the conscious mind as the “rational” part of the

97 Note that this ancient conception of the mind is much different than what is common today. In ancient Greece, for example, the mind was thought of as the rational part of the soul. In the present day and age, it is much more common in philosophy to think of the mind as a kind of computer (cf. Crane 1995/2003). This shall prove to be important in Section 5.
98 Derived from the Greek φαινόμενον (phainomenon), i.e. that which appears or is made manifest.
99 The honeybee example is found in Meno, 72a-72c. The most illustrative example of the nature of the Forms may be Plato’s Allegory of the Cave, found in The Republic, Book VII, 514a-520a. A similarly helpful description of the Forms comes from Socrates’ discussion of the nature of the soul in Plato’s Phaedo. Of course, Plato’s dialogues are rich in content, so unfortunately there is not sufficient space for further discussion on the nature of the Forms, or the soul.
100 The concept of the noumenal world is rather abstract. It includes things like freedom of the will, right and wrong, God, etc., things which – in contrast to the phenomenal world – do not require the senses in order to comprehend. It is the world of what Kant (1787/1933) calls “things-in-themselves,” and this is similar to Plato’s Theory of Forms. Note here that “noumenal” is derived from the ancient Greek νοος (“nous”) which refers to “intelligence,” “intellect,” etc.
soul, and as I just noted knowledge of the Forms – indeed most knowledge – is innate on Plato’s view.

Similarly, Aristotle’s seminal developments in metaphysics concern, in part, what makes a particular being a member of a more abstract category of those kinds of beings (i.e., “particulars” and “universals” respectively). Here, a particular – e.g. a chair or a table – is one being which falls into an ontological category – i.e. the universal of “chair” or “table.” Plato on the other hand would say that these things receive their “chair-ness” or “table-ness” from their respective perfect Forms. Similar discussion continued in medieval philosophy, e.g. in the Nominalist tradition, where the aforementioned universals only exist as names attached to kinds of particulars which share certain essential attributes. Here, a “universal” is not really an ontological category – universals do not exist in the sense that either Plato or Aristotle has in mind on this view – but they might be thought of as psychological categories. More recently, Kant (1787/1933) proposed that space and time are the necessary intuitive preconditions for understanding the phenomenal world using innate psychological “categories of understanding,” such as causality or substance. I could go on, but I think the above examples are probably sufficient to provide the philosophically inclined reader some insight into what ontological and psychological categories are.

---

101 ψυχή (psuchē), thus “psyche” and “psychology.”
102 Cf. Plato on knowledge as recollection, i.e. the example of the slave boy found in The Meno (82a-86c).
103 I mean the kind of philosophy Aristotle undertakes which we today call “metaphysics,” not his book of the same name, although he is undoubtedly concerned with the subject matter I am discussing in that book.
105 Let me briefly note here that I am not making a theoretical commitment here to the suggestion that all categories are innate; some of them are obviously learned. People may have an innate capacity to recognize faces or learn languages, for example, but who on Earth is born with an innate understanding of “chair-
To summarize the above: ontological categories are categories of different beings; psychological categories comprise the corresponding concepts or mental schemas used for determining the ontological category membership of particular beings. But it’s possible to be even more detailed that this, following Kang (2011). He is concerned with the place automata have held in the imaginative and intellectual history of Europe since the early renaissance. To elucidate on this he focuses on a particular type of category: “binary categories” *i.e.* pairs of opposite categories, the idea of which originates from Levi-Strauss. These categories can be concrete, *e.g.* human/not-human, night/day, male/female, or more abstract, *e.g.* moral/immoral, sacred/profane, natural/unnatural, *etc.* According to Kang, automata – but especially human-like automata – are “the ultimate categorical anomaly”:

[The automaton’s] very nature is a series of contradictions, and its purpose is to flaunt its own insoluble paradox. It is an artificial object that acts as if it is alive; it is made of inert material yet behaves like a thing of flesh and blood; it is a representation that refuses to remain a stable version of the represented; it comes from the inanimate world but has the characteristics of an animate creature; and, finally, it is a manmade thing that mimics living beings (36).

Thus automata are, on Kang’s (2011) theoretical framework, “transcategorical” (32) or “liminal” (35) objects (also *cf.* Ch. 1, especially 28-54).

There are noteworthy similarities between Kang’s remarks above and Jentsch’s (1906/2008) remarks concerning the “uncanny sensation” that arises as a result of

---

ness?” However, I do think that the ability to categorize is innate. But let us leave this question aside and continue considering what a category is.
psychological uncertainty as to whether something is animate or inanimate. Thus the terminology of cognitive psychology might help to contemporize some of Jentsch’s conclusions concerning the psychology of the uncanny. Of course, I think that both Jentsch’s and Kang’s (2011) remarks can be applied to modern robots and other contemporary human-like artifacts like robots, androids and CG characters, not just automata.

However, it is possible that difficulty categorizing a “transcategorical” object – such as the above mentioned artifacts – can also occur when the categories in question are not strictly binary. Think here of the categories “human” and “robot”: these are not binary relative to one another, whereas “human” and “not-human” or “robot” and “not-robot” are. Therefore, it is better to follow Burleigh et al’s (2013) example and refer primarily, as they do, simply to ontological categories, i.e. the earlier mentioned categories of different beings (e.g. humans, robots, tables, chairs, in addition to others such as trees, stalagmites, toasters, etc.), instead of strictly binary categories (e.g. night/day, male/female, etc.). In the following subsections, the reader will find that some of the empirical studies I review focus on binary categories while others do not. This will not be a problem, for binary categories are categories all the same, but not all ontological or psychological categories are binary. Anyhow, let us now begin to review some empirical literature on categorization difficulty and the uncanny valley phenomenon.

---

106 See Sections 2 and 3 for comparison.
107 Looser & Wheatley (2010), for example, focus on the binary categories of “alive” and “not-alive,” i.e. “animate” and “inanimate.” However, Burleigh et al (2013) – as I noted a moment ago – and Yamada et al (2013) do not limit their experiments to binary categories. This shall become clear below.
4.2: Categorization Difficulty and the Uncanny Valley Phenomenon

Ramey (2005) is the first to suggest that the uncanny valley phenomenon results from an uncanny object’s categorical ambiguity. His remarks are similar to Kang’s (2011), but he is especially concerned with Mori’s (1970/2012) theory and not with the place of human-like artifacts in Western cultural and intellectual history. Specifically, Ramey (2005) thinks that uncanny human-like robots can become a sorites-paradox sort of categorical anomaly if they are sufficiently anthropomorphic. That is to say that they can become eerie when they possess categorically atypical or anomalous attributes. When such an artifact possesses enough of these so that it approaches a categorical boundary, it becomes difficult for a subject to determine what ontological category the artifact belongs to (natural/artificial, human/not-human, animate/inanimate, etc.). Thus, on Ramey’s view, an artifact’s degree of human-likeness is one of the most salient factors in the uncanny valley phenomenon.

Empirical research on categorization difficulty and the uncanny valley phenomenon is just in its earliest stages, so the available data on the relation between categorical anomalies and the cause(s) of eeriness is limited. But remarkably, two recent empirical studies – viz. those of Yamada et al (2013) and Burleigh et al (2013) – independently arrive at very similar conclusions: difficulty categorizing an object can result in the subject evaluating it negatively; and, such categorization difficulty may also explain why such negative evaluation sometimes results in a subject experiencing a sense of eeriness. As I discuss further on, these studies also call into question whether the degree of human-likeness of an artifact is the cause of this categorization difficulty. This
is also supported by fMRI studies on the uncanny valley phenomenon, which I shall review in Section 5.

To begin, Yamada et al (2013) ran three trials to determine whether difficulty categorizing objects is related to the perceived “likeability”\(^{108}\) of the object in question. In all three of their trials, morphing (see Section 1.2) was used to change an image of one object (viz. a stuffed toy, a living being and a cartoon) into another. Yamada and colleagues measured likeability ratings, and used the latency of object categorization as an index of the category membership ambiguity of the morphed images. I shall focus primarily on the results of their first two trials here.

In the first trial, Yamada et al (2013) used the morphing technique to merge static images of a stuffed Charlie Brown Toy, a composite of two Japanese male faces, and the cartoon face of Shinji Ikari\(^{109}\) from *Neon Genesis Evangelion* (ca. 1995-1996).\(^{110}\) The subjects were asked to categorize each image (as ‘stuffed’ or ‘human’ *etc.*) and to rate the likeability of the images, which were randomized for the subjects. In all three conditions, the more categorically-distinct characteristics that the images possessed, the longer it

---

\(^{108}\) The same “likeability” that Bartneck *et al* (2009) suggest in the place of “familiarity” for *shinwakan*. Yamada and her colleagues (2013) do not reference this study, but their choosing of “likeability” as a measure in these trials may be because each group of authors chose to translate *shinwakan* in the same way (all of the authors of these two studies – except for one – are native speakers of Japanese). Note that the subjects rated the likeability of the morphed images using a 7-point Likert scale (from -3 to 3). 7-point Likert scales were also used for their other two trials.

\(^{109}\) The authors are not clear about why a cartoon of Charlie Brown was not used. The appearance – and indeed the artistic style – of Charles Schulz’ characters differs significantly from Japanese *Anime*, as does the realism of Western cartoons compared to their Japanese counterparts as a whole (*e.g.* Western cartoon characters like Mickey Mouse almost always have three rather than four fingers, plus the thumb; characters in *Anime* and *Manga* always have five digits). Shinji’s face looks much more like a real human face than Charlie Brown’s, which may explain the choice: the result was likely a smoother transition between the human and cartoon images.

\(^{110}\) For each condition (cartoon-stuffed, stuffed-real, and real-cartoon) Yamada and her colleagues generated 11 images, varying the morphing percentages between 0 – 100% (*i.e.* 10% increments of morphing). This approach was repeated in all three trials.
took for the subjects to categorize them. Likeability scores were also lowest when categorization latency was highest, suggesting a correspondence between negative evaluation and categorical ambiguity.

In the second trial, Yamada et al (2013) replicated the first trial, but used images of a dog instead of a human, viz. a Beagle, a cartoon image of the Peanuts character Snoopy, and an image of a plush toy of Snoopy. They found a statistically significant relationship between categorization latency and low likeability scores, further supporting the hypothesis that categorization difficulty and negative evaluation are related even when human-like appearance is not a factor. Rather, Yamada and her colleagues suggest that it’s not degrees of anthropomorphism, but category membership (in this case something like “real thing” vs. “representation of a real thing” perhaps), that determines whether such a stimulus is evaluated negatively. I briefly mentioned this possibility earlier in Section 1.3.

To continue, in the final trial only human faces were used. Yamada et al (2013) morphed two male faces, and a male and female face. The subjects were asked to identify the images (Male A or Male B/Male or Female depending on the condition) and to rate the likeability of the images. Yamada and her colleagues found an increase in categorization latency when the images were approximately 50% morphed, but interestingly, there was no statistically significant decrease in likeability. They therefore remark that the ambiguity of “real/stuffed,” or “human/robot,” has more of an effect on categorization difficulty than e.g. “male/female” categorical ambiguity, perhaps because

111 The authors note that, in the Peanuts cartoons, snoopy is a Beagle.
“real male human” and “real female human” both belong to the category of “real human,” whereas “stuffed human” and “cartoon human” are more ontologically distinct categories.

The results of the first two trials are helpful for this investigation, but I must point out that this utility is somewhat limited. For instance, small sample sizes\textsuperscript{112} mean further research is required to determine whether Yamada et al’s (2013) findings are really meaningful. At the very least, however, their results suggest that negative evaluation is correlated with category membership ambiguity and thus with categorization difficulty. Yamada and her colleagues concluded that lack of processing fluency is the causal mechanism that underlies this negative evaluation. In other words, categorizing categorically ambiguous things is difficult. It increases the processing load if one isn’t “fluent,” \textit{i.e.} has no previous knowledge of or experience with the uncanny stimulus. This in turn, according to Yamada et al (2013), may make that object less likeable (27-29).

