Perception of robot faces: Investigating the temporal aspect of the uncanny valley phenomenon using EEG

201607157 Anna Agermose Hinrichsen (AAH)

&

201609409 Amalie Holm Lund Sørensen (AHLS)

Cognitive Science at Aarhus University

Bachelor thesis

20.12.18

Characters:

AAH: 38885

AHLS 38300

In total: 77185

Summary

The ability to recognize and extract meaning from a face is crucial to human beings. From a face, we not only make inferences about the person it belongs to but also about his or her mental states. Mind perception enables us to understand each other and thereby is a necessary condition for social living. Recent advances in the technological development has brought about new social agents in the form of robots. Android robots are designed to resemble humans as much as possible both regarding appearance and actions. Generally, it is found that people react positively to robots with many human characteristics. However, it has been found that at some point increasing humanness will result in a sudden dip in affinity. This phenomenon is described as *the uncanny valley*.

In the format of a registered report, this thesis presents a study that can investigate the neural underpinnings of the perceptual experience of uncanniness relating to android robots. The behaviourally observed dip in affinity which is characteristic for the uncanny valley is in the present study sought grounded in brain responses, measured with electroencephalography (EEG). It is proposed that the signal value of different ERP components will depend on the degree of robotness in a face. More specifically, it is expected that uncanny stimuli will differ from fully human and fully robotic in the later components (>200 ms) reflecting reentrant dynamics.

A pilot EEG study was conducted to explore this prediction and results suggest that the curvature associated with the uncanny valley may be reflected in a peak in signal value around 232 ms after stimuli onset. That is, at time point 232 uncanny faces seem to be processed differently than fully human and fully robot faces. Furthermore, a decrease in signal value for more robotic faces appeared around 109 ms after stimulus onset. These time points are selected for further hypothesis testing in the pre-registered study. An additional aim of the pre-registered study will be to correlate ratings of uncanniness with signal values of later components at the individual level. This potential correlation will help disclose whether the observed difference in signal value indeed reflects the personal experience of the uncanny valley.

Keywords: *The Uncanny Valley, EEG, ERP, reentrant dynamics, face processing, android robots, mind perception*

Table of contents

1.0 The format of the thesis (AHLS)

This bachelor thesis is based on empirical work within the field of social robotics. Empirical hypothesis testing is a core principle of modern science (Gauch Jr, 2012, p. 1-19). This method relies on clear research questions, well-formulated hypotheses, clever experimental designs and data collection, meaningful statistical analyses and based on these elements, a precise conclusion (ibid.). This scientific method helps ensure that results are valid and thus researchers should follow the principles carefully.

Reproducibility is necessary in order to generalize findings, however there is evidence that especially in psychology, it is questionable whether this requirement is met (Chambers, 2013; Open Science Collaboration, 2015).

Also, the tendency for journals to be more interested in publishing papers with significant and extraordinary findings rather than those with null results poses a problem to science (Thornton & Lee, 2000). This publication bias may create incentive to cherry pick analyses and p-hack almost-significant results in order to get published (Nosek, Spies, & Motyl, 2012).

To prevent these problems, a new format for publishing called the *registered report* has been proposed by the Center for Open Science in 2014 (Nosek & Lakens, 2014). This allows researchers to submit a manuscript with a detailed description of methods and analysis before conducting the study. Pilot experiments can be included to test feasibility and establish proof of concept. The manuscript is then approved for publication based on good research questions and valid procedures. On condition that findings are interpreted reasonably the study will be published regardless of results. This minimizes publication bias dramatically by committing first the researchers to conduct their study as planned and second the publisher to accept also reproduced findings or null-results (ibid.).

The benefits of registered reports as sketched out here have guided the decision to adopt the format for this thesis.

2.0 Topic introduction and thesis outline (AAH)

Face perception is one of the most developed visual skills in humans. This is likely because faces convey crucial social meaning and are thus important for successful interaction (Haxby, Hoffman, & Gobbini, 2002). Alongside this, specific biomechanical structures make humans able to finely control their facial muscles and create various expressions for communication (Haxby et al., 2002; Wheatley, Weinberg, Looser, Moran, & Hajcak, 2011). Body language and voice can provide information about others' state of mind, but it seems that the most distinctive and important key to a person's identity and intentions, is the face (Bruce & Young, 1986; Winston, Strange, O'Doherty, & Dolan, 2002).

Recent technological development has introduced new actors to our social environment (Urgen, Kutas, & Saygin, 2018). We see that robots are incorporated into domains that were previously only occupied by humans (Broadbent, 2017). As a result of this change, the question arises as to how a robot should be designed in order to successfully take part in social interactions with humans. If faces are as important for social interaction as suggested, it seems natural that a social robot should have a face (Panchal, 2017). Furthermore, the face should be designed with features that allow humans to detect social cues (DiSalvo, Gemperle, Forlizzi, & Kiesler, 2002). In accordance with this it has been found that more human-like features in a robot face increase the perceivers affinity for the robot (M. J. E. Mori, 1970a). Thus, to facilitate positive social interactions, robots should resemble humans in terms of appearance (Broadbent, 2017).

But a phenomenon that interferes with the above logic is *the uncanny valle*y (Mori, MacDorman, & Kageki, 2012). The uncanny valley reflects the observation that people commonly report an eerie feeling towards entities such as robots or animated characters which look almost human (Karl F. MacDorman & Ishiguro, 2006; Mori et al., 2012). The uncanny feeling occurs at the moment it is realized that although initially believed to be human, the entity in question is in fact inanimate (Andersen, 2018).

This thesis explores how the uncanny valley relates to face perception focusing on the temporal aspect of the phenomenon. The first part of the paper consists of a literature review including theories of face perception, a presentation of the uncanny valley hypothesis as well as an overview of relevant early and late event-related potentials in the brain. The second part

describes a conducted pilot study. In the pilot study, it is tested via electroencephalography (EEG) recordings whether faces on a scale from fully human to fully robot elicit different brain responses.

The findings from the literature and the results from the pilot study together guide the exact experimental set up of the pre-registered study.

3.0 Literature review (AAH, AHLS)

3.1 Face processing (AHLS)

3.1.1 A specialized area

As noted in the introduction, face perception appears to be a very important ability for human beings. Evidence from cognitive neuroscience supports this intuition. Many cognitive functions are thought to be distributed throughout several areas of the brain connected in specific patterns. Face processing is however one of the few which has been found to reside in very specific areas committed to this one task.

In 1997, Kanwisher, McDermott and Chun conducted an fMRI study on face perception in humans and found higher activation in the fusiform gyrus for faces rather than other objects. The fusiform gyrus is located just above the inferotemporal gyrus and is now referred to as the fusiform face area (FFA). Increased activation of this area in response to faces has later been found consistently across various studies. Kanwisher, McDermott and Chun followed up on their initial finding by designing a series of experiments to eliminate other possible explanations for activation in FFA. Such alternative interpretations of the higher activation include difference in low-level feature extraction between face- and control stimuli, visual attention engaged more by faces, subordinate visual recognition of within-category stimuli and a difference in recognition of any human object compared to inanimate stimuli. As none of these suggestions were found to explain the results, the authors concluded that the area is selectively involved in face perception (Kanwisher et al., 1997).

FFA has consistently been shown to respond to faces but different hypotheses exist as to why this is the case. Studies have found that experts show higher activation in the FFA

when discriminating fine details in objects within their field of expertise such as cars or birds (Gauthier, Skudlarski, Gore, & Anderson, 2000). A possible explanation for this finding is that experts perceive objects from their area of expertise as distinct individuals rather than generic exemplars of a category which would be the case for non-experts. All human beings are considered to be experts in faces, hence it follows that the act of face processing would be expected to elicit higher activation in the FFA.

That face processing seems to rely on specific mechanisms is also suggested by the clinical condition of prosopagnosia. Patients who suffer from this condition show impairment in recognizing familiar faces, but often not other visual objects. A study of a sheep farmer who suffered from severe prosopagnosia following a stroke, showed that although he was not able to recognize his friends and family and did not recover this ability, he was still able to learn to recognize his sheep again. This result suggests that prosopagnosia is a disorder specific to human faces (McNeil & Warrington, 1993).

Duchaine and Nakayama (2005) conducted different tests, examining the ability to recognize faces and objects in seven developmental prosopagnosics. Along the lines of similar studies, their results demonstrated that face perception and object perception rely on different mechanisms. It is important to note that studies with brain injured patients necessarily have several uncontrollable parameters and often also a limited number of participants. Results should be interpreted with this in mind.

In face detection experiments, healthy individuals show faster and more accurate responses when recognizing human faces as opposed to other visual stimuli. Such results are taken as further evidence in favour of faces as a special category (Haan, Pascalis, & Johnson, 2002). As is the fact that people can remember a great amount of individual faces over a long period of time (Bahrick, Bahrick, & Wittlinger, 1975).

3.1.2 Bruce and Young's framework

Bruce and Young proposed in 1986 a cognitive model of face recognition. They developed a framework which put together and extended previous models. This framework builds on the idea that face recognition and identification involve an interaction between different functional components. The authors propose that the results of these components'

operations, that is, the information derived from seeing a face can be classified as seven different types; pictorial, structural, visually derived semantic, identity-specific semantic, name, expression and facial speech codes (Bruce & Young, 1986).

Pictorial information consists in information about lighting, grain, shades and colours. This information may be enough to recognize a picture of a face previously displayed for example in an experimental task. However, in real life pictorial information changes with e.g. face orientation and expression but we are nevertheless able to recognize familiar faces. This ability depends on a more abstract representations of the face which is encoded in structural information. Bruce & Young (1986) suggest that familiar faces are represented by a set of structural codes which allow recognition both from particular face parts and from the configuration of these. If the structural codes match a previously stored representation, the face is recognized as familiar.

Even for unfamiliar faces we can obtain information about a person's gender and age which in this framework is called a visually derived semantic code. This depends entirely on physical features. Facial expressions are also available to us for both familiar and unfamiliar faces revealing information about the person's emotional state.