However, this conclusion faces a similar problem to Jentsch’s (1906/2008) theory (cf. Freud 1919/2003).\textsuperscript{113} Namely, one does not always experience either eeriness with unfamiliar things, or evaluate unfamiliar things negatively, although people do seem to have a propensity for misoneism – \textit{i.e.} a dislike of the novel and unfamiliar which Jentsch (1906/2008) discusses in the first section of his essay – when they encounter categorically anomalous/ambiguous beings. Therefore, a lack of processing fluency cannot be a sufficient condition for the kind of negative evaluation that occurs in the

\textsuperscript{112} That is, ten participants in each of their trials. It was originally 12 participants. Yamada (2013) and her colleagues had to eliminate data from two participants who did not evaluate the images on the correct criteria (23-24).

\textsuperscript{113} Detailed discussion is given to this matter in Section 2.
uncanny valley phenomenon. The problem of identifying the necessary and sufficient conditions for eeriness with human-like artifacts shall appear again from time to time as I proceed through the present section of this investigation.

Moving on, Yamada et al (2013) further conclude that human-likeness is not a sufficient condition for this negative evaluation. For the results of their second trial, contra Ramey (2005), show that difficulty categorizing a nonhuman-like set of stimuli can also result in negative evaluation. Nonetheless, the tendency for negative evaluation is probably more pronounced when the categories in question are more ontologically distinct from one another, as the results of their third trial suggest. However, one final issue is that the above study did not measure eeriness. Negative evaluation ratings (obtained using “likeability” as a measure) may be related to high eeriness ratings, but they are not measures of the same thing. Objects and/or situations that are eerie may be less likeable, but not all things that have low likability are eerie. For example, it is conceivable that someone would rate a neighbour or a co-worker whom they do not like low on a Likert-type likeability scale, but that person probably wouldn’t say the neighbour/co-worker is “eerie,” unless perhaps he/she is a “creep.” Moreover, as I noted previously (see Section 3.5), eeriness and negative evaluation may co-occur, and it’s quite obvious that not everyone dislikes feeling eerie. Some people subject themselves to eeriness and other types of fright-anxiety intentionally, say, by visiting a “Haunted House” at an amusement park, listening to ghost stories around the campfire, or by watching a horror/thriller film. Since I am attempting to identify the cause(s) of a specific negative response toward human-like artifacts, viz. eeriness, let us now consider a study that examined the relation between categorization difficulty and this emotion.
I examined the results of one of Burleigh et al’s (2013) sets of trials in Section 3.2, which sought to clarify relationship between eeriness and fear. Presently, I shall examine some of the results of their trials concerning categorization difficulty and its possible role in the uncanny valley phenomenon. Burleigh and his colleagues used eeriness as one of the measures of the subjects’ emotional responses toward the stimuli presented in these experiments (along with fear, disgust and attractiveness, as I mentioned in Section 3.2). They tested whether difficulty ascribing category membership contributes to the subjects’ sense of eeriness. To investigate this possibility, a series of images was created which shifted from a CG human face to a CG-human with salient goat-like features. They also included a featured atypicality condition, incorporating an additional salient anomaly into the morphed image, e.g. one eye larger than the other. They also controlled levels of realism by manipulating the skin colour of some of the stimuli. Like Yamada et al (2013), Burleigh et al (2013) used 7-point Likert scales in order that the subjects could report how eerie, frightening, disgusting or attractive they found the stimuli. I’ll keep my focus specifically on eeriness below.

Overall, Burleigh et al (2013) observed that feature atypicality increased the level of eeriness reported by the subjects when the CG figure had lower realism, i.e. less realistic and more obviously artificial features like strange, unnaturally coloured skin. More importantly, however, they observed that reported eeriness was highest when category membership was most ambiguous for the “goat-human” image, i.e. “approximately the mid-point of subjective human likeness,” (770). These results appear to be consistent with those of Yamada et al (2013). Even though eeriness is not the same thing as likability, a subject may tend to evaluate eerie things negatively. However, it is a
matter of further research to determine how pervasive this tendency is, for people who enjoy feeling eerie from time to time may not evaluate all eerie things negatively, as mentioned above. In any case, I think that the empirical research reviewed thus far suggests that both eeriness and negative evaluation tend to occur the most strongly at what are, ostensibly, categorical boundaries. It is beginning to appear as if a categorically ambiguous/anomalous object – what Kang (2011) calls a “categorical anomaly” – could easily fall into the uncanny valley. However, it is as yet undetermined precisely which ontological and/or psychological categories cause categorization difficulty.

For now, consider further empirical evidence (Looser & Wheatley 2012) which suggests that subjects begin to attribute categorically distinct attributes that I mentioned earlier, *e.g.* attributing animacy or aspects of mind like agency or experience (Gray *et al* 2007) to inanimate, mindless representations of people, at categorical boundaries. This is an important finding for explaining the uncanny valley phenomenon, if one chooses to appeal to the CAH. For if subjects make such attributions at what are, in fact, categorical boundaries, then this suggests once again that it is not the degree of anthropomorphism of a human-like artifact *per se* which makes it eerie. Rather, it is the ambiguity concerning which ontological categories said artifact belongs to. The study described below also lends credence to Jentsch’s (1906/2008) claim that the perception of “certain bodily or mental functions” in a human-like artifact is important when it comes to explaining why they can elicit eeriness.

Looser & Wheatley (2010) conducted a series of trials using morphed images in order to determine the point on a continuum of human-likeness at which subjects begin to attribute life, *i.e.* animacy, to a face. The images used were of mannequins, and of
humans which looked similar to the mannequins. 11 images were created of each morph (10% degrees of change for each image from 100% mannequin to 100% human). A second continuum was also parsed into 2% increments. Both continua were used in the series of trials.\footnote{Note that these were static images, not videos of animate faces as was the case \textit{e.g.} in Tinwell \textit{et al}’s (2011) study (see Section 1.3).} In the one set of trials, one group of subjects rated the images in the first set of morph continua, which were randomized. Looser & Wheatley (2010) used a 7-point Likert scale (1 “definitely alive” – 7 “definitely not alive”) for this purpose. They then asked their subjects to scroll through the second continuum to pick out the “animacy boundary,” \textit{i.e.} when the subjects judged that the image they were viewing was an image of a living being. Participants then evaluated the “pleasantness” of images from both sets of continua. The subjects were asked to return later on to judge whether the images were of something which had a mind. Here, Looser & Wheatley (2010) asked subjects whether the image was of something/one that could “formulate a plan,” “feel pain,” or that “has a mind,” \textit{i.e.} whether the image possessed agency and experience (\textit{cf.} Gray \textit{et al} 2007; Gray & Wegner 2012). Finally, they tested to see whether perceptions of animacy or mind were related to the level of realism that their subjects ascribed to the morphed images.

Looser & Wheatley (2010) found that the boundary between perceived animacy and inanimacy occurred at a consistent point on the 10% morph continuum that was slightly closer toward full human-likeness than 50% human and 50% mannequin (64% human). When the subjects freely scrolled through the 2% continuum, they likewise judged the image was animate at a consistent point closer to human-likeness than 50%
human and 50% mannequin (viz. 67% human). Looser & Wheatley (2010) did not find an “uncanny valley” in their data, and they did not measure eeriness. However, they observed that the point at which the subjects attributed animacy to the morphed images was strongly correlated with the point at which their subjects began to make attributions of mind. The subjects – by and large – began making these attributions at the ostensive categorical boundary between “animate” and “inanimate,” and we often attribute minds to things which appear animate, even if they do not possess the complex beliefs and desires we do, and sometimes even when there is no mind at all.\(^{115}\)

Looser & Wheatley (2010) conducted further trials to explore how ontological category perception might explain the above results, and to determine where exactly in the face animacy is perceived the most, \(i.e\). what are the most salient facial features when it comes to judging whether something is alive or not? For example, in another set of trials, images were randomized and then presented in pairs. Looser & Wheatley (2010) asked subjects to judge whether the pairs of images they were presented with were identical or different. Their results indicated that the subjects perceived the images as either belonging to the psychological category “animate” or “inanimate,” and that the latency for categorization increased the closer the images were in appearance, becoming strongest what Looser & Wheatley concluded was a precise category boundary, the same point at which the stimuli where perceived to be the most categorically ambiguous.\(^{116}\)

\(^{115}\) For example, \cf\. Dennett (1987) for more discussion on this propensity, which he calls the “Intentional Stance.” I shall return to the Intentional Stance in Section 5.

\(^{116}\) Looser & Wheatley (2010) also used cropped images of different areas of the face (\eg\. nose, eyes, mouth, \etc\.) to see what areas of the face subjects relied on the most when attributing animacy. They found that the subjects were disproportionately sensitive to perceived animacy in the eyes more than any other
As I mentioned a moment ago, Looser & Wheatley (2010) did not test for an uncanny valley-like phenomenon, and the “pleasantness” of the images (one of the other measures used besides animacy and mind) was found to have a linear rather than non-linear relationship to the images’ levels of human-likeness. They note, however, that some of the subjects anecdotally remarked that some of the images at or near the category boundary between “animate” and “inanimate” were “creepy or unsettling” (1860). Had they measured eeriness, they may have observed a non-linear relationship; additional experimentation is required to address this question. In any case, I want to focus on the significance of their finding that attributions of both animacy and mind occurred at a categorical boundary.

As I noted, these findings appear to support Burleigh et al’s (2013) and Yamada et al’s (2013) finding that categorization difficulty plays some role in the negative evaluation of, and/or eeriness with, human-like artifacts. They also appear to be compatible with Jentsch’s (1906/2008) conclusions concerning the psychology of the uncanny, *viz.* that emotions such as eeriness are elicited by uncertainty as to whether something is animate or inanimate, or whether it exhibits apparent “mental functions” (223). Once again, it appears as though when it comes to human-like artifacts, a subject perceives attributes such as animacy and mind, and these belong to beings from a different ontological category than that which the uncanny artifact is a member of. A subject’s perception of such an artifact thus clashes with what she knows about that kind of being.

area of the face. This finding may be consistent with Tinwell et al’s (2011), where unpleasantness ratings were highest when motion in the upper part of the face – which includes the eyes – was limited.
So far, I have been somewhat vague about how I think the psychological state associated with the uncanny valley phenomenon is constituted. I have adopted Jentsch’s (1906/2008) terminology and used “psychological uncertainty” in places because this provides some idea of the psychological state I have discussed in this section. In some way or another, a subject experiences uncertainty as to whether something (a robot) possesses something else which it should not (animacy and/or mind) according to her knowledge about those ontological categories. So now that I have discussed the emerging empirical research on categorization difficulty and the uncanny valley phenomenon, I can work toward determining what exactly the state that Jentsch of psychological or “psychical uncertainty” is during encounters with the negative uncanny, and whether or not it might be the same psychological state characteristic of encounters with categorically ambiguous human-like artifacts.

4.3: Cognitive Dissonance, Alief, and the Uncanny Valley Phenomenon

To gain a more precise understanding of the psychological state that occurs during the uncanny valley phenomenon, let’s return to Burleigh et al (2013), who make another important contribution that can be used in my theoretical overhaul of the uncanny valley. Namely, they draw an analogy between the psychological tension that categorically anomalous stimuli elicit and the psychological state known as “cognitive dissonance,” in that this tension can be likened to the psychological tension created by entertaining dissonant cognitions (761; 770-71). This analogy is similar to an earlier idea of

---

117 Cf. Festinger (1957), who introduced cognitive dissonance theory into psychology.
MacDorman et al.’s (2009b). They suggest that different areas and/or levels of mental processing, which are associated with categorically different stimuli, may be activated by “liminal objects” such as human-like artifacts. For example, a robot could be perceived as an artifact while certain salient human-like attributes simultaneously activate areas of processing associated with perceiving humans. MacDorman et al (2009b) further claim that this dissonance is “more cognitive than perceptual” (697). This remains to be determined, for as I discuss below, it is not clear exactly how well cognitive dissonance theory applies to uncanny valley phenomena cases. It may be the case that the psychological state associated with the uncanny valley phenomenon is similar to that of cognitive dissonance in that there are some cognitive or perceptual elements which co-occur, and are dissonant with each other; however, as I show below, the kind of dissonance that occurs during the uncanny valley phenomenon is quite different from that which occurs in the usual examples of cognitive dissonance.

Nevertheless, an analogy between the psychological tension caused by categorization difficulty and cognitive dissonance – with further conceptual development – is still useful for my purposes here. For it can help to contemporize the account of the psychology of the uncanny that emerged from my exegesis of Freud (1919/2003) and Jentsch (1906/2008) in Section 2, and it may also be compatible with the definition that eeriness as type of anxiety that I provide in Section 3. More specifically, Jentsch’s (1906/2008) notion of psychological or “psychical” uncertainty, both his and Freud’s
(1919/2003) remarks on the perception of animacy and the animistic worldview, and the proposal that anxiety (cf. Lazarus 1991) and eeriness (see Sections 3.3 and 3.5) are constituted in part by psychological tension and ambiguity are all compatible with a dissonance-type theoretical framework, as I show below. I’ll discuss Festinger’s (1957) theory a little further below before returning to Burleigh et al’s (2013) and MacDorman et al’s (2009b) analogy between cognitive dissonance and the psychological tension associated with categorization difficulty.

As I noted above, cognitive dissonance theory was articulated by Leon Festinger (1957). On Festinger’s theory, cognitive dissonance is a psychological state that occurs when a subject (S) experiences inconsistencies in knowledge, which results in an incongruity between what she believes and how she acts or what she does. These bits of knowledge are called “cognitions” or “cognitive elements.” I shall use “cognitions” here, although whether such “cognitions” actually count as knowledge is a matter for the epistemologists, and lies outside of the scope of this investigation. Anyhow, on Festinger’s (1957) theory, “cognitions” include all the knowledge that one has about him/herself and his/her environment (which can include the physical as well as the social environment). When these cognitions are inconsistent with one another, Festinger (1957) says they are “dissonant.” Note that Festinger also thinks cognitive dissonance is a motivational state because these inconsistencies motivate the subject to reduce

---

118 That is, provided that the psychoanalytic theme of the return of the repressed is left out when examining cases of the negative uncanny known from experience and not one uncanny that one “fancies or reads about.” Cf. Freud (1919/2003, 154).
119 An oft encountered definition of knowledge is that a bit of knowledge is a “justified true belief.” That is, S’s belief P counts as knowledge iff: S believes that P; it is the case that P is true; and S is justified in believing P (i.e. S has some evidence that it is the case that X). This definition of knowledge dates, at least, from Plato. See, for example, Socrates’ and Meno’s discussion concerning knowledge and true opinion in Meno, especially 97c-100c.
dissonance. This psychological state therefore has a significant behavioural aspect to it, which I shall return to in a moment.

Wicklund & Brehm (1976, 1-2) use the example of a fearful person who is aware that she has nothing to fear to elucidate on Festinger’s (1957) theory. Her knowledge that she is afraid, for instance, is dissonant with her knowledge that there is nothing to fear. This may sound strange, for as I discussed in Section 3.1, on Prinz’s (2004) view emotions are primarily perceptual. However even if the relational themes\textsuperscript{120} that Prinz adopts from Lazarus (1991) are cognitive or cognitive-perceptual, instead of mostly just perceptual, the phenomenology of the emotions – that is, what emotions feel like from a first-person subjective standpoint – are better described as perceptual rather than cognitive. This is because these sensations are the phenomenological aspects of the body-states associated with different emotions. If, for example, I am struck by some kind of fright-anxiety – startled by something sudden, say – the phenomenology of this experience is similar to the phenomenology of other bodily sensations like hunger, or pain; they do not “happen to a subject,” so to speak, in the same way that a thought does. They are automatic and beyond conscious control, unlike thoughts are.\textsuperscript{121} But, on Festinger’s (1957) theory, it is not the fear sensation itself, but the subject’s (S) knowledge \textit{that she is afraid} of X, which is one of the dissonant cognitive elements in Wicklund & Brehm’s (1976) example.

\textsuperscript{120} Recall that these are what tell a subject how a significant event in her environment impacts her. As I discuss in Section 3.1, on Lazarus’ (1991) view these relational themes are highly cognitive; on Prinz’s (2004) view, they are perceptual. They are what the particular body states of simple emotions represent to a subject.

\textsuperscript{121} Wittgenstein (1953/1963), for instance, returns to observations like these throughout Book I of the \textit{Philosophical Investigations}. 
So let us expand upon the example of a fearful person in order to illustrate how cognitive dissonance works more precisely, bearing in mind that it is not fear itself but rather S’s knowledge that she is afraid of X which counts as a “cognition”\(^{122}\) in this example. Say that I’m afraid of seeing the doctor, \textit{i.e.} I have a case of iatrophobia. My knowledge that I ought to see a doctor regularly to maintain good health – call this “cognition A” – is inconsistent with my knowledge that I am afraid of seeing the doctor. Let us further say that my fear of the doctor is unfounded or irrational: perhaps my fear of the doctor is rather like someone’s fear of the dentist, who doesn’t want to be poked and prodded by those horrible looking instruments; or perhaps, even though I maintain a healthy active lifestyle, I worry that the doctor will discover I have a terminal illness, even though I feel perfectly fine. Call this second cognition – the knowledge that I have an irrational fear of the doctor – “cognition Z.” In the above example, cognitions A and Z are dissonant.\(^{123}\)

Recall that on Festinger’s (1957) theory, cognitive dissonance is also a motivational state. That is because cognitive dissonance is uncomfortable. If the discomfort that S’s dissonance causes for her is strong enough, she is motivated to reduce it. In the preceding example, my cognitive dissonance might motivate me to re-evaluate either cognition A or Z. Perhaps, after careful reflection, I realize my fear is unfounded.

\(^{122}\) Here, the fear itself is a perception, following Prinz’s (2004) theory.\(^{123}\) I should note that on Festinger’s (1957) theory, most cognitions are not dissonant. Cognitions can also be \textit{consonant}, \textit{i.e.} some particular cognition implies having another. In this case cognition A – knowledge that regular medical exams contribute to proper health – is implied my knowledge that good health is desirable (call this “cognition C”). Therefore, A and C are consonant. Of course, most cognitions are simply \textit{irrelevant} to one another. For instance, my knowledge that Jones is an ideal candidate for political office is irrelevant to my knowledge that Brown is a skilled cook. Note further that the level of dissonance experienced depends upon the relevance of two cognitions to one another, and on the importance of the cognitions to the subject.
Cognition Z then becomes less relevant to A, thereby reducing dissonance. I could, alternatively, introduce cognition P: here, I form the belief that, even if I cannot find a rational explanation for my fear of the doctor, it is better to know if I really do have some sort of illness so that it can be treated rather than not knowing about it until it’s too late. On this example I might still feel fear about visiting the doctor. However, by eliminating cognitive dissonance, I may be able to motivate myself to see my doctor in spite of my fear and the dissonance between cognitions A and Z, whereas before my fear of the doctor might have kept me from seeing him despite my knowledge that I ought to do so.

The ultimate aim of Festinger’s (1957) theory of cognitive dissonance – which I think is clear from the above example – is to explain why people act contrary to what they claim to believe. Above, I maintain that I believe going to the doctor is beneficial, yet I do not do so because of my fear of the doctor. My knowledge of this fear is inconsistent with my belief that I ought to see the doctor. But if my cognitive dissonance becomes strong enough, then I re-evaluate my beliefs. This can reduce dissonance, and result in a change in my behaviour. Where I would otherwise appear a hypocrite by not going to the doctor despite my acknowledgement of the importance of getting a check-up, I might instead schedule an appointment with him in spite of my iatrophobia. So now that cognitive dissonance theory has been explained, let us return to the question of how it might apply to the uncanny valley phenomenon.

I think that Burleigh et al’s (2013) and MacDorman et al’s (2009b) analogies between the psychology of the uncanny valley phenomenon and cognitive dissonance are somewhat helpful for casting light onto the psychological state that occurs in the uncanny valley phenomenon. But the explanatory framework for the uncanny valley phenomenon
I’ve begun to sketch in this section – drawing on the Categorical Anomaly hypothesis (CAH) – is still incomplete. I have yet to find a way to account for the affective component of the uncanny valley phenomenon. In other words, as Misselhorn (2009) formulates it, how the positive emotional phenomena associated with human-like artifacts – which she thinks are best understood as empathic responses\(^{124}\) – become negative when one encounters an uncanny stimulus such as a human-like artifact. The misfiring of these empathic abilities may be a sufficient condition for eeriness. Moreover, I think the “dissonance” that Burleigh et al (2013) and MacDorman et al (2009b) reference is better understood not as resulting from clashing cognitions or perceptions of a human-like artifact. Rather, one of the dissonant psychological elements in the uncanny valley phenomenon is what Gendler (2008) calls an “alief.” By appealing to this mental state, I shall be able to solve some of the lingering deficiencies in the theory of eeriness that I am formulating here. Specifically, her concept of alief can help me to address the affective component of the uncanny valley phenomenon.

According to Gendler, an alief is “an innate or habitual propensity to respond to an apparent stimulus in a particular way.” This mental state differs in several key aspects from other mental states like belief, desire, or imagination. To elaborate a little further, on Gendler’s view aliefs are: “associative” (i.e. they encode particular response patterns to particular mental images, which are “prompted” by sense data from the external world); non-conscious and “automatic” (i.e. they operate without the subject’s conscious

\(^{124}\) As I mentioned early on, my account of the uncanny valley phenomenon differs in a number of key respects from Misselhorn’s (2009). See the above mentioned sections for a summary of Misselhorn’s hypothesis; also see Section 5 for discussion on the differences between my account of the uncanny valley phenomenon and hers.
intervention); “arational” (*i.e.* aliefs are not rational, but they are not irrational either); “shared with other non-human animals” (*i.e.* any life-form with a sensory system has aliefs); “antecedent” to other mental states (*i.e.* aliefs are more primitive than, and cannot be “factor into,” beliefs and desires); “action-generating” (*i.e.* aliefs dispose a subject to behave in a certain way, but they don’t have to generate action all the time); and finally, “affect-laden” (*i.e.* aliefs, by and large, tend have an emotional component to them) (2008, 557-58).

Gendler (2008) elaborates on alief a little further. Specifically, she identifies two sorts of aliefs: those which are “norm-discordant” and those which are “norm-concordant.” I’ll explain this is short order. Moreover, in contrast to the typical cognitivist view that treats mental states as “two-place affairs,” where a subject has a mental state, and that mental state has an intentional object – *e.g.* a subject believes that B, or hopes that H, or desires that D, or fears that F, *etc.* – Gendler maintains that aliefs include representational content, affective states, dispositions to behave in certain ways, and so forth. So when one alieves something, she does not simply alieve that A. Rather, she “alieves r, a, [and] d” (558-59). I’ll introduce a brief example below, with further examples to follow in Section 5. For now, I’ll say that by appealing to the mental state of alief instead of belief, desire, or imagination, Gendler thinks that otherwise “perplexing phenomena” such as the following example can be explained relatively simply. I think

125 Since the concept of alief is not a universally accepted one amongst philosophers, I should note that I am not very deeply committed to the concept. Alief – as a concept for describing mental states – might fail to find a place within the conceptual toolbox of the philosophy of mind; however, I think that it provides a number of advantages over Misselhorn’s (2009) concept of imaginative perception, as I discuss later on in Section 5, when it comes to explaining the uncanny valley phenomenon. Even if alief should go largely unadopted within the philosophy of mind, I still think that some sort of alief-like mental states can be used to explain the psychology of eeriness with human-like artifacts.
that the uncanny valley phenomenon is one perplexing phenomena which can be explained by appeal to alief. But before getting to that, I shall explain how alief works a little further.