Face recognition as it happens in everyday life can be described in terms of access to the seven different information types. In Bruce and Young's functional model, structural encoding produces a description of the face parts and configuration as well as a more abstract representation. Analyses of facial speech and expression are made based on this description and also face recognition units are thought to obtain information from it. The concept of face recognition units refers to the idea that each face someone knows has its own recognition unit which can retrieve information about that person from the so-called person identity node. The more the current face percept resembles the stored representation the stronger the signal of the recognition unit. The signal is also elevated if top-down knowledge primes us to expect to see a certain person in a given situation. When the signal is strong enough, we experience not only having recognized a face but also identified the person to whom the face belongs. Bruce and Young note that this process of deciding whether the face we see does actually belong to a person we know or just someone who look-alike is probably more complex than sketched out here (Bruce & Young, 1986).

3.1.3 Core and extended system

Haxby, Hoffman and Gobbini described in 2002 how face perception relates to the functional anatomy of relevant brain areas. The areas involved in face processing is considered to be structured hierarchically with a core system performing a visual analysis and an extended system that with help from other cognitive systems extract meaning from the face perceived.

In the core system the authors focus on the distinction between variant and invariant parts of the face as these two feature categories are thought to rely on processing in different brain areas. Returning to the framework of Bruce and Young (1986) we also find the distinction between the types of information used for identification and that which is employed in recognition of facial expressions and speech codes. The idea gets anatomical grounding by Haxby, Hoffman and Gobbini (2002) who suggest that the face-sensitive regions in the extrastriate visual cortex are organized according to this distinction. Hasselmo et al found in 1989 single neurons in the macaque brain responding to changes in either facial expressions or face identity. The neurons most responsive to expression were mainly found in the superior temporal sulcus whereas those sensible to identity were primarily located in the inferior temporal gyrus (Hasselmo, Rolls, & Baylis, 1989). Haxby, Hoffman and Gobbini (2002) suggest that the corresponding regions in the human brain would be the superior temporal sulcus and the lateral fusiform gyrus. Their functional imaging study from 2002 constitutes a functional dissociation between these two areas. The superior temporal sulcus responded more to eye gaze whereas the lateral fusiform gyrus was shown to be more responsive to individual identity (Haxby et al., 2002).

The extended system of face perception consists of several brain regions that also contribute to cognitive functions not related to faces. The auditory cortex is thought to be involved in lip reading, the amygdala, insula and limbic system in processing of emotional face expressions and anterior temporal regions in retrieving name and biographical information for the person to whom the face belongs. It is important to understand and model how these cognitive functions related to face processing rely on regions interacting with each other in different constellations (Haxby et al., 2002).

To sum up, neuroimaging studies confirm the importance of face perception by localizing the process to distinct and specialized areas in the brain. Furthermore, several models have been proposed as to how information about faces is integrated within these regions. The framework proposed by Bruce and Young (1986) as well as the core and extended system proposed by Haxby and colleagues (2002) include a distinction between the processing of physical face properties and the extraction of meaning from the face.

3.2 The Uncanny Valley (AAH)

That faces are very salient perceptual stimuli is not surprising taking the human evolutionary history into account. Recognizing another as a human being is crucial for social interaction and the ability of quick and confident recognition of other humans is present already in infancy (Bushnell, 2001; Field, Cohen, Garcia, & Greenberg, 1984; Johnson, Dziurawiec, Ellis, & Morton, 1991). However, not only human beings have faces. Also stuffed animals, cartoon characters, dolls and robots can have a more or less human-like representation of a face. Dolls may resemble human beings in appearance, but our perceptual system is fine-tuned enough to avoid being fooled by such look-alikes (Wheatley, Kang, Parkinson, & Looser, 2012).

Nevertheless, as technology advances, the line between artificial and biological is blurred by android robots designed to look as human as possible. When faced with such a robot the process of categorizing it as either human or inanimate is complicated by its human-like appearance (Looser & Wheatley, 2010). The classification of faces as either human or inanimate is important as it determines whether there is a mind behind the face or not; an inference which in turn advises us how to react. Especially eyes are considered informative when assessing animacy (Looser & Wheatley, 2010).

Situations in which android robots look so convincing that our perceptual system is briefly confused, can result in a feeling of eeriness which the Japanese roboticist, Masahiro Mori, has described as uncanny (Karl F. MacDorman & Ishiguro, 2006; Mori et al., 2012).

In 1970 Mori hypothesized that the more human-like a robot becomes, the more affinity people have for it. But also that our affinity only rises until a certain point. Beyond this point the robot looks so much like a real human being that it is almost but not perfectly

convincing. At this point, affinity will drop dramatically. Only if the robot becomes so perfect that we cannot tell it apart from a real human being, affinity will rise again. This dip for very human-like robots were by Mori called the uncanny valley (Mori et al., 2012)[1]. Humans have an innate ability to recognize biological movement, which might have developed due to evolutionary importance of recognizing animacy and thereby better recognizing potential threat (Gazzaniga, Ivry, & Mangun, 2014, p. 442). In accordance, (Mori et al., 2012) additionally hypothesized that movement will change the shape of the uncanny valley graph by steepening the slopes, and thus make the feelings associated with the uncanny valley more extreme.

Recently developed technology allows us to build robots with convincing human resemblance and as a consequence the uncanny hypothesis again attracts the attention of researchers and designers. The people basing their research on Mori's idea come from a broad range of disciplines including robotics, engineering, cognitive science, psychology, philosophy and even evolutionary biology (Andersen, 2018).

Mori's original paper was written in 1970 in a Japanese journal called Energy. An unofficial and unpublished translation of the paper (M. Mori, 1970) was circulating in 2005 and a great deal of research and interpretations of the uncanny valley has been based on this. An official translation, edited by Mori himself, was not published until 2012 (Mori et al., 2012). One line of research in the field intends to find the uncanny valley empirically while another proposes explanations which can account for the phenomenon. Such explanations include theories of threat avoidance arising from the need for self-preservation, cognitive dissonance from objects close to category boundaries, prediction error and terror management (K F MacDorman, Green, Ho, & Koch, 2009).

Findings are mixed, both when it comes to proving the existence of the uncanny valley and regarding possible explanations. Some of this variability may be explained by the lack of well-established and common conceptual ground of the uncanny valley (Andersen, 2018; Redstone, 2013). There has been disagreement regarding how to translate the Japanese term 'shinwakan' (Mori et al., 2012) used to describe the feeling elicited by the robot. Suggested translations in English include the words familiarity, empathy, affinity, likability and

[1] Mori's original graph is included in appendix 1

acceptability (Andersen, 2018). Similarly, the dimension often plotted at the x-axis has been translated to human likeness, anthropomorphism, human nature, realism and human realism (ibid.)

Despite the confusion about which words to use when describing the feelings towards uncanny entities, several studies have found behavioural evidence showing a valley for specific kinds of robots which are more human-like than others (Hanson, 2006; Karl F. MacDorman & Ishiguro, 2006). MacDorman and Ishiguro (2006) asked participants to first choose which of 31 pictures morphed between human and robot they found to be eerie and second rate them in terms of familiarity, human likeness and eeriness. The familiarity rating plotted against the morphed pictures from robot to human showed a valley corresponding to the one proposed by (Mori et al., 2012).

An important, yet not thoroughly investigated part of the uncanny valley hypothesis is the temporal aspect (Andersen, 2018). Mori (2012) writes about the uncanny valley in relation to a prosthetic hand: "However, once we realize that the hand that looked real at first sight is actually artificial, we experience an eerie sensation." (p. 99). This quote indicates that the uncanny valley has a time course. First, there is no drop in the level of affinity; we simply believe that what we see is alive (Andersen, 2018). But then we realize that our first intuition was wrong, and what we are looking at is artificial. It is not before this realization that we get the eerie sensation (Andersen, 2018; Mori et al., 2012).

In relation to this, it has been suggested that the uncanny experience is not only linked to visual, bottom-up properties of the robot (Andersen, 2018) but also to later, top-down processes such as extracting meaning from its face (Wheatley et al., 2011). The meaning we extract may relate to what intentions, emotions and mental abilities that potentially lie behind the face(Winston et al., 2002).

3.3 Mind perception (AHLS)

Throughout the history of philosophy thinkers have investigated what makes human beings different from other entities. Discussions on the topic often end up revolving around the human mind and its supposedly unique ability to think, reason and self-reflect (K. Gray & Wegner, 2012; Suddendorf, 2013, pp. 215-229). The human mind withholds a great deal of

secrets and its workings is also today the subject of different fields of study. However, the very existence of the mind is rarely doubted; As Descartes (1641/2011) famously stated "Cogito ergo sum" [I think, therefore I am] (Descartes, 1641/2011).

Essentially the only mind we can know for sure exists is that of our own (Dennett, 1996, p. 2). Nevertheless, with the possible exception of the most sceptic philosophers, people generally believe other human beings to have minds (Suddendorf, 2013, p. 39). I.e. although we cannot directly observe the inner lives of others, we ascribe to them mental capacities similar to our own. This belief rests on the fact that we belong to the same species and thereby are thought to share basic capacities. The idea is further strengthened by the causal relationship between thoughts and actions we perceive to exist (Davidson, 1963). There seems to be a link between mental states and behaviour in that one in some way causes the other. Actions can be understood as results of preceding mental operations i.e. when someone moves his arm it is at least to some degree *because* he in his mind decided to do so (ibid.). Thus, when we observe the behaviour and expressions of another, we can infer something about his intentions, feelings and beliefs. This act can be referred to as mind perception. It is a very important and fundamental process that enables us to understand each other, interact and live together as social beings (Amodio & Frith, 2006; Epley & Waytz, 2010). Children as young as 18 months are thought to have the ability of mentalizing implicitly about intentions (Frith & Frith, 2003).

The process of mentalizing is so essential to our social lives that we sometimes even apply the ability in relation to non-human entities. Some people treat their pets as if they were human and others regard their car as a helpful friend (Waytz, Gray, Epley, & Wegner, 2010). Waytz, Gray, Epley and Wegner (2010) argue that causal uncertainty is directly connected to mind perception in others. If a computer works exactly as expected it seems mindless and mechanical. If, however it does not react to commands as it should, uncertainty of the cause of the deviating behaviour is raised in the perceiver. This causal uncertainty in turn entails the idea that the computer has a mind of its own as this idea seems to explain the unanticipated behaviour (Waytz et al., 2010).

The coupling of minds and computers is not completely arbitrary or unreasonable. Some cognitive scientists would say that a computer is actually a good analogy for human cognition; the hardware corresponds to the brain and the software to the mind (Rescorla,

2017). If a computer processes information in the same way as the brain it seems logical to ask whether a computer, then can be said to ‘think’.

Alan Turing presented in 1950 a test to examine this question without having to define what it actually means to think. The scenario of the Turing test is two agents asking each other questions in writing. One is human, and the other is a computer. If a third human judge cannot distinguish between the two the machine has passed the test and said to exhibit intelligent behaviour (Turing, 1950). That a Turing test is enough to determine if a machine has human intelligence is a claim which has later been refuted e.g. by John Searle with his famous Chinese Room argument from 1980. The argument relies on a thought experiment involving a non-Chinese speaking person sitting in a room. He receives questions in Chinese from a person outside and his job is to answer in Chinese. He has a basket with Chinese symbols and a rule book in his own language explaining how to put the symbols together. This enables him to construct an answer that would make the person outside think that he knows Chinese. What the argument is meant to show is that even though the person in the room has passed the Turing test he cannot be said to know Chinese. By analogy, neither a computer can be said to actually think even though it seems to do so. It is concluded that there must be something special to the human mind which essentially cannot be replicated by a computer program. A machine can resemble a real human mind to a high degree but simulation is not the same as duplication (Searle, 1980).

In 2000 Nass and Moon examined the phenomenon of mind perception towards computers. They created an experimental setting where participants interacted with computers not resembling humans at all. Although all participants verbally rejected anthropomorphism, that is no-one said that computers should be understood and treated as human beings, their behaviour towards the computer showed otherwise. It was found in a series of experiments that participants did apply social rules and expectations to the computer. They over-used human social categories like ethnicity and gender and engaged overly in learned social behaviour such as politeness and reciprocity (Nass & Moon, 2000). The authors suggest that the reason for this behaviour is, that people, instead of carefully analysing and constructing categories based on all features, seem to apply overly-simplistic scripts already created. They simply ignore the indications that the computer is entirely mechanical and treat it as a social being.

Along the same lines Nowak and Biocca (2003) found that participants responded socially to agents in a virtual environment no matter if they were perceived to be human or computer-controlled. Three conditions were tested where either no image, a less-anthropomorphic image or a highly anthropomorphic image represented the interaction partner. Results showed that copresence and social presence were rated highest in the less-anthropomorphic condition. This is suggested to show that a highly anthropomorphic image sets up expectations of human-like actions which it cannot meet leading to an experience of less presence (Nowak & Biocca, 2003). Such expectation violation might also be a relevant aspect when explaining the causes of the uncanny valley (Urgen et al., 2018).

Mind perception may not only have one dimension but rather be composed of different aspects. Gray, Gray and Wegner (2007) examined the question through a survey asking participants to compare different human and non-human characters with regard to assumed mental abilities. Characters included robots, animals and humans at several ages from foetus to adult. 18 questions of mental abilities and 6 of personal judgments were included in the survey. Examples of mental abilities are ability to feel pain and ability to understand the feelings of others. Personal judgment was expressed in question of e.g. how much the participant liked the character or if the character would deserve punishment for a crime. A factor analysis of the 24 questions of mind perception showed that the concept can be expressed in two dimensions; agency and experience.

Gray and Wegner (2012) have also investigated mind perception as a possible cause of the eerie feeling which can be evoked by almost human-like robots. They proposed that the reason an uncanny robot elicits an eerie feeling is because it pretends to be human and have a human mind but that at the same time these qualities are seen as fundamentally impossible for machines to have. This mismatch of expectation and reality (prediction error) is thought to elicit the feeling of unease. Again the distinction between the two dimensions agency and experience was emphasized (H. M. Gray et al., 2007). Agency, which is the ability to act could be found in the uncanny robots. On the other hand, experience or the ability to feel was seen as essentially lacking in robots. It is suggested that a mind capable of agency but without experience is ‘disturbingly incomplete’ (K. Gray & Wegner, 2012).

3.4 Event-related potentials (AAH)

One way to investigate what happens in the brain when people see a face, is by tracking brain responses and characterize their sensitivity to stimulus manipulations. This can be been done using electroencephalography (EEG) (Luck, 2014; Teplan, 2002).

The EEG method records the electrical signals produced by electrical activity in the brain (Darvas, Pantazis, Kucukaltun-Yildirim, & Leahy, 2004). The EEG reading is made from the scalp, hence it is a non-invasive method that can be applied repeatedly to participants with no risk (Teplan, 2002). The greatest advantage of EEG is its temporal resolution, that is of 1 ms or better under optimal conditions (Luck, 2014). Other methods such as fMRI, which uses hemodynamic measures, has much poorer temporal resolution limited to several seconds (Gazzaniga et al., 2014, p. 103; Luck, 2014). A disadvantage, however is that EEG signals cannot provide exact information about localization of the signals (Luck, 2014).

If the electrical activity is tied to specific stimuli and averaged across many trials, it results in what is called an event-related potential (ERP) (Luck, 2014; Picton et al., 2000). Event related potentials are electrical changes in EEG recordings and as they are time-locked to stimuli they can provide knowledge about cognitive, sensory or motor mechanisms in the brain (ibid.).

In order to uncover an ERP component, the stimulus of interest has to be shown many times. After EEG recording, the data is then segmented into data portions that capture the electrical signals short before and after the stimulus onset. This segmentation is often referred to as epoching. After epoching, each of the data-portions (or epochs) are averaged sample by sample, resulting in an average time-course (Luck, 2014).

ERP components are often labelled by their approximate latency in milliseconds and their polarity. Thus, the N170 component refers to a negative deflections of the signal with a peak around 170 ms after stimulus onset. Some ERP components are associated with specific psychological processes (Gazzaniga et al., 2014, p. 100).

3.4.1 ERPs and reentrant dynamics in the brain

Most theories of perception, including face perception (see section: 3.1.3 Core and extended system), involve interacting brain areas (Di Lollo, Enns, & Rensink, 2000). In dealing with the question of how this interaction takes place, researchers have investigated the flow of information in the brain. It has been found that ongoing, parallel signalling among different brain areas occurs, which is referred to as reentrant dynamics within cortical hierarchies the brain (Edelman & Gally, 2013). Hence, when visual information enters the brain, a wave of information goes through the visual system, starting in hierarchically low areas, such as the primary visual cortex. In parallel, back-signalling, or reentrant processing, serve to test predictions about the incoming information (ibid.).

The sequence of different ERP peaks can also reflect information flow in the brain (Garrido, Kilner, Kiebel, & Friston, 2007; Luck, 2014). Garrido and colleagues (2007) found that forward connections from hierarchically lower areas, affect the brain responses throughout the presentation of a stimulus, whereas later ERP components, later than 200 ms after stimulus, are mediated by reentrant dynamics.

It is found that reentrant processes play a role in attention and awareness (Hamker, 2003). The brain can focus resources on processing the intrinsic character of a sensory input but also on expectations from higher cortical areas. The strongest reentrance signals, will be the ones that we attend to (Edelman & Gally, 2013).

3.4.2 Early components

Some early components include the P100 and the N170 (Luck, 2014).

The P100 component is elicited by visual stimuli (ibid.). It has been shown to be sensitive to variation in the physical characteristics of a stimulus, such as contrast and spatial frequency. However, it is not solely derived from sensory processing but can be affected by attention and the subject's state of arousal (Luck, 2014; Saavedra & Bougrain, 2012). Other top-down processes do not affect the P100. Thus the component seems to be more dependent on external features of a stimulus rather than internal influences (ibid.).

The N170 component can be captured from the occipito-temporal brain regions. It is larger in response to face stimuli than other stimulus categories. The N170 has been hypothesized to be a marker of a face-specific system (Carmel & Bentin, 2002; Rossion & Jacques,

2008). It has been suggested by many that the N170 component merely reflects early processing such as the detection and structural encoding of human faces (Bentin, Allison, Puce, Perez, & McCarthy, 1996; M. Eimer, 2000), and that the response is prior to processes involved with face identification and recognition ((Bentin et al., 1996; M. Eimer, 2000). Supporting this hypothesis, it has been found that N170 is not affected by the familiarity of a face (Bentin & Deouell, 2000; M. Eimer, 2000). Neither do other aspects of face processing that relate more to top-down processes, such as differences in gender and age, seem to affect the amplitude of the N170 (Mouchetant-Rostaing & Giard, 2003).

It has been found that inversion of faces affects N170 differently than inversion of objects. (Rossion & Jacques, 2008) found that the N170 was not affected when inverting objects but when faces were inverted the N170 was both enhanced and delayed. They proposed two explanations for the enhancement of amplitude, either it has to do with higher difficulty in the processing of inverted faces or processing of inverted faces might recruit both object and face-processing systems. The latency shift is compatible with other findings suggesting that faces are perceived as wholes and as the inversion disrupt the configural information in the face, it affects the possibility of fast holistic processing (Liu, Harris, & Kanwisher, 2010).

It has been argued that eyes are critical social stimuli and therefore eyes are specifically important when perceiving a human face (Taylor, Edmonds, McCarthy, & Allison, 2001). Bentin and colleagues (1996) found that the N170 response to isolated eyes was bigger than the response evoked by whole faces. Their experiment was not designed to investigate whether the N170 component could indicate a specialized eye processor and they therefore proposed that more research should investigate this. They did suggest from their results that eyes are the most representative feature in the face (Bentin et al., 1996). Taylor, Edmonds, McCarthy and Allison (2001) tested the N170 component in children and its sensitivity to eyes compared to whole faces. They found that the N170 was larger and faster for eyes than for whole faces and by those results they proposed that eye processing might develop before face processing in children.