Let us say that a subject is suspended in a cage over a precipice, or a deep valley. The cage has been very well secured and there is no danger of the subject falling; she is completely safe, and she knows it. But as Gendler points out, this subject would still tremble in fear at the prospect of falling. On traditional cognitivist accounts of these sorts of cases, it can be said that on the one hand, the subject a false belief that she is safe since she is apparently afraid because of her situation. On the other hand, even if she is in fact safe, the subject might continue to tell herself the cage is safe. Thus she entertains two inconsistent beliefs on two quite different grounds. Gendler calls these two assumptions “hesitation precludes belief” and “behaviour reveals belief,” respectively (561-62).

If a subject feels fear when suspended in a perfectly safe cage, and she knows that the cage is perfectly safe, then she has just had what Gendler (2008) calls a “norm-discordant” alief. That is, she ought not to be afraid in this situation given what she knows about the security of the cage, and yet she feels fear. Note that this is similar to Festinger’s (1957) theory of cognitive dissonance. On his framework, the behaviour which results from the fear a person who is suspended in a perfectly safe cage above a high precipice feels seems to be inconsistent with her knowledge that there is no danger.

---

126 Not the uncanny valley of course.

127 I just mentioned that Gendler (2008) is discusses two types of alief: norm-discardant and norm-concordant. She focuses primarily on the former. I shall also be concerned with norm-discardant aliefs, for reasons that will become clear soon.
So, this might be explained as a result of the subject entertaining two dissonant cognitions or beliefs about her situation and its impact on her welfare. But on Gendler’s (2008) framework, there is no need to say that the person in the cage has a false belief, or that she entertains two dissonant beliefs. Instead, she has a norm-discordant alief: she should not be afraid and she is aware of this fact. That she ought not to be afraid in such a situation is a norm with which here fear is dissonant or discordant. Yet in spite of this knowledge, she feels fear owing to what she alieves about being in the cage.

Following the comprehensive account of aliefs which Gendler (2008) establishes, the representational, affective and behavioural content of an alief such as that discussed above might be: “dangerous fall below”; “unnerving and/or frightening situation”; and “stay calm…don’t rattle the cage,” respectively.128 Now, Gendler, as I have noted, is primarily concerned with norm-discordant aliefs, especially where they pertain to such paradoxical phenomena as, e.g. the case of an anti-racist who exhibits an implicit bias against members of *homo sapiens* that are not a part of his or her ethnic group. I shall not be concerned with these kinds of norm-discordant aliefs here, although I am obviously interested in the aliefs subjects entertain when they encounter artifacts that resemble conspecifics.129 The reason why is that aliefs about conspecifics become discordant – *i.e.* dissonant – with norms associated with interacting with conspecifics in cases of uncanny human-like artifacts. To put it simply, the uncanny valley phenomenon is one example of the “perplexing phenomena” that are otherwise difficult to account for without making use of the concept of alief. For reasons that will become fully clear in the final section of

128 Here I am following Gendler’s (2008) style of discussing aliefs. Again, note here that a subject entertains the representational, affective and behavioural content of her alief simultaneously.
129 For discussion about this, *cf.* Sections 3 and 4 in Gendler (2008), 570-78.
this investigation, it is better to employ a dissonance-type framework that uses Gendler’s (2008) concept of alief, rather than the traditional conception of cognitive dissonance that follows from the work of Festinger (1957) or Misselhorn’s (2009) concept of “imaginative perception” (introduced in Sections 1.5-1.7).

4.4: Psychological Uncertainty Reconsidered

Following the material I’ve discussed thus far, I think that describing something as “uncanny” expresses very much the same thing as saying it is categorically anomalous, minus the emphasis on how the stimulus is processed psychologically; for as I have discussed above, categorical anomalies are anomalous because they appear to possess features which are “not-at-home” within the proper ontological category into which those stimuli fall. This is why I think that the research results concerning categorization difficulty are so useful for contemporizing Jentsch’s (1906/2008) account of the psychology of the uncanny. Indeed, the research results concerning categorization difficulty and the uncanny valley phenomenon reviewed here support the idea that categorically ambiguous or atypical features of an object/situation can elicit eeriness, and the intentional objects of the examples of psychological uncertainty that Jentsch discusses always possess what cognitive psychologists today would call categorically anomalous attributes, e.g. animacy and mind in a lifeless, mindless automaton.

Now I’d like to address a potential criticism against my reading of Jentsch before moving on. Someone might accuse me of reading the language of categorization theory and cognitive psychology into Jentsch’s (1906/2008) essay. This is not the case. Instead, I
am suggesting that Jentsch – very successfully I should add – anticipated the basic idea behind the CAH before the conceptual tools of cognitive psychology were available for him to use to describe how the psychology of the uncanny is constituted beyond just “uncertainty.” For Jentsch’s identification of the attributes of animacy and mind as the intentional objects of this “psychical uncertainty” is consistent with the data reviewed in this section, which is obtained from cognitive psychological research. It is also consistent with the empirical and theoretical material that I shall review in the following section. Finally, as I mentioned above, describing something as “uncanny” expresses very much the same thing as does calling it a “categorical anomaly,” sans emphasis on the cognitive and perceptual processes which underlie these experiences. In both cases, the uncanny object appears to possess features or attributes which are out of place given what people typically expect, believe, or know about that kind of object.

So much for this potential criticism of my reading of Jentsch (1906/2008); now, I shall address another problem with the CAH. As the results of Yamada et al’s (2013) third set of trials demonstrate (see Section 4.2), categorically anomalous stimuli do not always elicit eeriness, just as uncanny objects/situations do not always elicit a negative response.130 So if cognitive dissonance of one type or another occurs during categorization difficulty, then neither of these things, by themselves, completely explains why eeriness occurs in the uncanny valley phenomenon. This is further evidenced by Kang’s (2011) observation that, although automata and robots are “transcategorical,” they are also often objects of fascination and curiosity rather than of eeriness. To summarize:

130 See my discussion of the positive and negative uncanny in Section 2.
categorization difficulty and/or some kind of cognitive dissonance might be necessary conditions for eeriness, but it’s pretty clear that they are not sufficient conditions.

The above discussed problems are similar to the problem that Freud (1919/2003) thinks Jentsch’s (1906/2008) theory of the psychology of the uncanny suffers from (see Sections 2.2-2.3). Namely, if uncertainty and/or unfamiliarity are Jentsch’s essential conditions for a subject to experience an uncanny response to something that is unfamiliar or psychologically disorienting, then why does she not feel an uncanny sensation in each and every sentence that she feels uncertain in this way? Recall that I don’t think this criticism against Jentsch is a very good one, for Jentsch specifies two particular intentional objects of this state of psychological uncertainty which I’ve discussed at length and have just mentioned above, viz. animacy and mind (123).

I think that the reasons why these particular attributes make something seem eerie, as I’ve mentioned at various points herein, are probably due to the evolutionary importance of being able to successfully detect them. That is, it is important to know what is alive and what it wants from you, or what it its intentions are for you (see Section 3.5). It is so important that we regularly misattribute intentions and other mental states to things which don’t have these things. Human-like artifacts cause these capacities to misfire somehow, and I think this is another case of misattributing animacy and mind that we regularly make toward other inanimate things in the physical world, i.e. taking something like the “animistic worldview” discussed by both Jentsch (1906/2008) and Freud (1919/2003). The perception of certain of these artifacts’ human-like features

---

131 Cf. Dennett 1989; see also Section 5 for detailed discussion.
might, as MacDorman *et al.* (2009b) suggest, register as features of something which *e.g.* is alive, or is a human, if the artifact in questions is at or near a categorical boundary such as “animate/inanimate,” “human/not human,” *etc.* In fact, as Kang (2011) argues, they can also violate more than one category simultaneously. These categories may be binary or non-binary (see Section 4.1).

I think the capacities for detecting animacy, and mental/emotional states in others – when they misfire under certain conditions which I shall articulate fully toward the end of Section 5 – provide the means for completing the picture of the uncanny valley phenomenon. Now, I am nearly finished revising Mori’s (1970/2012) uncanny valley theory. In the final section, I return to Misselhorn’s (2009) explanation of the uncanny valley phenomenon, as well as to the discussion of empathy and the uncanny valley, and to a few more of Gray & Wegner’s (2012) empirical research results. Misselhorn (2009) – as I discussed in Sections 1.5-1.7 – tries to account for the uncanny valley phenomenon using something she calls “imaginative perception.” While it might be possible to explain the affective component of the uncanny valley phenomenon using this concept, I think a clearer and more parsimonious theoretical model can be constructed using Gendler’s (2008) concept of norm-discordant alief. I’ll explain exactly how I think aliefs play a role in the uncanny valley phenomenon in the following section. But even though my account of the uncanny valley differs from Misselhorn’s (2009), I can still use my theory to show – as she likewise attempts to – that empathy ought to replace *shinwakan* in Mori’s (1970/2012) theory of the uncanny valley. Once I’ve done so – thereby completing my conceptual overhaul of the uncanny valley theory – I shall also be able to explain the uncanny valley phenomenon.
5.0 – Of Minds and Machines

But can’t I imagine that the people around me are automata, lack consciousness, even though they behave in the same way as usual? – If I imagine it now – alone in my room – I see people with fixed looks (as in a trance) going about their business – the idea is perhaps a little uncanny. But just try to keep hold of this idea in the midst of your ordinary intercourse with others, in the street, say! Say to yourself, for example: “The children over there are mere automata; their liveliness is mere automatism.” And you will either find these words becoming quite meaningless; or you will produce in yourself some kind of uncanny feeling, or something of the sort.


Over the course of the previous sections, I’ve attempted to provide a general idea of how the psychological state that is characteristic of eeriness with human-like artifacts is constituted, viz. by some type of psychological dissonance or discordance that occurs when one encounters a categorically anomalous human-like artifact. However, my explanation of the uncanny valley is, as of yet, incomplete. I have yet to explain how the positive and negative emotional responses toward human-like artifacts are related to one another, and how the perception of apparent aspects of mind factors into the uncanny valley phenomenon. But since I have obtained a better idea of the kind of emotion that eeriness is in Section 3 – specifically, a species of anxiety – I am now much closer to achieving my overall objective.

Thus far I have focused primarily on one particular uncanny saliency, viz. apparent animacy in an inanimate object. As far as human-like artifacts e.g. robots and automata are concerned, animacy is uncanny insofar as such artifacts are inanimate objects. In other words, the perception of animacy is “not-at-home” in an artifact that a subject knows or believes is inanimate. This is one of the things which make them
categorically anomalous, but again, it only provides part of the picture of the psychology of the uncanny valley phenomenon. Categorical ambiguity might be a necessary condition for being the cause of a subject’s sense of eeriness, but it is not a sufficient condition.

At the end of the previous section I gestured at a possible means of addressing this problem. Recall that Jentsch (1906/2008) thinks “bodily and mental functions” are what give rise to a subject’s experience of an “uncanny sensation” (223) where human-like artifacts such as wax figures or automata are concerned. Indeed, when it comes to the animistic worldview, which both Jentsch and Freud (1919/2003) discuss, they are not simply talking about a universe that seems somehow alive, but that also has intentions, can carry out actions, etc. (see Sections 2.2-2.3). Jentsch’s (1906/2008) dryads who live the trees (225), as it were, are not just alive – they also appear to possess agency. Recall further that according to Gray et al (2007) agency is an attribute of mind, along with experience. I think much of the material I have discussed in the preceding sections of this investigation is illustrative of the fact that a human-like artifact can appear to possess both animacy and agency. According to Gray & Wegner (2012), they may also appear to possess the capacity to experience, and this may be an important contributing factor to eeriness and the uncanny valley phenomenon.