A study by Eimer (1998) found that the N170 was delayed for faces where the eyes were removed. But the amplitude of the N170 did not seem to be affected by the removal of eyes and therefore, they concluded that N170 is not caused by the activation of cortical regions specialized for eye processing.

The above suggests that the N170 reflects detection and encoding of structural properties of faces. One of the reasons why face perception might be an important and specialized aspect of human cognition is that faces often convey affective information which is crucial for social behaviour and from an evolutionary perspective, might have been important for survival (Gazzaniga et al., 2014, p. 246). It is therefore interesting if the N170 can be affected by top-down processes. Blau, Maurer, Tottenham and McCandliss (2007) found that the N170 response to pictures of emotional faces was larger in amplitude when compared to neutral faces, and thereby suggested that besides structural, exogenous processing the N170 might reflect top-down modulations from emotional systems. However, their findings are not conclusive, as other literature regarding this topic has returned conflicting evidence (Rellecke, Sommer, & Schacht, 2013).

3.4.3 Late components

Some of the late components that have been associated with face processing include the P300 and the N400 wave (Barrett & Rugg, 1989; Bentin & Deouell, 2000; Meijer, Smulders, Merckelbach, & Wolf, 2007; Schacht, Sommer, & cognition, 2009)

In contrast to the P100 wave, the P300 does not seem to be affected by the physical properties of a stimulus, but instead it is dependent on internal factors. Thus it can be considered an endogenous potential (Picton et al., 2000). There has been several studies testing what factors influence the amplitude of the P300 component (Luck, 2014). Many findings point to a connection between the P300 wave and attention. One of the most robust findings is that probability in detection tasks affects the amplitude of P300 so that when the probability of a target stimulus is low, the P300 amplitude associated with it, is larger (Luck, 2014). Also, the difficulty of a task can affect the component (ibid.).

The N400 component is recognized as the electrophysiological index of semantic processing. The component has mostly been used to investigate semantic expectations in language paradigms. However as both words and pictures are visual symbols that carry meaning, it has been argued that N400 is relevant when investigating both linguistic and non-linguistic stimuli (Nigam, Hoffman, & Simons, 1992). Supporting this argumentation Nigam and colleagues (1992) found that the N400 for word- and picture stimuli was identical in terms of

amplitude. It is therefore argued that N400 reflects activity in an amodal conceptual memory system that can be accessed by both words and pictures (Kutas & Federmeier, 2011; Nigam et al., 1992).

Researchers are debating what kind of comprehension- and neural processes that lie behind the N400. *The integration view* states that the N400 is happening after the recognition of a word or a picture. It occurs due to difficulties in integrating a recognized symbol into a context or into pre-existing information contained in memory. However, this theory cannot account for the finding that words with no meaning has been found to elicit the N400 (Deacon, Dynowska, Ritter, & Grose-Fifer, 2004). Another theory is *the lexical view.* According to this theory, N400 reflects processing stages prior to recognition stages and also prior to semantic access (Kutas & Federmeier, 2011; van Vliet, Mühl, Reuderink, & Poel, 2010). Studies have found that N400 seems to be sensitive to manipulation associated with both prior and post recognition stages. More recent accounts of the N400 posit that both theories may be relevant, and that N400 possibly indicates a broad range of processes related to semantic memory (Kutas & Federmeier, 2011).

Kutas and Federmeier (2011) argue that language, perception, attention and memory all contribute to the neural events behind the N400 component. They conclude that we can understand the component as a window into the brain's way of making predictions about the world and using these to comprehend meaning. This conclusion coincides well with what was described previously; that later components, such as N400, relate to reentrant dynamics where top-down predictions are made from bottom-up information (Garrido et al., 2007).

From what has been said about the N400 component, it might not be straightforward how it can be connected to face processing. However, many have made the assumption that face processing not only include bottom-up processes such as structural encoding. Finding meaning in the face is also an essential part of the process (Balconi & Pozzoli, 2005; Klatzky, Martin, & Kane, 1982). Haxby, Hoffman and Gobbini's theory of a core and an extended system for face processing, relies on this exact inference. They describe how the core system performs a visual analysis and the extended system extract meaning from the faces.

The fact that faces carry meaning, have motivated researchers to investigate the semantic-related N400 in this domain as well (Barrett & Rugg, 1989; Bentin & Deouell, 2000). In fact, human faces seem to be a category where meaning is extremely important, as faces are often the basis for successful navigation in a social world (Haan et al., 2002). We can find meaning in facial expression and use emotional cues in the face for appropriate interaction,

and we can identify faces as belonging to specific stereotypical categories (Bruce & Young, 1986; Klatzky et al., 1982).

Furthermore, the N400 component has been used in research about the uncanny valley. Based on the inference that the component is the brain's way of making predictions about the world, Urgen, Kutas and Saygin (2018) have studied whether EEG responses to a real human, a mechanical robot and a realistic (human-like) robot differed. They added a dimension of movement to each category. The real human moved like a human, the mechanical robot moved like a robot is expected to move, but the robot that looked like a human moved like a mechanical robot. The realistic robot elicited an N400 response and the two other categories did not. The authors concluded that a prediction violation is likely an underlying mechanism of the uncanny valley and when investigating the time course of the phenomenon, it might be relevant to look at later processing, such as the N400 ERP component.

4.0 Pilot (AAH,AHLS)

4.1 Introduction to pilot study (AAH)

This registered report proposes an empirical study investigating how face processing in the brain interacts with the phenomenon of the uncanny valley.

In order to formulate precise hypotheses for the pre-registered study, a pilot experiment was carried out. The pilot study investigated the expectations obtained from the literature in an exploratory manner.

The reviewed literature of the uncanny valley suggests that the time course of the phenomenon might include two stages. One early stage in which the perceiver is fooled to believe that they see a human, and one in which they realize that something is wrong. The latter, most likely engage more top-down processing, related to re-entrance dynamics. Thus it is expected that uncanny stimuli will cause a different activation in later ERP components such as P300 and N400 compared to fully human and fully robot stimuli. Earlier components are not expected to show this difference.

4.2 Method (AAH, AHLS)

4.2.1 Stimuli (AHLS)

The stimuli used for the pilot experiment consisted of pictures of android robots morphed with human faces. Each robot face was matched with a human face that had similar facial features and head orientation. These pairs were morphed together in FantaMorph 5.5.8 (Abrosoft, 2002-2018) based on the exact location and form of face parts. Eight different morphs were created each split in 15 pictures equally spaced on the human-to-robot spectrum. All pictures were converted to grayscale to avoid colour differences to impact results. Most pictures of human faces were retrieved from the ADFES Stimulus Set (Van der Schalk, Hawk, Fischer, & Doosje). The rest of the pictures including those of robots were found on the internet.

To better suit the practical demands of EEG-studies it was decided to reduce the number of stimuli by selecting the five best morphs and only include five pictures from each. The following section describes the selection process.

The morphs were designed with the aim of capturing the phenomenon of uncanniness. How uncanny a picture was judged to be was expected to depend on where on the spectrum the picture was located. A picture at one end of the spectrum was fully human and a picture at the other end was fully robot. Neither of these were expected to be perceived as uncanny. In contrast, pictures in the middle of the spectrum were expected to be judged as more uncanny as they reflected a mix of human and robot. Thus, when plotted, the uncanny judgments were predicted to follow a curve.

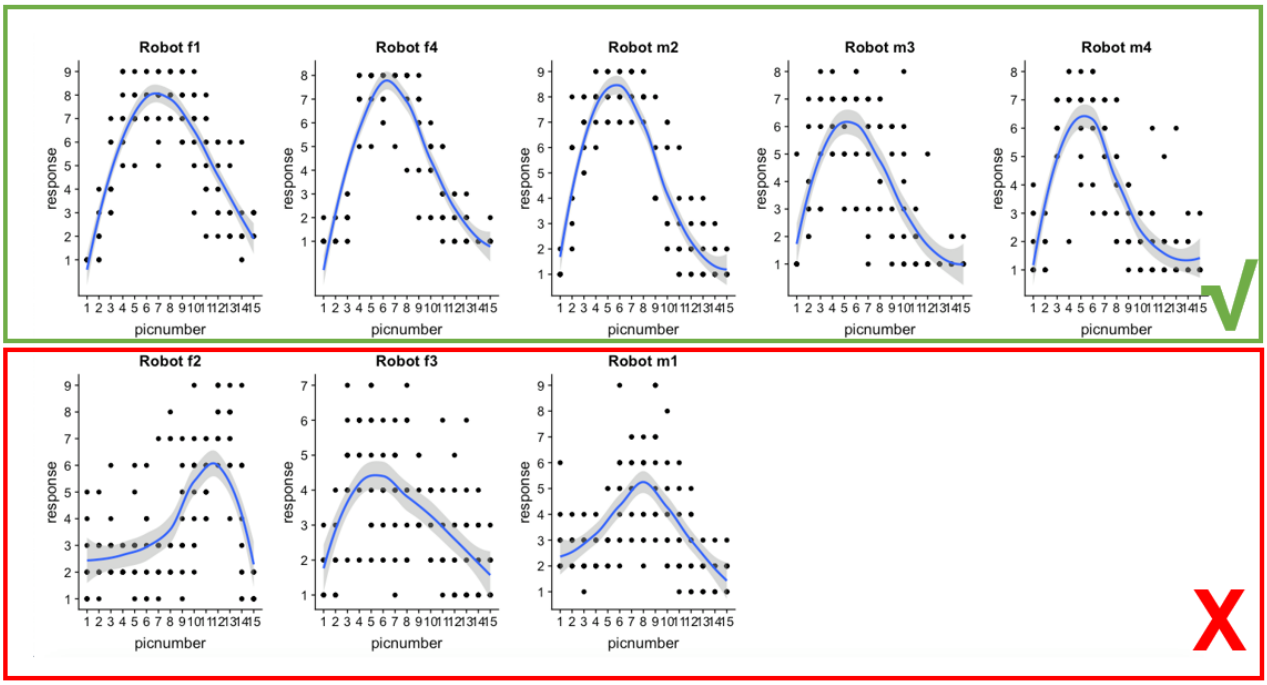

Figure 1: Uncanniness rating for the eight different human-robot pairs.

Based on uncanniness ratings from the two experimenters a curve was created for each morph. As expected not all morphs elicited equally clear curves. The five morphs displaying the best curves were selected as final stimuli (upper five in figure 1).

After choosing the five best morphs each of them had to be reduced from 15 to 5 picture slices. Again, the curves of uncanny ratings were used.

As each curve reflected a different human-robot pair all curves were not expected to exhibit the same steepness or to peak in uncanny judgement at the same image slice. I.e. image slice number 4 from pair 1 might have gotten a different uncanny-judgement than image slice number 4 from pair 2. To account for this variability, the five images for each morph were selected based on perceptual distance. For each of the five selected morphs the following procedure was adopted: Picture 1 and 15 were chosen. The picture rated highest was chosen. The rating value for this was divided in two, and the two pictures with ratings closest to the resulting value were chosen. This process ensured in a simple way that the different morphed pairs were directly comparable. The method did however rest on the assumption that the two end-points both get 0 in uncanny rating which was not entirely true.
Below is an example of one of the resulting morph spectra (Figure 2).

Figure 2: Example of stimuli

4.2.2 Preliminary work (AAH)

Before the EEG study a preliminary behavioural experiment was conducted. The purpose of this was to test if the introduction shown prior to the EEG experiment was precise enough to make the phenomenon of the uncanny valley and the rating task understandable to naive participants. The introduction consisted of text explaining the uncanniness as well as examples of robots commonly judged to be uncanny.
Eight people participated in this comprehension experiment (mean age = 35.5; SD = 17.93). Every participant had to judge 2 morphs each, meaning a total of 30 image slices.

After the experiment, many participants reported uncertainty about the general phenomenon of uncanny indicating that the concept can be hard to grasp. The study was conducted in English but with Danish participants having English only as their second language. The word uncanny does not have a direct translation in Danish which seemed to further confuse participants. In the context of this experiment it was important that participants had an understanding of the phenomenon of uncanny as described by Mori rather than of the word as it is used in everyday language. The introduction text was elaborated, and examples were changed to accommodate this.

Although a general understanding of uncanniness is necessary it should also be stressed that individual participants are encouraged to form their own idea. The personal experience is essential to the uncanny valley and is interesting especially in conjunction with information about brain responses. If brain responses to an uncanny stimulus reflect the subjective experience of it, will be explored in the pre-registered study.

4.2.3 Participants (AAH)

Two female 23-year-old participants took part in the pilot study.

Participants needed to have normal or corrected-to-normal vision and no history of neurological disorders.

4.2.4 Experimental design (AAH)

The experiment was designed using the software Presentation (NeuroBehavioral Systems, 2004)[2]. First, participants were shown an introductory text about the general purpose and form of the experiment. They were also shown the modified introduction to uncanniness together with examples as described above. In the end of the introduction, participants were instructed in which keys to use when responding to the behavioural task (figure 3).[3]

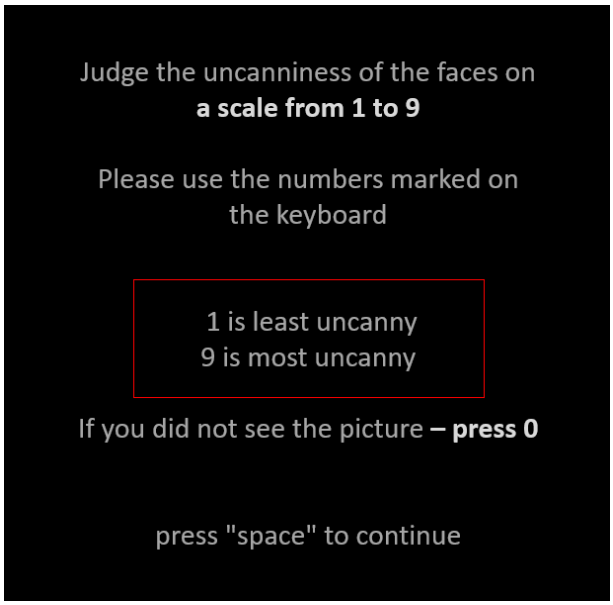

Figure 3: Example of instruction text

The paradigm (figure 4) comprised 1000 trials which each consisted of a fixation cross (400 ms), a picture (500 ms) and a rating screen asking participants to rate the picture according to uncanniness on a scale from 1 to 9 using keys on a standard computer keyboard.

[2] The full code can be found in appendix 2
[3] The exact elements of the introduction are included in appendix 3

If participants did not see the picture, they were asked to press 0. The first fixation cross lasted 4 seconds to make sure participants were ready to begin the experiment.

The durations of stimuli presentation are close to but not exactly e.g. 500 ms as it is dependent on the monitors refresh rate. However, the triggers sent to the EEG equipment are accurate and reflect the exact timing of stimuli.

Figure 4: A graphic representation of the paradigm

Each robot pair had a trigger code assigned to ensure that the EEG recordings could be time locked to the onset of each picture and to be able to differentiate between the different pictures. The key press responses that the participant could gave each picture also resulted in triggers being sent to the recording computer.

4.2.5 Data acquisition (AHLS)

Participants signed an informed consent form stating their rights regarding the experiment and that participation was voluntary. They were seated comfortably in a room with dimmed light. Participants were instructed not to move or clench their jaws unnecessarily as the artefacts resulting from these movements can interfere with the EEG signal. Oculomotor activity was measured with two electrodes attached with sticky tape. One was placed on the outer canthus of the left eye and the other above the same eye at the supraorbital ridge.

EEG was recorded at 1000 Hz using an electrocap (ActiCAP, Brain Products, GmbH) with 32 Ag/Ag-Cl electrodes (see figure 5), an amplifier and Brain Vision Recorder ® (Brain

Products, GmbH). The electrodes were placed according to the 10-20 international system. The reference electrode was placed at Cz and the ground electrode at Fz. Electrolyte gel was used to secure conductance from scalp to electrodes. The impedance for most of the electrode-offsets was kept below 25 k-ohm.

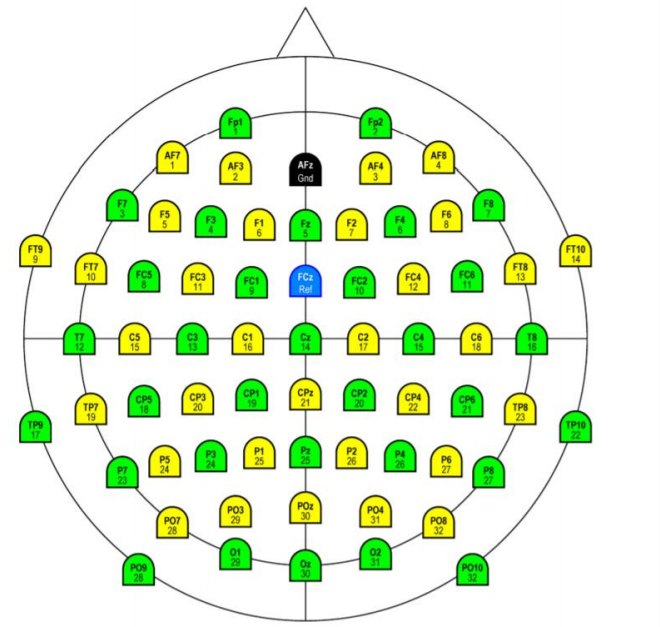

Figure 5: the electrode distribution on the ActiCAP (ActiCAP, Brain Products, GmbH). Only the 32 electrodes marked in green, were used together with the ground and reference electrodes (marked with blue and black in the picture).

4.2.6 EEG pre-processing: Data exclusion (AAH)

For the pre-processing of the raw EEG data, the open-source Python software, MNE-python was used (Gramfort et al., 2013)

The data was high-pass filtered at 1 Hz and low-pass filtered at 40 Hz. Signals below 1 Hz are considered too fast to reflect brain responses and therefore excluded (Luck, 2014).

Insufficient connection between electrodes and scalp and other technical issues can result in poor signal quality from certain electrodes. Such bad channels were found and rejected using digital filtering algorithms automatically applied by the Python toolbox Autoreject (Jas, Engemann, Raimondo, Bekhti, and Gramfort, 2017).

In the pilot unintended duplication of one of the trigger codes, resulted in further data exclusion. Both the key press response of "0" and the onset of picture 1 from robot pair A were assigned the same trigger code. As it was not possible to separate the two, robot pair A was excluded from the data analysis.

4.2.7 Electrode selection (AAH)

Signal values were extracted from 12 electrode sites. From the left and right occipital (O1, O2, Oz), parietal (Pz, P3, P4, P7, P8) and centro parietal regions (CP1, CP2, CP5, CP6).

These electrodes were chosen based on results reporting that N400 is largest above central parietal sites (Kutas and Federmeier, 2011). N170 has primarily been located at occipito-temporal sites (Rossion 1999, Bentin 1996). The largest signals for N170 are often found at electrodes T5 and T6, which on the ActiCAP are referred to as P7 and P8 (Bentin 1996, Eimer 1998, Eimer & Holmes 2002).

4.2.8 Epochs and ERPs (AHLS)

Epochs of 700 ms were created from the preprocessed data. Each epoch was time-locked to the image presentation starting 200 ms before and lasting until 500 ms after stimulus onset. The mean signal value for the 200 ms preceding stimulus onset was subtracted from the following values as baseline correction.

To initially explore the evoked data, plots were created for both participants. As expected they showed random fluctuations around zero until stimulus presentation. This is a sign of good data. See example from one participant below (figure 6).