For the reasons outlined above, in this section I shall concern myself with the perception of mind – and the capacities for affective and cognitive empathy, i.e. the capacities used to understand other minds – toward the end of accounting for the affective

132 Review Kang’s (2011) remarks in Section 4; see also my remarks in Section 4.4.
133 Emphasis added.
component(s) of the uncanny valley phenomenon. In other words, I’ll explain the very same thing that Misselhorn (2009) attempts to explain. Namely, how/why the positive, empathic responses toward human-like artifacts become negative when such artifacts become categorically ambiguous/anomalous, or “uncanny.” But instead of appealing to her concept of “imaginative perception,” (see Sections 1.5-1.6), I shall argue that the uncanny valley phenomenon is ultimately attributable to beliefs or “cognitive elements” about human-like artifacts that are dissonant with what Gendler (2008) calls “aliefs.” As noted in Section 4.5, an alief is “an innate or habitual propensity to respond to an apparent stimulus in a particular way” (557). For now, I shall say that the account of the uncanny valley phenomenon that I present below posits that human-like artifacts are capable of eliciting aliefs that ought to be elicited only by conspecifics. These are dissonant with a subject’s knowledge and/or beliefs about the ontological nature of the human-like artifact. The uncanny valley phenomenon therefore results from the earlier discussed “norm-discordant aliefs” which a subject has when she interacts with human-like artifacts (see Section 4.4). The result of this discord – caused by the categorically ambiguous features of the artifact – is a sense of eeriness. Of course, once I have explained all of this fully, I can show conclusively why “empathy” – both the affective and cognitive variety – ought to replace “familiarity”/“affinity,” i.e. shinwakan, in Mori’s (1970/2012) theoretical framework (see Section 1 for discussion).

5.1: Perceiving Artifacts with Minds

What happens when a subject perceives that a human-like artifact possesses a mind, or certain attributes thereof? First off, I’m not suggesting here that the artifact
really does have a mind and that this is what the subject perceives, for artificial intelligence (AI) is simply not advanced enough at present that engineers can make a robot that is human-like in appearance and intelligence. But just as it’s conceivable that a machine that cannot really think could fool a human into attributing thought to it – thereby passing the so-called “Turing test” (cf. Turing 1950) – it’s also conceivable that a robot with sufficiently life-like human features could fool a subject, in some sense or another, into perceiving that it is alive. If it has certain kinds of apparent motion, like goal-oriented movements, then this might cause a subject to attribute intentions, goals – and thus a mind – meaning that this robot would also pass the Turing test in a certain sense.

Recall the results of Gray & Wegner’s (2012) first trial, which I discussed in Section 1.2. Here, the subjects who watched the videos of the robot Kaspar made greater attributions of experience to the human condition – which depicted it from the front showing a human-like face – and approximately equal attributions of agency to the human and mechanical conditions. They also found the videos of Kaspar in the human condition caused them greater discomfort. Note that in the mechanical condition, Kaspar was turned around to reveal its wires and circuit boards. Gray & Wegner (2012) conducted several further trials to see how the perceptions of these attributes can explain eeriness with human-like artifacts, which I shall now review.

In another set of trials, perceptions of both agency and experience were manipulated. The authors gave two descriptions of a supercomputer that they called the “Delta-Cray supercomputer.” In one condition it was simply described as a powerful computer. In the other, however, subjects were told that it could experience sensations
like those that a human can, such as pain, hunger, thirst, etc. The subjects were then asked to rate their perceptions of experience, agency and unnervingness using questionnaires. Gray & Wegner (2012) found that their subjects rated the human-condition description of the computer as more unsettling, and that perceptions of experience were more salient to the subjects than perceptions of agency. In yet another set of trials, Gray & Wegner (2012) showed their subjects a picture of a man instead of describing a sentient computer to them. In the first condition in this trial, he was described as normal. In the second, he was described as lacking either agency or experience. Subjects were then asked to rate how “unnerving” they perceived the man in the photograph using Likert scales. The results were similar to the first two trials that I discussed. Here, subjects reported that perceiving a human who lacks experience was more uncomfortable than perceiving a human who lacks agency. Gray & Wegner (2012) think these results evidence the hypothesis that perceptions of experience play a much larger role in the uncanny valley phenomenon than perceptions of agency do.

I think that caution is required if I am to use Gray’s & Wegner’s (2012) empirical results. In their trials the subjects weren’t actually interacting with a human-like artifact/human, but were simply given a description/representation of one. The results might have been different had a human actor been used instead of simply a photograph in the third set of trials, for instance. The authors acknowledge the narrowness of the stimuli used in their trials and are of course optimistic about future research. In any case, Gray & Wegner (2012) ultimately conclude that attributing experience to things which characteristically do not have experience – or, perceiving a lack of experience where such a lack is unexpected – may be a factor in the uncanny valley phenomenon. This appears
consistent with the Categorical Anomaly hypothesis (CAH) discussed in the previous section. That is, the perception of attributes like the capacity to experience in a mindless, inanimate object makes that object categorically anomalous; or in other words, “uncanny.”

Categorization difficulty – and possibly a negative evaluation like eeriness – would be the result of perceiving such an artifact, as I discussed in Section 4. It also seems to be compatible with Jentsch’s (1906/2008) observations on animacy, mental functions, and the cause(s) of uncanny sensations, which I discuss in Section 2.

Gray’s & Wegner’s (2012) conclusion turns on the fact that, as they point out, “higher cognition” has been regarded as an essential human characteristic since at least the time of Aristotle, who famously thought man was a “rational animal.”

The unease associated with human-like artifacts is, then, a result of perceiving apparent characteristics where they don’t usually appear, or where they are not expected to be present (129). As Gray & Wegner (2012) put it:

[B]ecause humans are fundamentally expected to have experience, a person perceived to lack experience […] should be seen as unnerving. This would not only explain the unnervingness of those who seem incapable of fear or love (e.g., psychopaths), but also the strangeness of the idea of philosophical zombies – otherwise normal people without conscious experience. Thus, the uncanny valley may apply to both humanlike robots and robotic people (126).

It is debatable whether “robotic people” are always unnerving in this way. However despite the limitations I discussed above, I think these results are encouraging, and

---

134 “Uncanny” in the sense that I sketched out in Section 2.

135 I am not trying to sound exclusionary here. Aristotle thought that older men possess the most reason. Women, children, slaves and “barbarians,” on his view, are less rational. Thankfully, times have changed.

136 Take psychopaths for example. Familiar examples from fiction like Hannibal Lecter are quite unnerving to be sure. But real psychopaths aren’t very unnerving when one interacts with them. According to Hare (1993), a subject typically has no idea that she is in the presence of a psychopath, for psychopaths are often
support the suggestion that perceiving some aspect of mind where it shouldn’t be present can elicit eeriness. But I must point out that it is possible that agency plays a stronger role in eeriness with human-like artifacts than Gray’s & Wegner’s (2012) results suggest. Recall, for example, the animistic worldview (see Section 2). Here, it seems that the world itself – or the gods and spirits that are supposed to populate it – possess agency. Recall further that Gray et al (2007) found that their subjects attributed a large amount of agency, but little experience, to God.\textsuperscript{137} It seems as though this might also apply to gods, spirits, demons, ghosts, and other such imaginary entities. If one encounters the negative uncanny in everyday life – e.g. one has an encounter with something that he or she deems paranormal – then apparent agency may play a larger role than apparent experience.

Of course, this is an empirical question, and further research is required to determine the extent to which perceptions of agency and experience play a role in encounters with the negative uncanny that do not involve human-like artifacts. However, it seems to be the case that perceptions of these attributes play some role in the psychology of the uncanny in a general sense, although Gray & Wegner (2012) might be correct that perceived experience plays a greater role in eeriness with human-like artifacts than perceived agency, as I discuss further on. It may also be possible to mitigate the limitations of Gray’s & Wegner’s (2012) empirical results appealing to the similar conclusions of Jentsch (1906/2008) and Freud (1919/2003) about the animistic worldview (see Section 2.2).

\textsuperscript{137} Specifically, the large, bearded man in the sky found in the holy books of Judaism, Christianity and Islam.
In any case, Gray’s & Wegner’s (2012) empirical results suggest that, although human-like artifacts like Kaspar may not completely pass the Turing test, they do seem capable of convincing a human, in some sense, that they possess attributes of mind. This is also compatible with the empirical results I reviewed in Section 1.5, where subjects felt compassion for artifacts that were tortured in a Milgram-style experimental design or otherwise mistreated, even though they knew that this response was unwarranted (cf. Bartneck et al 2005; Slater et al 2006; Rosenthal-von der Pütten et al 2013a; 2013b). Here, the capacity to experience pain was clearly attributed to the artifacts in these experiments. This is why the subjects were able to show compassion for them. The capacity for this kind for experience, however, is out of place within the ontological categories that human-like artifacts inhabit. The subjects in the above-mentioned experiments did find report that they found these artifacts uncanny or eerie; however, the fact that subjects showed compassion demonstrates that subjects do make attributions of experience to human-like artifacts. Now, I shall begin to explain why attributing experience to things which are not known and/or believed to have experience elicits eeriness. For this capacity being out of place in such an artifact does not completely explain why human-like artifacts elicit eeriness.

5.2: Knowing Other Minds

The above-described mind perception – i.e. perceiving attributes like agency and/or experience (cf. Gray et al 2007; Gray & Wegner 2012) – is so reflexive that one has to make a conscious effort not to exercise this ability, whether it is with conspecifics or with highly realistic human-like artifacts. In fact, it is difficult to imagine that
successfully doing so is even possible. It would, for example, be analogous to making a conscious effort not to understand the words on this page that you are reading now if you are a native/fluent speaker of English.\textsuperscript{138} This is probably because the automatic and unconscious perceptual mechanisms used for interaction with conspecifics are the same that are activated by perceiving a human-like artifact. Think here of Fodor’s (1983) theory that the “input systems,”\textit{i.e.} our sense modalities, are modular; two of their distinguishing properties are informational encapsulation, and domain specific. Moreover, these information processing systems operate without conscious direction.\textsuperscript{139}

The mental mechanisms used for perceiving humans might be “fooled” by the inputs from a sufficiently realistic human-like artifact.\textsuperscript{140}

Similarly, when it comes knowing other minds it’s easy to confine oneself to the armchair and imagine that the people passing by outside of one’s window are mere automata, as René Descartes famously does in the\textit{Discourses} (1637/1968) and in the

\textsuperscript{138} If the reader wishes, he or she should try this exercise for him/herself before reading any further.

\textsuperscript{139} More specifically, Fodor (1983) proposes that these input systems possess a number of properties that are not common to higher-level cognitive systems (47). These properties are: (1) domain specificity (each module has a specific domain of inputs that it can process); (2) mandatory operation (modules process inputs reflexively/automatically); (3) limited accessibility to central processing systems (one only has limited conscious access to the representations that modules compute); (4) speed (input systems compute their outputs quickly); (5) informational encapsulation (processing in a module is not affected by information from higher levels of processing); (6) “shallow” outputs (the outputs of modules are simple); (7) fixed neural architecture (modules are associated with specific structures/areas of processing within the brain/nervous system); (8) specific and characteristic breakdown patterns (modules malfunction is particular ways,\textit{viz.} specific neurological disorders); and finally, (9) they are determined ontogenetically (modules have their own characteristic ontogenetic pace and sequencing). He also thinks that our language processing centres are modular.