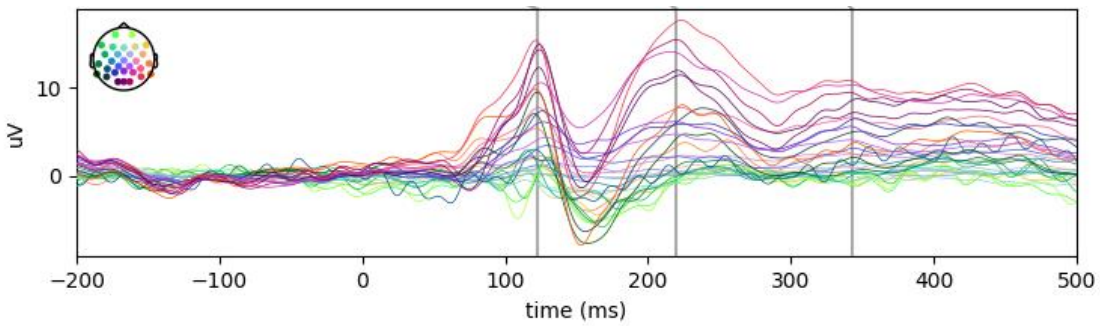

Figure 6: Plot of evoked signals for participant 1 in robotness condition 3 which is at the midpoint be-tween fully human and fully robot. The coloured lines display signals from 30 different elec-trodes

Using these epochs, specific time points were extracted for analysis. The time points were derived from the global field power plot which displays the root mean square of the signal value across all electrodes at all conditions simultaneously (figure 7). The plot indicated effects around 109 ms, 232 ms and 413 ms thus time windows from 99-119 ms, 222-242 ms and 403-423 ms were selected.

Signals from all samples within each of the four time windows were averaged across all selected electrodes. This results in four values (one for each time window) linked to each stimulus presentation and reflects the average peak amplitude.

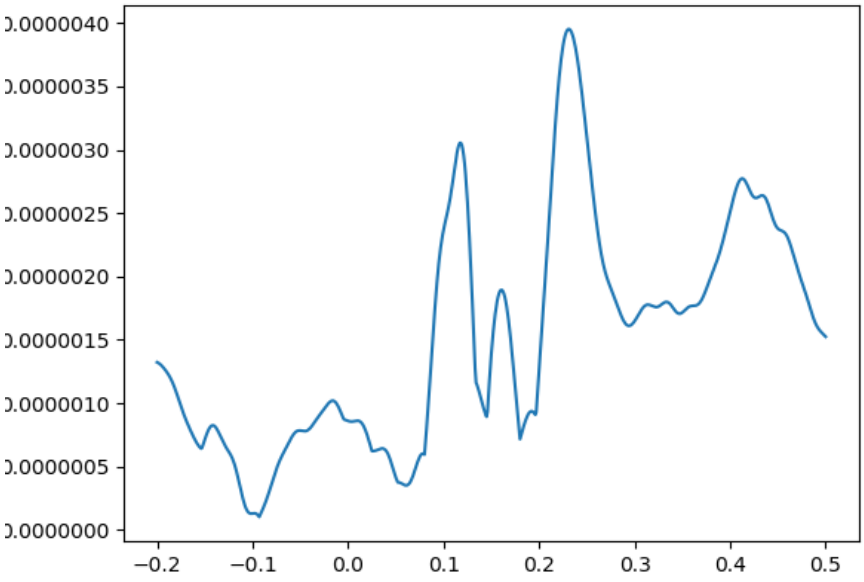

Figure 7: Global field power plot. Y-axis displays signal value in µV, x-axis displays time from 200 ms before stim-ulus onset (-0.2) to 500 ms after stimulus onset (0.5)

4.3 Analysis and results (AAH, AHLS)

4.3.1 Statistical analysis (AAH)

After the data was pre-processed, it was analysed via the statistical software R (R Core Team, 2017) using the package lme4 (Bates, Maechler, Bolker, & Walker, 2015). Two multilevel linear models were created and a log-likelihood comparison between the two models was performed using the R-function anova (R Core Team, 2017). This was followed by a further analysis of the best performing model.

The function aggregate from the R stats package (R Core Team, 2017) was used to look at medians and standard deviations across conditions and time points. These were plotted with ggplot2 (Wickham, 2016).

4.3.1.1 The models

A scale from fully human to fully robot (see stimuli section) was used to explore whether the brain responds differed depending on how robotic or human a face is.

In accordance with the reviewed literature, the models were created on the assumption that early and late ERP components rely on different brain dynamics. Hence, changing the degree of robotness will possibly have a different effect at different time points. This is apparent in the model as an interaction effect between the degree of robotness and time point.

Model 1 was the simplest model predicting signal values from an interaction between degree of robotness and time point. As random effect, the model had by-participant random intercept for signal values. This was included due to the fact that the general strength of the EEG signal often varies across participants (Luck, 2014).

Model equation (m1):

$$Sv = \beta_{0p} + \beta_1 robotness * \beta_2 timepoint + \beta_3 robotness*timepoint$$

For the second model, a quadratic term was added to the equation. This accounts for the expectation that responses to faces in the middle of the robotness scale, will differ from responses to faces in each end of the scale (fully human and fully robot), reflecting the drop in affinity which has previously been found for very human-like robots. Thus, the signal values is suggested to depend on robotness following a quadratic polynomial.

Model equation (m2):

$$Sv = \beta_{0p} + \beta_1 robotness + \beta_2 robotness^2 + \beta_3 timepoint + \beta_4 robotness*ERP + \beta_5 robotness^2*timepoint$$

4.3.2 Results (AHLS)

The result of the model comparison, showed a trend towards significance favoring model 2 over model 1 $\chi 2(3)=6.95$, $p=0.07$. It was not significant ($p > .05$) but due to a low number of participants and the exploratory nature of the pilot study, model 2 was selected for further investigation.

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	2.724e-06	7.632e-07	3.570
condition	2.380e-07	4.197e-07	0.567
condition2	-7.799e-08	6.864e-08	-1.136
time0.232	2.139e-06	7.771e-07	2.753
time0.413	1.126e-06	7.771e-07	1.449
condition:time0.232	-1.207e-06	5.936e-07	-2.033
condition:time0.413	-1.184e-07	5.936e-07	-0.199
condition2:time0.232	2.374e-07	9.707e-08	2.446
condition2:time0.413	4.212e-08	9.707e-08	0.434

Figure 8: Table of estimates, standard error and t-values for main effects and interactions in model 2 including all time points.

As displayed in the two bottom lines of the table of estimates (figure 8), there is an interaction effect between squared robotness (condition2) and time point 232 when compared

to 109, β =2.374e-07, SE = 9.707e-08, t-value = 2.446. But when time point 109 is compared to time point 413 there is no effect of robotness, β = 4.212e-08, SE = 9.707e-08, t-value = 0.434. In other words, squared robotness seems to have a different effect on signal value at time point 232 compared to time point 109. A t-value greater than 2 is taken to signify an effect.

There is also an effect of the interaction between robotness (not squared) (condition) and time point 232 β =-1.207e-06, SE = 5.936e-07, t-value = -2.033.

Because an interaction effect is included, it is not meaningful to interpret the main effects of time alone.

To be able to further interpret the interaction reported above, a visualization (figure 9) of the data was made using the R-package ggplot2 (Wickham, 2016).

The y-axis in the plot has been scaled in order to better visualize the effects.

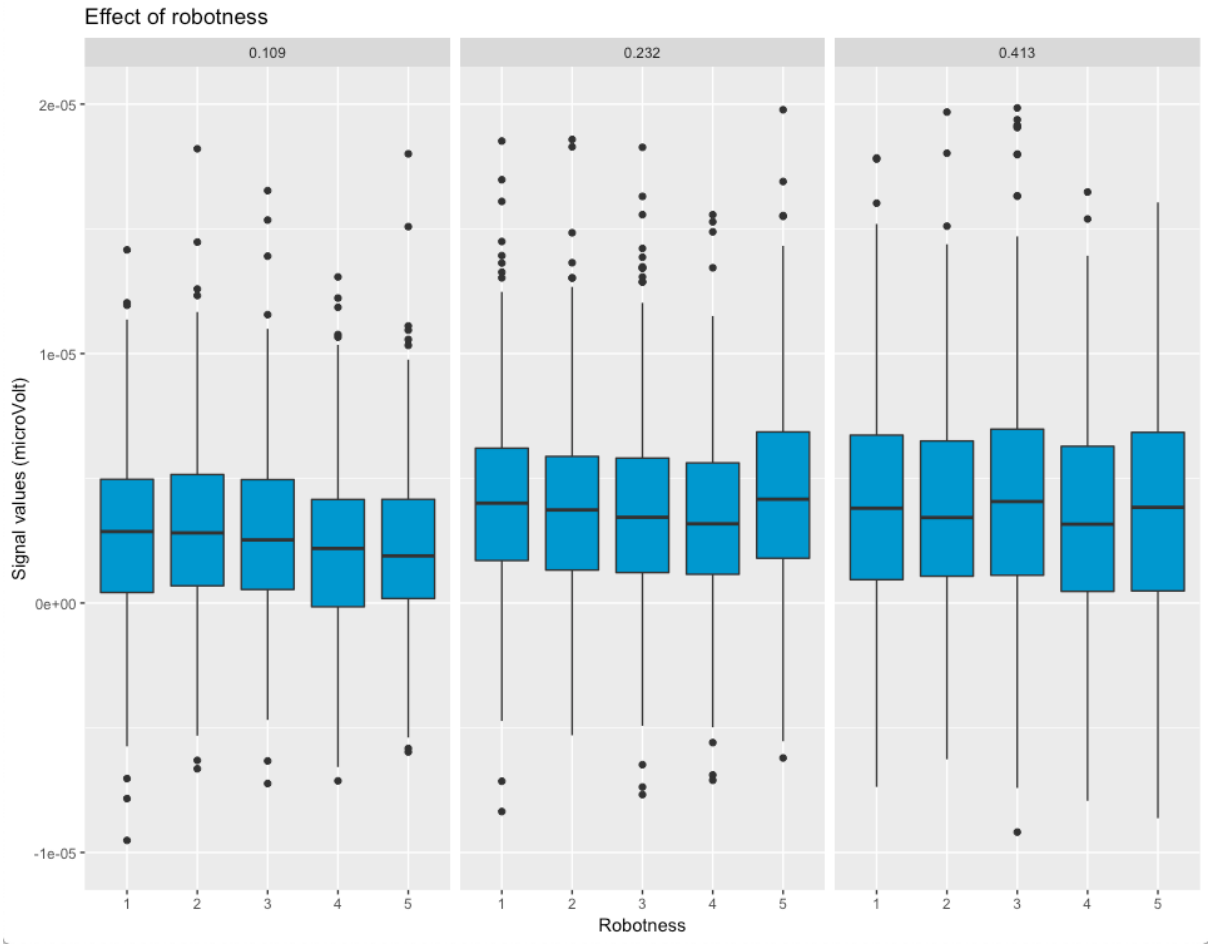

Figure 9: A boxplot showing the effect of robotness on signal values (μV). The black line in the middle of the boxes depicts the median of the values. The boundaries of the blue boxes represent the upper and lower quartiles. The line extending from the box indic

The model comparison showed that adding a quadratic term improved the model fit implying a curvature in the data. The boxplot indicates a vague curve in the data at time point 232 meaning that faces in the middle of the robotness scale seem to differ from faces in each end of the scale (fully human and fully robot). Therefore, time point 232 is potentially what drives the overall curvature effect. At the 109 time point, there is a trend towards declining EEG signals for more robotic faces. At the 413 time it is hard to see any effect of robotness.

Even though, trends appear in the plot, there is reason to be critical. The error bars are very long, the boxes are overlapping, and several outliers are present in the plot. However, results serve as indications of effects. Especially when taking into account the general messiness of EEG data as well as the fact that the data is obtained from only two participants.

	Robotness	time	median	sd
2	1	0.232	3.950329e-06	4.819048e-06
5	2	0.232	3.731150e-06	4.693779e-06
8	3	0.232	3.438127e-06	4.535452e-06
11	4	0.232	3.172370e-06	4.523828e-06
14	5	0.232	4.122271e-06	4.286762e-06

Figure 10: Signal value medians for time point 232

The different median numbers show that the vertex of the curve is found at a robotness level of 4 (figure 10).

The above results suggest that robotness is affecting EEG signals but only at time points of 109 and 232, and not at time point 413. To explore this further, part of the analysis was run again using a data set that only contained data from time point 109 and 232.

4.3.3 Results: Time point 109 and 232 (AAH)

Model comparison between model 1 and model 2 (with the quadratic term) now showed statistical significance favoring model 2 $\chi 2(2)=7.5065$, $p=0.02$. This confirms the idea that time point 413 did not drive the curvature effect.

```
Fixed effects:
                       Estimate Std. Error t value
(Intercept)           2.731e-06  6.466e-07   4.224
condition             2.433e-07  3.948e-07   0.616
condition2           -7.906e-08  6.457e-08  -1.224
time0.232             2.139e-06  7.310e-07   2.927
condition:time0.232  -1.207e-06  5.583e-07  -2.161
condition2:time0.232  2.374e-07  9.131e-08   2.600
```

Figure 11: Table of estimates, standard error and t-values for model 2, only including time point 109 and 232

Excluding time point 413 did not change the direction of the effects but increased t-values indicating more confident estimates.

The interaction between squared robotness (condition2) and time point is now, β =2.374e-07, SE = 9.131e-08, t-value = 2.600 and the interaction effect of robotness (condition) and time point is, β =-1.207e-06, SE = 5.583e-07, t-value = -2.161.

5.0 The pre-registered study (AAH, AHLS)

5.1 Method (AHLS)

The method in the study will closely follow the method used in the pilot. This includes experimental design, data acquisition and participant- and data exclusion criteria.

Results from the pilot established the quality of the paradigm. Therefore the preliminary comprehension study and the stimuli selection process will not be repeated.

The electrodes used for analysis in the study will also be the same. Epochs will be created at time points 109 ms and 232 ms (+/- 10 ms) as the field power plot from the pilot suggested. The time point 413 will not be included, based on results from the pilot.

The statistical analysis will also be carried out in a similar manner including comparison of models with and without a quadratic term and analysis of the best performing model. Boxplot and table of medians and standard deviations will be included as they were in the pilot.

In the pre-registered study, the greater number of participants will make it possible to look at individual differences in response to uncanniness. In every trial a picture is shown,

and participants are asked to rate how uncanny they perceive it to be. This task helps ensure that the participants stay awake and ready during the paradigm. However, the aim of the rating task is first and foremost to assess the perceived uncanniness as experienced by participants. In the pre-registered study, it will be explored whether there is a correlation between individual rating and EEG signal values. The correlation test will be carried out in R with the function cor.test (R Core Team, 2017). As the rating variable is ordinal and not continuous Spearman's ranked correlation coefficient will be used to assess results. The coefficient will be judged according to standard benchmarks for weak, medium and strong effects shown below (A. Field, Miles, & Field, 2012, pp. 211-225) (figure 12).

Correlation coefficient *(r)*	Interpretation of effect size
0	No effect
± .1	Small effect
± .3	Medium effect
± .5	Large effect

Figure 12: Standard interpretation of effect size

A potential correlation between personal rating of uncanniness and responding brain signals will speak in favour of the anticipation, that the indicated effect found in the pilot study was indeed induced by uncanniness.

5.2 Hypotheses (AAH)

The exploratory work on the pilot and the reviewed literature form the basis of the hypotheses for the pre-registered study.

5.2.1 Literature findings

The literature points to face perception as one of the most developed and specialized systems in humans. This specialization might be due to the social importance of facial information. It seems that we use facial information for mind perception, which is an important process when it comes to understanding and interacting with each other.

Reentrant dynamics can describe how face processing elicits different brain responses. Bottom-up processes, such as visual analysis of physical elements in a face, have been related to earlier ERP-components. Top-down processes, such as extracting meaning from a face have been linked to later ERP-components. The top-down predictions may be connected to the uncanny valley in that prediction violation is suggested as its underlying mechanism.

In regard to the uncanny valley, Mori suggested a temporal aspect of the hypothesis. He assumed that initially when faced with an almost-humanlike robot, there is no drop in the level of affinity because people believe that what they see, is alive. But then they realize that their first intuition was wrong, and what they are looking at is an artificial agent with no mind attached. The mismatch between early expectations and the reality, then causes the eerie sensation upon realization.

5.2.2 Pilot results

The results from the pilot study first of all establish proof of concept. Secondly, the exploratory work provides some specificities of method and analysis and guides formulation of hypotheses in the pre-registered study.

The global field power plot was explored, and highest activations were found at latencies of 109 ms, 232 ms and 413 ms. However, the analysis showed no effect of robotness on signal value at 413 ms.

Based on these considerations, hypotheses for the pre-registered study are formulated as follows:

H1)

Signal values at time point 232 will be distributed as a curve with more extreme values for faces approaching the middle of the robotness scale, and less extreme values for fully human or fully robot faces. Thus, a model that predicts signal values from the degree of robotness will improve if a quadratic term is added.

H2)

Signal values at time point 109 will depend on robotness in a linear manner. When degree of robotness increases, signal value will decrease.

6.0 Preliminary discussion (AAH)

The following sections go beyond the registered report and are included for the purpose of the thesis.

6.1 What the pre-registered study can and cannot say

Expectation violation arising from mind perception in robots have been proposed as a possible explanation of the uncanny valley. This is touched upon in the literature review as considerations about potential causes are interesting and relevant to discuss when investigating a phenomenon. However, the main focus of this study has been to empirically ground the uncanny valley with neural dependent measures and does not attempt to explain exactly why the phenomenon arises.

Whether the effects found in the pilot study are truly expressing Mori's idea of the uncanny valley, cannot be concluded from the study. However, If results from the pilot replicate it indicates that the brain has a special response towards something that looks human, but is not convincing enough to be perceived as fully human. The study therefore provides empirical evidence that the curve found in behavioural studies can be observed in brain responses.

Moreover, the investigation tells us that pictures of faces alone, as opposed to e.g. entire robots physically present, may be enough to induce different brain responses to almost-but-not-quite-human agents compared to fully human and fully robot. This is further evidence that faces provide a great deal of meaningful information to the perceiver. Although Mori's original hypothesis proposed movement as a amplifying effect this aspect cannot be investigated with the proposed paradigm, as it exclusively uses still pictures of faces.

As follows, the uncanny valley is a multifaceted phenomenon which can be examined with different aims and methods. With neural dependent measures the present study helps collect empirical evidence that the uncanny valley exists. It is thus considered as contributing to the understanding of the phenomenon within the field of social robotics.

7.0 Conclusion (AHLS)

This thesis has presented a registered report of a proposed EEG study examining the uncanny valley phenomenon. The registered report consisted of a literature review and a description of method, analysis and results of a conducted pilot study. Conclusively, the method of the pre-registered study was set up and hypotheses were presented.

The literature review concerned the topics of face processing, the uncanny valley, mind perception and early and late ERP-components in relation to faces. It was found that face processing is a fundamental mechanism relying on specific neural structures. What is more, perception of a face leads to inferences about the mind behind it. Mind perception applied in the case of very human-like robots may serve as an explanation of the uncanny valley phenomenon i.e. a dip in affinity for very human-like robots. These theories formed the initial ground for designing a pilot study using the method of EEG. Method and analysis of this pilot were described for later application in the proposed pre-registered study.

The results from the pilot study showed that the brain responses elicited by face stimuli ranging from fully human to fully robot may be better explained by a quadratic polynomial than by a straight line. This aligned with the expectation that processing is different for very human-like faces than for fully human or fully robot. The effect was found at a latency of 232 ms after stimulus presentation, and hence this time point was chosen for analysis in the pre-registered study as well. Also time point 109 showed a peak in overall activity, which was later found to reflect a decrease in signal value as stimuli become more robotic. These results guided the formulation of hypotheses for the pre-registered study.

Although only two people participated in the pilot, results indicated an effect of robotness on signal value. This is in itself a promising result which implies that there is an effect to be investigated. It augurs well for an approval to continue with the pre-registered study which in turn may contribute to a better understanding of the uncanny valley.

References

Abrosoft. (2002-2018). Abrosoft FantaMorph. Retrieved from www.fantamorph.com

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature reviews neuroscience, 7*(4), 268.