\textsuperscript{140} Sperber & Hirschfeld (2007) propose that modules have “proper” and “actual” input domains. The input systems used for perceiving humans have human features as their proper domain of inputs; however, their actual domain of inputs might include inputs from artifacts which are very human-like. Sperber & Hirschfeld use the example of a “snake detection” module that can be “fooled” by a stick which resembles a snake. I do not think that we have such a module, for either snake or conspecific detection. That is, I’m not attempting to argue that the mind is massively modular, a view to which the aforementioned authors subscribe. However, whatever modular inputs systems are involved with perceiving humans can probably be fooled in the way I’ve described above, and can therefore be thought of as having “proper” and “actual” input domains.
Meditations (1641/1968). But as Wittgenstein (1953/1963) keenly points out, this exercise becomes quite absurd during one’s “everyday intercourse with others” (§420). However we come to know that other minds exist, the obviousness of their existence is such that imagining that a human that does not have a mind – e.g. imagining the “robotic people” mentioned by Gray & Wegner (2012) – is a very uncanny exercise indeed. For a mind is “at-home” within a human being, and we seem to attribute minds to other humans automatically. It is so automatic a tendency that when we encounter someone who lacks some aspect of mind that is shared by others, we sometimes find this strange, such as the man in the photograph who lacked experience in Gray’s & Wegner’s (2012) empirical trials. The above observations can help me to hone in on the manner in which mind perception functions – and how it fails – with human-like artifacts.

One way to articulate the nature of this function and failure – and to overcome the limitations of the Categorical Anomaly hypothesis and the perceptual-cognitive dissonance framework introduced in the previous section – may be to borrow some of the ideas found in theories of “Folk Psychology.” According to Crane (1995/2003), folk psychology – sometimes called “common-sense psychology,” “cognitive empathy,” or “theory of mind” – is one answer to the so-called “problem of other minds.” The problem of other minds asks: how does one know that other minds exist? Moreover, how does one know what other minds believe, desire, feel, think, etc.? “Folk psychology” refers to the common-sense set of principles that people use to know about other minds in their day-to-day transactions with other people.

There are a number of different possible ways in which we employ folk-psychological reasoning. One possibility is that folk-psychological reasoning works like a
theory; here, a subject forms a “theory” about what an agent is going to do in a certain situation, or why she expresses a certain emotion in a given situation, etc., based upon his observable behaviour, and what the subject knows about how her own behaviour is causally connected to her thoughts. This is often called “Theory theory.” Another possibility is that a subject “simulates” by imagining herself in another’s situation, and then predicts what the agent will do, how he feels, what he is thinking, etc., given what she knows about how she might think/feel in his situation. This is known as “Simulation theory.” I, however, shall appeal to a different account of folk-psychological reasoning for the purposes of this investigation. I don’t wish to outright reject either the Theory theory or the Simulation theory, but I also do not have the space – nor is it the object of this inquiry – to determine which of these theories, or what possible combination of the two, better accounts for folk-psychological reasoning.\textsuperscript{141} I simply think that this alternative – namely, Daniel C. Dennett’s concept of the “Intentional Stance” – may be more easily applied to explaining empathy with human-like artifacts, at least for the present time.

This concept is part of a larger explanatory framework that very easily applies to objects such as human-like artifacts, as I shall make clear below. The intentional stance, according to Dennett (1989), is the level of abstraction that a subject adopts in order to determine the beliefs, desires and behaviours of other rational agents. On Dennett’s theory, a subject decides to treat the object in question as a rational agent, and then determines what beliefs and desires it ought to have given what kind of thing it is, and what its place in the world is. Usually, this stance is adopted because the stimulus in

question appears in some way to be animate, and further, because its movements appear to us as though they are goal-/intention-oriented. That is, the stimulus appears to have intentions. Of course, to behave this way, that thing must have a mind. Thus the subject treats the object in question as an autonomous, rational agent. The subject then predicts what the agent will do based on that knowledge. By doing the above, one applies what Dennett calls the “Intentional Strategy” (17-18).

Dennett (1989) thinks that some of our folk-psychological abilities are probably given an “innate perceptual boost” (9). In other words, the ability to understand what other minds are like – whether this is done by theorizing, simulating or adopting the intentional strategy – is a lot like the ability to acquire language. Certain noteworthy exceptions, like many individuals on the Autistic spectrum (cf. Baron-Cohen, Leslie & Frith 1985) aside, we are generally equipped with the mental mechanisms necessary to pick up the above described cognitive-empathic skills.142 Furthermore, it is probably the case that some of our folk-psychological intuitions are generated at an automatic level of processing, without the conscious direction of the subject. This appears to be compatible with the general consensus in psychology that many of our social cognitive abilities are automatic and do not require the conscious direction of the subject. In other words, the computational mechanisms that underlie these abilities can be thought of as what Bargh & Chartrand (1999) call “mental butlers,” i.e. they process information so as to leave higher cognitive abilities free to cope with novel stimuli and situations. This automaticity might also explain why, according to Dennett (1989), we adopt the intentional stance toward things that aren’t really “Intentional Systems,” or “True Believers,” i.e. rational

142 Cf. Chomsky (1966/2009) for further discussion on the topic of language acquisition.
agents whose behaviour can be predicted. For example, people often adopt the intentional stance when they speak of apple trees which “want” to blossom in the spring, or a bold of lighting that “wants” to find the fastest way to the ground, etc. (22).

I think that when a subject attributes a mind – or some particular aspect of mind – to a human-like artifact, what is happening can be thought of as a misapplication of the intentional strategy. As I just noted, Dennett (1989) thinks that such misapplications happen all the time. We often speak of inanimate objects or natural phenomena as if they have beliefs and desires, even though we’re well aware that they don’t. However, certain of a human-like artifact’s features – if sufficiently realistic – may cause a subject to treat it as a rational agent with a mind even if the subject is quite aware that this is absurd, since the processes that underlie these abilities are by and large unconscious and automatic. The human-like features that cause a subject to misapply the intentional strategy in this way are, as the reader may have guessed, realistic human-like appearance, apparent animacy and apparent mentation. Where a subject would normally adopt the “Design Stance” – the level of abstraction that Dennett (1989) thinks she adopts in order to understand designed artifacts and features of the natural world which are “designed” by the process of evolution by natural selection (16-17) – she instead adopts the intentional stance, because human-like artifacts are sufficiently anthropomorphic so as to “trick” her into taking this stance even though she knows better. This is a rather straightforward way of explaining how human-like artifacts pass the above-mentioned Turing test using the language of folk psychology.

143 My use of “trick” here is yet another example of misapplying the Intentional Strategy, and of how pervasive this misapplication can be.
I think this provides a rough idea of how mind perception and human-like artifacts might work, but it’s possible to explain this phenomenon with even greater precision. For example, it is possible to isolate a neurological causal mechanism for the perception of mind – in addition to perceptual animacy – which is compatible with what I’ve argued above. MacDorman et al (2009b) suggest that cognitive empathy is possible with human-like artifacts because of the Mirror Neuron System (MNS) (see Section 1.4-1.5). Mirror neurons are neurons which fire when an action is being performed, as well as when an action is observed. According to Iacoboni (2009), the MNS is another answer to the aforementioned problem of other minds: that is to say, because the same neurons fire when an action is performed, and when the same action is observed, a subject is able to represent the contents of other minds; i.e., their goals, beliefs, desires, etc., at the neurological level. MacDorman et al (2009b) further suggest that the MNS may be activated by perceived robotic actions as well, and this appears to be supported by preliminary fMRI studies. For example, they cite Gazzola et al (2007), who found that the MNS was activated when their subjects watched a robotic arm performing simple actions such as picking up and moving around objects. Even though the robotic arm was not very human-like (it resembled the arm of an industrial robot), the subjects apparently represented its movement in terms of goal-oriented action at the perceptual level, as indicated by activity throughout the MNS. Considering these empirical results, it seems reasonable to speculate that the MNS should be activated by the perceived actions of a very human-like robot, and not just an industrial-looking one.

Saygin, Chaminade, Ishiguro, Driver & Frith (2011) examined the neural correlates of perception of human and robot motion in the APS, or “Action Perception
I’ll discuss their empirical results in rather full detail, for they are quite important. Here, “action” is moving with the intention of fulfilling a goal or accomplishing a task, *i.e.* acting with intention. The Action Perception System, as the name suggests, includes brain structures associated with the perception of action, *i.e.* the lateral temporal, anterior intraparietal, and inferior frontal/ventral premotor cortices. According to Saygin *et al.* (2011), areas in the APS are also associated with the MNS. By examining the level of activity in the APS when perceiving the actions of an android, a human and a robot, Saygin *et al.* (2011) sought to uncover how the APS processes the perceived actions of different human-like agents. To investigate the selectivity of the APS on the appearance and motion of these agents, they used fMRI neuroimaging to examine their subjects’ APS’ while they viewed videos of a robot, an android, and a human performing simple motor tasks like raising a hand in greeting. The robot condition had mechanical appearance and movement; the android condition had biological appearance and mechanical movement; and the human condition had biological appearance and movement.

Saygin *et al.* (2011) found an increased level of activity in the APS of subjects in the android condition, but that the robot and human conditions both instantiated levels of APS activity similar to one another. Saygin and his colleagues therefore concluded that the robot’s and the human’s appearance corresponded to what this part of the brain predicted or “expected” their actions would be like. But in the case of the android, the

---

144 Specifically, the study used video of Repliee Q2 and the human after whom she was designed. For the human and android conditions, the human and android were used, respectively. For the robot condition, Repliee Q2’s clothes, hair, and silicon skin were removed, and her face covered by a metallic mask (Saygin *et al.* 2011, 3-4).
perceived biological appearance was inconsistent with the perceived mechanical movement. Here, the APS tries to predict what the android will do based on its human appearance, but accomplishing this increases the processing load, since the behaviour of the agent does not meet the subjects expectations. Saygin et al (2011) conclude that activity in the APS increases when the brain tries to interpret the conflicting perceptions of mechanical movement and human-like appearance.\(^{145}\) To put it very simply, certain salient features – type of motion in this case – clash with what the brain normally “expects” the human – which is really an android – to do.

Saygin et al (2011) are not directly concerned with testing the predictions of the uncanny valley theory. Nevertheless they offer the following suggestion to help explain the uncanny valley phenomenon based on their findings: when the brain is presented with a stimulus which isn’t predictable because its features are categorically ambiguous, like an android, the above-described “predictive-coding errors”\(^{146}\) can occur in the APS. Burleigh et al (2013) suggest that Saygin et al’s (2011) findings might be consistent with their categorical anomaly hypothesis (discussed in Section 4.2); however, they maintain that caution is required here, for it is not clear that a mismatch between the perception of the android clashed with a single neural schema for “human” and “robot,” or many related schemas which are all involved in one aspect of human/inanimate object perception or another (761). But whether this conflict occurs between one or multiple

\(^{145}\) This phenomenon may also occur the other way around. That is, if the android’s movement is more salient than its appearance, the “expectations” about what the agent will be based on its movement, and further will be dissonant with the perception of its human-likeness. This is one point that highlights the need for additional research, and for being cautious about what conclusions are drawn from these data until further studies are completed.

\(^{146}\) In this case, failing to predict how the android will behave because the brain can’t properly categorize it as either a human or a robot.
neural models of different categories is a matter for further study. For now, I think there are a number of helpful insights to be found in the results of the fMRI research I have just discussed. Mainly, they are illustrative of the fact that humans often attribute intentions to things which move as if they are capable of having the same kinds of intentions that a human being does. This provides a neurological causal mechanism behind the misapplication of what Dennett (1989) calls the intentional stance, where the design stance is really the ideal stance for a subject to take to understand the artifact.

Together with the material reviewed in Section 1.5, the material I’ve discussed above illustrates that humans have a capacity for empathy with human-like artifacts. This includes the capacity for cognitive empathy or folk psychology, in addition to affective empathy. Now let us briefly return to MacDorman et al. (2009b). They do not follow up on empathy’s relation to eeriness in the uncanny valley phenomenon very thoroughly, even though they are among the first to suggest that empathy and cognitive dissonance might play some role in it. However, their suggestion that the MNS – and also, in light of what I’ve considered above, the APS – facilitates cognitive empathy with human-like artifacts may have more explanatory power than they realize. Indeed understanding how these empathic responses fail, as I am about to show, can explain the changing relationship between the positive and negative emotional responses that subjects show toward human-like artifacts. To explain this failure precisely, I shall incorporate Gendler’s (2008) concept of alief (see Section 4.5) into the framework for explaining the psychology of the uncanny valley phenomenon that I am developing. I can

---

147 Again, the artifact has neither emotions nor mental states, so a human who “empathizes” doesn’t truly empathize; rather, it is merely as if the subject empathizes with an artifact that apparently has emotions and/or mental states.
then explain why I think this approach is superior to Misselhorn’s (2009) “Imaginative Perception” hypothesis (outlined throughout Sections 1.5-1.7).

5.3: “Re-imagining” the Imaginative Perception Hypothesis

I have primarily limited myself to cognitive empathy in this section, whereas in Sections 1.5-1.7, I am largely concerned with affective empathy. Recall further my summary of Misselhorn’s (2009) Imaginative Perception hypothesis from earlier: she thinks that the misfiring of empathy might explain the phenomenology of the uncanny valley, *i.e.* eeriness. On her view, this psychological state is constituted in part by rapid oscillation between activation and deactivations of concept(s) associated with G caused by salient similarities between G and F. She likens to oscillation to a gestalt switch, except that with human-like artifacts, this psychological switching back and forth persists, and occurs very rapidly. But as I discussed earlier – and as I discuss further below – I don’t think that empathy misfires in exactly this way, although it should be obvious that I do think that the misfiring of human empathic abilities plays a significant role in the uncanny valley phenomenon.

One reason why I shall not appeal to the Imaginative Perception hypothesis here is that I do not think the gestalt switch-like oscillation that Misselhorn (2009) describes is what takes place in a subject’s mind when she encounters an uncanny human-like artifact. I also think that there is a more parsimonious way of explaining the uncanny valley phenomenon. Namely, employing Gendler’s (2008) concept of norm-discordant alief – discussed in Section 4.3 – is just as useful as Misselhorn’s concept of imaginative perception for my purposes, but it does not require such an unorthodox use of the concept.
of “imagination.” In other words, Gendler’s (2008) concept of alief offers a clearer and simpler way to describe the psychology of the uncanny valley. It also allows me to incorporate the previously discussed cognitive dissonance – provided of course that it is understood that this dissonance is not quite like that described by Festinger (1957) – into the theory of eeriness that I am developing. Let me begin to explain by discussing just a few additional empirical findings from android science research.

As Ishiguro (2006; 2007) points out, some human-like artifacts are able to “fool” human subjects into thinking that they are human, although the subjects in his empirical trials aren’t usually fooled in this way for very long. Ishiguro typically manages to fool his subjects for approximately a second or two using his android designs, such as those used in Bartneck et al’s (2009) and Saygin et al’s (2011) empirical trials. In a typical experiment, either Ishiguro or his robotic doppelganger waits behind a curtain. A subject then enters the room and interacts with an experimenter. While this takes place, the curtain opens to reveal either Ishiguro or his robot, depending on the experimental condition a subject is assigned to. The subject is later asked questions about what he or she noticed after the curtain was opened. Typically, subjects report that they become aware that the thing behind the curtain is a robot very quickly, which is corroborated by eye movement data and other physiological measures. Completely fooling a human for sustained periods of time means that a robot passes what Ishiguro (2007) calls the “Total Turing Test,” i.e. the robot’s appearance and movements, as well as its (apparent) mind, would fool the human observer into believing the artifact is a real human being, unlike in Turing’s (1950) “imitation game” where the computer has no human-like appearance, and is hidden from the view of the other players in the game, using only its intellect – or
the programmed semblance thereof – to convince a human player that it has a mind and is capable of thought.\textsuperscript{148}

To date, no robot or other such human-like artifact exists which can pass Ishiguro’s (2007) Total Turing Test. But I have shown above that some human-like artifacts – in part by virtue of their apparent animacy and mentation – can sometimes convince a subject on an unconscious level that they do have minds, emotions, mental states. Ishiguro (2006; 2007) himself suggests that a sufficiently realistic robot should continually activate the input systems typically employed during our interaction with other humans,\textsuperscript{149} although he cautions that such subconscious recognition is purely hypothetical at this point (326-31). But it seems clear from the data I’ve reviewed here – which has been obtained since Ishiguro proposed this hypothesis – that human-like artifacts are perceived using the very same input systems used for humans and other animate, autonomous agents that appear to have minds, as was the case in Saygin et al’s (2011) study of the perception of android motion and activity in the APS, and Gazzola et al’s (2007) study of robot action and the MNS. The computational processes in these perceptual or “input” systems, as Fodor (1983) calls them, are not affected by higher-order cognitive processes because they are, as he says, “informationally encapsulated.” That is to say, information from the same level of processing, or a higher level, cannot

\textsuperscript{148} Alan Turing (1950) proposes that if a computer can fool a human into thinking the computer is another human, then for all practical purposes, the computer can be said to “think.” To see if a computer can do this, a subject can play “The Imitation Game,” where the computer tries to convince a human player that it is a human and not a computer by answering the human player’s questions, e.g. “What length is your hair?” “Are you a man or a woman?” or “What is 1768 x 9145?” (Note that this last question could easily give the computer away – in such a case the computer would have to “play dumb.”) Yet another human participant plays the game as well; her role is to convince the first human player that she is the real human and that the computer is just a computer.

\textsuperscript{149} Ishiguro (2006; 2007) doesn’t use this term. I have employed it here following Fodor’s (1983) terminology.
affect the perceptions which these mechanisms process. This means that even if one knows the robot is just a robot, this does not affect the phenomenal feel that perceiving it takes on owing to its similarities to a human being. The phenomenal feel that perceiving a human-like artifact elicits is therefore bottom-up rather than top-down. This appears to be consistent with Misselhorn’s (2009) observations concerning the phenomenology of human-like artifacts.

The reasons why uncanny artifacts have this distinct phenomenology, however, are better explained using “alief” rather than “imaginative perception.” For Misselhorn (2009) uses the concept of “imagination” in a way that this mental state is not typically understood. In fact she is trying to incorporate elements of two very different mental processes. As I discussed in Section 1.7, imagining something requires effort. Perception requires no such effort, and “actively imagining” is clearly not what Misselhorn has in mind here anyway. Perhaps “imaginary perceptions” rather than “imaginative perception” can better describe what she is trying to say. But this still does not solve the problem I just mentioned, i.e. it seems very odd to try to incorporate elements from mental processes as different as imagination and perception if there is a simpler way to explain the mental states that occur during interaction with human-like artifacts. If, for example, I use the mental state that Gendler (2008) calls “alief,” then I have identified a more parsimonious means of explaining the uncanny valley phenomenon which doesn’t require such an intuitive and/or conceptual strain over the concept of “imagination.” Gendler’s (2008) theoretical model also allows me to retain a cognitive dissonance-type theoretical framework like that which I outlined in Section 4.3. By making this change, I am now finally able to explain the uncanny valley phenomenon.
5.4: Completing the Map of the Uncanny Valley

In this investigation, I have considered several mainstream explanations of uncanny valley phenomenon, viz. the Mortality Salience hypothesis, what I’ve called the Categorical Anomaly hypothesis, and the Imaginative Perception hypothesis. Further, I have had occasion to review a number of related areas of research in philosophy, cognitive psychology, historical psychology, intellectual history, android science and cognitive science, all toward the end of conceptually overhauling Mori’s (1970/2012) uncanny valley theory. It seems clear from the variety of possible explanations for the uncanny valley that I’ve considered that there is a certain “family resemblance” among the different possible causal factors of the uncanny valley phenomenon, in addition to how these psychological and environmental factors work such that the result is a sense of eeriness. More precisely, there are many similar sorts of conditions which, when fulfilled, result in an uncanny valley-type phenomenon. These separate threads can now be woven together into a coherent theory of the uncanny valley. In other words, I am ready to attempt to make my own contribution toward “completing the map of the uncanny valley” (Mori 1970/2012, 100), or at least, toward filling in some of the as-yet uncharted territory.\textsuperscript{150}

\textsuperscript{150} Although it is rather late in the investigation, I’d like to remark that I seldom feel eeriness around human-like artifacts. However, I have felt a sense of eeriness from time to time, and I have observed other people who are “creeped out” by human-like artifacts. I seem to fall into the camp of people who find highly realistic robots wonderful rather than eerie, although I should point out that my direct encounters with human-like artifacts such as robots has been somewhat limited. Why some people are less sensitive to the uncanny valley phenomenon than others is a matter which deserves empirical attention in the future, and which I plan to pursue in another investigation. However, I will not offer any further remarks on this matter here, other than that the reason why might have something to do with the fact that people tend to differ in their sensitivity to eerie sensations, as Jentsch (1906/2008) and Freud (1919/2003) both observed (see Sections 2 and 3 for discussion).
Let’s now briefly review the accounts of the uncanny valley phenomenon that I have considered herein. One of the first contenders to be discounted is MacDorman’s (2005) Mortality Salience hypothesis. Recall that on this hypothesis, robots are reminders of death and can cause subjects to utilize terror management strategies to cope with this angst. MacDorman’s results are illustrative of the fact that corpse-like robots are eerie or frightening, but I showed that there are reasons to think that non corpse-like robots may not be frightening in the same way (see Section 3.4). There are some further ideas which follow from this that I have not discussed in much detail thus far, viz. that the eeriness elicited by uncanny human-like artifacts is related to a disgust response linked to, say, pathogen avoidance or evolutionary aesthetics (cf. MacDorman et al 2009a; 2009b; also cf. MacDorman & Ishiguro 2006; Etcoff 1999). The reason for this is that eeriness seems to have a very distinct phenomenology, even when compared to disgust. To my mind, feeling eerie does not feel like experiencing disgust, or beholding an ugly conspecific. So even though these two emotions might co-occur, it is better to regard them as phenomenologically distinct. In this investigation, I’ve simply primarily been concerned with eeriness. Moreover, as I discussed earlier, eeriness and disgust can co-occur in some cases.

I concede that in a small number of cases – where, for instance, robots do seem particularly corpse-like as in MacDorman’s (2005) experiment – a disgust response of some kind might play a role in the uncanny valley phenomenon. But in order to be disgusted by a diseased or ugly conspecific, one first has to categorize it as such. So perhaps corpse-like robots – insofar as they have a corpse-like appearance – can elicit disgust, and this co-occurs with the eeriness that I have argued is a result of the robot’s
categorical ambiguity. Indeed, this may be a more precise use of the word “creepy”; think here of how the word creepy is often applied to frightening things which are perceived as both eerie and disgusting, such as certain snakes, insects, arachnids, and other “creepy-crawly” things. In any case, my point here is that determining what something is, is primary to determining whether it is, e.g. diseased.

However, robots and other such human-like artifacts have something other than corpse-like appearance (when they appear corpse-like, that is) in common with corpses. Namely, a freshly deceased corpse and a robot which is turned on and moving are categorically ambiguous. A freshly deceased corpse seems uncanny because of its immediately previous animate state, as Jentsch (1906/2008) argues. That is, it looks as though it might open its eyes, move its hand, or even get up and walk around, or so you’d think if you didn’t know any better. Likewise, artifacts are not supposed to move as if they are alive; yet human-like artifacts are often designed and built to give this appearance. This is the reasoning behind Mori’s (1970/2012) placing of a zombie down in the depths of the uncanny valley (100). I therefore think that the Categorical Anomaly hypothesis shows much more promise than the Mortality Salience hypothesis (see Sections 3.4 and 4.1-4.2 for discussion) for explaining the uncanny valley phenomenon.