Andersen, S. S. (2018). *The Uncanny Valley or the Valley of Eeriness: Reconstructing the Uncanny Valley Hypothesis for Interdisciplinary Research Exchange*. Unpublished manuscript. Cited with permission from the author.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of experimental psychology: General, 104*(1), 54.

Balconi, M., & Pozzoli, U. (2005). Morphed facial expressions elicited a N400 ERP effect: A domain-specific semantic module? *Scandinavian Journal of Psychology, 46*(6), 467-474.

Barrett, S., & Rugg, M. D. J. N. (1989). Event-related potentials and the semantic matching of faces. *27*(7), 913-922.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:doi:10.18637/jss.v067.i01

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological Studies of Face Perception in Humans. *8*(6), 551-565. doi:10.1162/jocn.1996.8.6.551

Bentin, S., & Deouell, L. Y. (2000). Structural encoding and identification in face processing: ERP evidence for separate mechanisms. *Cognitive neuropsychology, 17*(1-3), 35-55.

Blau, V. C., Maurer, U., Tottenham, N., & McCandliss, B. D. (2007). The face-specific N170 component is modulated by emotional facial expression. *Behavioral brain functions 3*(1), 1.

Broadbent, E. J. A. R. o. P. (2017). Interactions with robots: The truths we reveal about ourselves. *68*, 627-652.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology, 77*(3), 305-327.

Bushnell, I. W. R. (2001). Mother's face recognition in newborn infants: Learning and memory. *Infant Child Development: An International Journal of Research Practice, 10*(1-2), 67-74.

Carmel, D., & Bentin, S. (2002). Domain specificity versus expertise: factors influencing distinct processing of faces. *Cognition, 83*(1), 1-29. doi:https://doi.org/10.1016/S0010-0277(01)00162-7

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609-610.

Darvas, F., Pantazis, D., Kucukaltun-Yildirim, E., & Leahy, R. (2004). Mapping human brain function with MEG and EEG: methods and validation. *NeuroImage, 23*, S289-S299.

Davidson, D. (1963). Actions, reasons, and causes. *The journal of philosophy, 60*(23), 685-700.

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology, 41*(1), 60-74.

Dennett, D. C. (1996). *Kinds of minds: toward an understanding of consciousness*: Basic Books.

Descartes, R. (1641/2011). Meditations on first philosophy. In L. P. Pojman & L. Vaughn (Eds.), *Classics of Philosophy* (pp. 489-516): Oxford University Press.

Di Lollo, V., Enns, J. T., & Rensink, R. A. J. J. o. E. P. G. (2000). Competition for consciousness among visual events: the psychophysics of reentrant visual processes. *129*(4), 481.

DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002). *All robots are not created equal: the design and perception of humanoid robot heads.* Paper presented at the Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques.
Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of cognitive neuroscience, 17*(2), 249-261.
Edelman, G. M., & Gally, J. A. (2013). Reentry: a key mechanism for integration of brain function. *Frontiers in integrative neuroscience, 7*, 63.
Eimer, M. (2000). Event-related brain potentials distinguish processing stages involved in face perception and recognition. *Clinical Neurophysiology, 111*(4), 694-705. doi:https://doi.org/10.1016/S1388-2457(99)00285-0
Eimer, M. J. N. (1998). Does the face-specific N170 component reflect the activity of a specialized eye processor? *, 9*(13), 2945-2948.
Epley, N., & Waytz, A. (2010). Mind perception. *Handbook of social psychology, 1*(5), 498-541.
Field, T. M., Cohen, D., Garcia, R., & Greenberg, R. (1984). Mother-stranger face discrimination by the newborn. *Infant Behavior Development, 7*(1), 19-25.
Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*.
Garrido, M. I., Kilner, J. M., Kiebel, S. J., & Friston, K. J. (2007). Evoked brain responses are generated by feedback loops. *Proceedings of the National Academy of Sciences, 104*(52), 20961-20966.
Gauch Jr, H. G. (2012). *Scientific Method in Brief*: Cambridge University Press.
Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*(2), 191.
Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2014). *Cognitive neuroscience : the biology of the mind* (Fourth edition ed.). New York: Norton.
Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python.
Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619-619.
Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125-130.
Haan, M. d., Pascalis, O., & Johnson, M. H. (2002). Specialization of neural mechanisms underlying face recognition in human infants. *Journal of cognitive neuroscience, 14*(2), 199-209.
Hamker, F. H. (2003). The reentry hypothesis: linking eye movements to visual perception. *Journal of Vision, 3*(11), 14-14.
Hanson, D. (2006). *Exploring the aesthetic range for humanoid robots.* Paper presented at the Proceedings of the ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science.
Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural brain research, 32*(3), 203-218.
Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological psychiatry, 51*(1), 59-67.
Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40*(1-2), 1-19.
Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience, 17*(11), 4302-4311.
Klatzky, R. L., Martin, G. L., & Kane, R. A. (1982). Semantic interpretation effects on memory for faces. *Memory Cognition*

10(3), 195-206. doi:10.3758/bf03197630
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology, 62*(1), 621-647. doi:10.1146/annurev.psych.093008.131123
Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of face parts and face configurations: an fMRI study. *Journal of cognitive neuroscience, 22*(1), 203-211.
Looser, C. E., & Wheatley, T. (2010). The Tipping Point of Animacy: How, When, and Where We Perceive Life in a Face. *Psychological Science, 21*(12), 1854-1862. doi:10.1177/0956797610388044
Luck, S. J. (2014). An Introduction to Event-Related Potentials and Their Neural Origins. In. Cambridge MA: MIT Press.
MacDorman, K. F., Green, R. D., Ho, C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3), 695-710. doi:10.1016/j.chb.2008.12.026
MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*(3), 297-337. doi:10.1075/is.7.3.03mac
McNeil, J. E., & Warrington, E. K. (1993). Prosopagnosia: A face-specific disorder. *The Quarterly Journal of Experimental Psychology, 46*(1), 1-10.
Meijer, E. H., Smulders, F. T., Merckelbach, H. L., & Wolf, A. G. (2007). The P300 is sensitive to concealed face recognition. *International Journal of Psychophysiology, 66*(3), 231-237.
Mori, M. (1970). The Uncanny Valley. *Energy, 7*(4), 33-35.
Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine, 19*(2), 98-100. doi:10.1109/MRA.2012.2192811
Mori, M. J. E. (1970). The uncanny valley. *7*(4), 33-35.
Mouchetant-Rostaing, Y., & Giard, M.-H. (2003). Electrophysiological correlates of age and gender perception on human faces. *Journal of Cognitive Neuroscience, 15*(6), 900-910.
Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues, 56*(1), 81-103.
NeuroBehavioral Systems. (2004). Presentation. Albany, CA.
Nigam, A., Hoffman, J. E., & Simons, R. F. (1992). N400 to Semantically Anomalous Pictures and Words. *Journal of Cognitive Neuroscience, 4*(1), 15-22. doi:10.1162/jocn.1992.4.1.15
Nosek, B. A., & Lakens, D. (2014). Registered reports A Method to Increase the Credibility of Published Results. *Social Psychology, 45*(3), 137-141. doi:27/1864-9335/a000192
Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631.
Nowak, K. L., & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments, 12*(5), 481-494.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.
Panchal, N. (2017). Face of Robotics. Retrieved from https://becominghuman.ai/face-of-robotics-1bbbd78599af
Picton, T., Bentin, S., Berg, P., Donchin, E., Hillyard, S., Johnson, R., . . . Rugg, M. J. P. (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *37*(2), 127-152.
R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Redstone, J. (2013). *Beyond the Uncanny Valley: A Theory of Eeriness for Android Science Research.* Carleton University,
Rellecke, J., Sommer, W., & Schacht, A. (2013). Emotion effects on the N170: a question of reference? *Brain topography, 26*(1), 62-71.
Rescorla, M. (2017). The Computational Theory of Mind. In N. Z. Edward (Ed.), *The Stanford Encyclopedia of Philosophy*.
Rossion, B., & Jacques, C. (2008). Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain? Ten lessons on the N170. *NeuroImage, 39*(4), 1959-1979. doi:https://doi.org/10.1016/j.neuroimage.2007.10.011
Saavedra, C., & Bougrain, L. (2012). *Processing stages of visual stimuli and event-related potentials.* Paper presented at the The NeuroComp/KEOpS'12 workshop.
Schacht, A., Sommer, W. J. B., & cognition. (2009). Emotions in word and face processing: early and late cortical responses. *69*(3), 538-550.
Searle, J. R. (1980). Is the Brain's Mind a Computer Program? *Scientific American*(January), 26-31.
Suddendorf, T. (2013). *The Gap: The Science of What Separates Us From Other Animals*: Basic Books (AZ).
Taylor, M. J., Edmonds, G. E., McCarthy, G., & Allison, T. (2001). Eyes first! Eye processing develops before face processing in children. *Neuroreport, 12*(8), 1671-1676.
Teplan, M. (2002). Fundamentals of EEG measurement. *Measurement science review, 2*(2), 1-11.
Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: its causes and consequences. *Journal of clinical epidemiology, 53*(2), 207-216.
Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 49*, 433-460.
Urgen, B. A., Kutas, M., & Saygin, A. P. (2018). Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia, 114*, 181-185. doi:10.1016/j.neuropsychologia.2018.04.027
Van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. J. Moving faces, looking places. The Amsterdam Dynamic Facial Expressions Set (ADFES), from Emotion
van Vliet, M., Mühl, C., Reuderink, B., & Poel, M. (2010). *Guessing what's on your mind: using the N400 in Brain Computer Interfaces.* Paper presented at the International Conference on Brain Informatics.
Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in cognitive sciences, 14*(8), 383-388.
Wheatley, T., Kang, O., Parkinson, C., & Looser, C. E. (2012). From mind perception to mental connection: Synchrony as a mechanism for social understanding. *Social and Personality Psychology Compass, 6*(8), 589-606.
Wheatley, T., Weinberg, A., Looser, C., Moran, T., & Hajcak, G. (2011). Mind perception: Real but not artificial faces sustain neural activity beyond the N170/VPP. *PLoS ONE, 6*(3), e17960.
Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.
Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature neuroscience, 5*(3), 277.