But recall that while categorization difficulty might be a necessary condition for eeriness, it is not a sufficient condition. In the material I reviewed in Sections 2 and 3, however, two salient intentional objects of the aforementioned categorization difficulty become apparent which might be sufficient conditions for eeriness with human-like artifacts. They are, as I have discussed at length herein, the perception that a lifeless

---

There is no better example of something creepy in this way, to my mind, than the “face-huggers” from Ridley Scott’s Alien (1979) and James Cameron’s sequel, Aliens (1986).
artifact is apparently alive, and the perception of there being a mind within that lifeless, mindless artifact. There is also a third, i.e. a subject must be sensitive to the negative uncanny (see Section 2). How to explain the variance of this sensitivity amongst different people is a matter for another investigation. In any case, I am now able to say that: (1) if a subject encounters a human-like artifact/object that appears to possess categorically anomalous or ambiguous attributes, and she subsequently has difficulty categorizing the object; and (2) if the intentional objects of this categorization difficulty, i.e. the categorically anomalous features of the artifact are apparent animacy and/or mental functions; and finally (3), if she has a disposition to react negatively to uncanny stimuli (i.e. she is not one of those people I have mentioned who enjoy being “creeped out”), then she will experience the uncanny valley phenomenon, and feel eeriness in the presence of the artifact.

Above I have described the conditions under which the uncanny valley phenomenon occurs, but what of the psychology of the uncanny valley phenomenon? I think that when a subject (S) experiences eeriness with human-like artifacts, this can be explained as follows. When S interacts with a human-like artifact (A), she either (1) becomes aware that it is not really a human almost immediately, or (2) already knows that the artifact is just an artifact. If certain features such as the movement and appearance of A are sufficiently human-like, however, then A is categorically anomalous. In this case, even though S knows, believes or judges that A is an artifact and not a conspecific, the categorically anomalous features of its movement and appearance cause her to believe that she is interacting with a conspecific. Here the representational, affective, and behavioural contents of her alief might be something like: “conspecific-
looking being”; “respond empathically”; and “behave as if conspecific (there are many ways of behaving this way).” These are rather vague, but they provide some idea of what S could possibly alieve when she is interacting with A. What matters here is that, on the level of alief, S has a tendency to respond to this apparent stimulus as if it were a conspecific; she alieves that it is a human, that it has mental states, intentions, agency, the capacity to experience emotion, some combination of these attributes, etc.; further, S is motivated to behave as if the stimulus really is a conspecific, for the processes which govern this alief are not consciously governed, and are reflexive and automatic (cf. Gendler 2008). Further, these aliefs are norm-discordant. That is to say they run contrary to what S knows/believes about how she ought to act, or about how people general ought to act, around things such as inanimate artifacts. Using the language of the cognitive dissonance-type framework I briefly explored in Section 4, S’s alief – its representational, affective and behavioural content – are dissonant, or as Gendler would say “discordant,” with what she knows/believes/judges about the ontological nature of the artifact. Thus S experiences a kind of psychological uncertainty that cuts across the perceptual and cognitive levels of processing. This uncertainty elicits a type of anxiety, viz. eeriness (see Sections 2 and 3 for discussion). Thus the uncanny valley phenomenon occurs.

Someone may object here on the grounds that I have placed too much unnecessary focus on empathic abilities. I do not think this is the case. For empathy – broadly defined here as the capacity to understand other minds, either affectively or cognitively – functions as part of a theoretical framework that explains a variety of related causes for the uncanny valley phenomenon. The misfiring of empathic abilities in

---

152 Note that by “empathically” I am referring to both cognitive and affective empathy. Here the subject might alieve that the artifact has mental states, or that it experiences emotions, or both.
this way also seems to account for what gives this kind of anxiety, *i.e.* eeriness, its particular phenomenological “tone,” given the evolutionary importance of being certain about what is alive, has a mind, and what its intentions for you are, in addition to the distinctive phenomenology associated with perceiving and interaction with our fellow human beings. Using empathy also helps shed light on Gray’s & Wegner’s (2012) results, where perceptions of apparent experience were very salient to the subjects, and where this attribute (when it seemed to be out of place) caused them to evaluate the stimuli as more unnerving. As I gestured in Section 5.2, we use our own experiences – specifically, how the contents of our mind are related to what occurs in our physical and social environments – to understand the minds of others. Moreover, we do this whether we are empathizing cognitively or affectively. When this approach toward understanding other minds is misapplied in the above-described unconscious and automatic way to something which does not really have the capacity for experience – and further, when one knows it doesn’t have experience – this creates psychological uncertainty very similar to that which Jentsch (1906/2008) discusses, for instance.

Someone might also say that empathy, as I have employed the term, is not strictly necessary for the uncanny valley phenomenon to occur; perhaps one could make the argument that only dissonant aliefs and beliefs about an uncanny artifact’s apparent animacy are required in order for the uncanny valley phenomenon to occur. However as I maintained at various points throughout this investigation, animacy is one of the most salient indicators of the presence of a mind. Specifically, certain kinds of movement – for instance action-oriented motion (*cf.* Saygin *et al* 2011; Gazzola *et al* 2007), or the motion of the features of a human face (*cf.* Tinwell *et al* 2011) – indicate to a subject that she is
perceiving a being with a mind. So even if it is not immediately apparent that the misfiring of empathy plays a role in the uncanny valley phenomenon, it still plays a role in the above-described fundamental sense in that the movements of something in one’s immediate environment can indicate that it has a mind. Of course, as is clear from what I’ve discussed, I think empathy plays a much bigger role than this. In any case, all that remains now is to explain in further detail why “empathy” ought to replace “familiarity” (shinwakan) in Mori’s (1970/2012) uncanny valley theory, and to offer a few concluding remarks.
Conclusion

“When a man’s an empty kettle / He should be on his mettle / And yet I’m torn apart / Just because I’m presumin’ / That I could be kind o’ human / If I only had a heart.”

– The Tin Man, in *The Wizard of Oz* (1939)

In this investigation, I have sought to give Masahiro Mori’s (1970/2012) uncanny valley theory a conceptual overhaul. Toward that end, I have adopted a philosophical approach and argued – following Catrin Misselhorn (2009) – that the misfiring of empathic abilities, which is caused when an artifact possesses categorically ambiguous human-like features, is what explains the uncanny valley phenomenon. Moreover, I have provided a theory of eeriness which is compatible with the research results on the uncanny valley phenomenon that I’ve reviewed here. Specifically, I have argued that eeriness is a type of anxiety constituted by uncertainty concerning the ontological nature of beings in/features of the world. In the case of the uncanny valley phenomenon, salient attributes like apparent mentation and animacy are the intentional objects of this uncertainty. This conclusion appears to be consistent with some of the remarks of Jentsch (1906/2008) and Freud (1919/2003) concerning the psychology of the uncanny, as well as with the research results of Gray & Wegner (2012) and some of the themes found in Misselhorn’s (2009) work.

Following the discussion in Section 5.4, I conclude that “empathy” ought to replace *shinwakan*, *i.e.* “familiarity,” “affinity” or “rapport,” in the uncanny valley theory. On the theory I’ve presented, “empathy” works much better than the other translations of *shinwakan* because it captures the subject’s positive emotional response toward human-like artifacts, and because the failure of empathy can explain eeriness.
This is very much in line with Misselhorn’s (2009) argument, although she does not articulate this idea, and my account of the uncanny valley phenomenon differs in terms of the concepts and some of the theorized causes/constituents it employs. Moreover, it is clear from what I have discussed herein that uncanny stimuli – those which are categorically anomalous – are so because certain of their attributes cause human empathic abilities to misfire. This answers some of the questions related to the problem of eeriness (see Section 1.2), while remaining consistent with the results of empirical research on the uncanny valley phenomenon. The conception of eeriness provided in Section 3 may also facilitate future empirical research on this emotion, for it is clear from what I’ve discussed here that eeriness is not a well-researched emotion, and deserves further empirical attention.

The theory of eeriness I’ve developed here might also apply to cases of eeriness beyond the uncanny valley, *i.e.* eeriness with human-like artifacts. Recall the example of the uncanny photograph discussed in Section 3.5, or my occasional references to purportedly paranormal phenomena. I think that the uncanniness of both of these things is explainable in terms of the theory I present here. Caution is required here, for these are separate phenomena from the uncanny valley, although the causes of eeriness in these cases may share some family resemblance with those of the uncanny valley phenomenon, *e.g.* apparent agency where it shouldn’t be present. I hope that future research on the emotion of eeriness addresses cases like these. Until then, I shall not make any firm conclusions on this matter.

As mentioned, I think that the failure of empathy with human-like artifacts which I have articulated here occurs in a slightly different way than on Misselhorn’s (2009)
Imaginative Perception hypothesis. She thinks that the uncanny valley results from the activation and de-activation of “imaginatively perceived” human-like features – like a very fast, repetitive gestalt-switch – which oscillate back and forth from active to inactive. I have argued that this failure is constituted not by oscillation, but by dissonance between aliefs, and beliefs, judgements or knowledge (i.e. “cognitions”) concerning the ontological nature of the uncanny stimulus. Understand, however, that this dissonance is not exactly of the kind discussed by Festinger (1957), i.e. it is somewhat different from how the term “cognitive dissonance” is usually understood. Nevertheless, what I have sketched out above is consistent with Burleigh et al’s (2013) and MacDorman et al’s (2009b) suggestion that some sort of psychological dissonance contributes to the uncanny valley phenomenon.

Moreover, following the results empirical research discussed here (especially in Section 4), it is clear that it is not a matter of an artifact’s position at a certain point on a continuum of human-likeness that places it in the uncanny valley; rather, it is whether it lies on a categorical boundary. Here, I follow some of the thinking of Ramey (2005), Kang (2011), Burleigh et al (2013) and Yamada et al (2013); that is, I concur that one additional change to the uncanny valley theory ought to be shifting focus from degrees of human-likeness to ontological category membership. Furthermore, I’ve shown in this investigation that the eeriest human-like artifacts, as suggested by Ishiguro (2006; 2007), seem to pass the Turing test at the unconscious level. This is probably enabled by neural mechanisms like the Mirror Neuron System, as suggested by MacDorman et al
(2009b), and as evidenced by the fMRI studies reviewed in Section 5. The apparent minds of these artifacts cause an eerie response because they cause a subject to attribute the capacity to have emotions, intentions, beliefs and desires to the artifact in a persistent, automatic and unconscious way, even though she knows it doesn’t have these things. Thus a subject experiences “norm-discordant aliefs” (Gendler 2008) when the uncanny valley phenomenon occurs. These aliefs are norm-discordant in that they are primarily the sort of aliefs a subject only has when she interacts with conspecifics and not with inanimate artifacts, and they are likewise dissonant with what a subjects knows/believes/judges about the artifact’s ontological nature. Finally, I think that this feeling of eeriness becomes most pronounced when the subject alieves that the artifact in question has agency and – following Gray & Wegner (2012) and Misselhorn (2009) – the capacity to form beliefs, desires, intentions, and to experience emotions and bodily sensations (e.g. pain).

So I say that when it comes to human-like artifacts with apparent mentation and animacy, much of the discussion of “brains” ought to be saved for The Scarecrow. What I mean here is that these artifacts evoke the range of emotional responses that they do not because of their minds’ computational prowess, or any attributions of mind like that. That would be rather strange. Rather, it is the aforementioned appearance of intentions, beliefs and desires – which are made apparent to a subject because of the kinds of animacy which these artifacts exhibit – and even the capacity to experience things such as emotions, or pain, etc. which cause the psychological dissonance or discord that results in

---

153 MacDorman et al (2009b) mention several additional studies, but Gazzola et al’s (2007) is probably the most pertinent for my purposes here, along with Saygin et al’s (2011), and the work of Rosenthal-von der Pütten and her colleagues (2013a; 2013b), discussed in Section 1.
the uncanny valley phenomenon. In other words such artifacts, in their own strange and uncanny way, appear to possess what the Tin Man had all along: a heart.
References


