Computers in Human Behavior Reports 10 (2023) 100288

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Uncanny valley effect: A qualitative synthesis of empirical research to assess the suitability of using virtual faces in psychological research

Anna Flavia Di Natale [a,b,*], Matilde Ellen Simonetti [a], Stefania La Rocca [a,b], Emanuela Bricolo [a,b]

[a] *Department of Psychology, University of Milano-Bicocca, Piazza Dell'Ateneo Nuovo 1, 20126, Milan, Italy*
[b] *Mind and Behavior Technological Center, Univesity of Milano-Bicocca, Milan, Italy*

ARTICLE INFO

Keywords:
Virtual faces
Uncanny valley
Affinity
Eeriness
Face perception

ABSTRACT

Recently, virtual faces are often used as stimuli to replace traditional photographs in human face perception studies. However, despite being increasingly human-like and realistic, they still present flaws in their aspects that might elicit eerie feelings in the observers, known as the Uncanny Valley (UV) effect.

The current systematic review offers a qualitative synthesis of empirical studies investigating observers' subjective experience with virtual compared to real faces to discuss the possible challenges that the UV effect poses when virtual faces are used as stimuli to study face perception.

Results: revealed that virtual faces are judged eerier than real faces. Perception of uncanniness represents a challenge in face perception research as it has been associated with negative emotions and avoidance behaviors that might influence observers' responses to these stimuli.

Also, observers perceive virtual faces as more deviating from familiar patterns than real faces. Lower perceptual familiarity might have several implications in face perception research, as virtual faces might be considered as a category of stimuli distinct from real faces and therefore processed less efficiently.

In conclusion, our findings suggest that researchers should be cautious in using these stimuli to study face perception.

1. Introduction

Faces are essential to many research areas in psychology as they provide valuable cues about the people with whom we interact. Faces indeed reveal information about age, gender, ethnicity, emotional state, and many other attributes used to guide our behaviour during verbal and non-verbal communication (McKone & Robbins, 2011). Traditionally, the stimuli used by researchers examining face processing are photographs of faces retrieved from databases available online (Gross, 2005). These databases contain portraits of unique individuals with possibly different characteristics, like age, gender, and ethnicity. In some databases, several images are available for each identity. For example, the models can be captured in different positions and emotional expressions. Nevertheless, these databases are often customized according to the researchers' specific needs and do not always fulfil the needs of other experimenters.

Conducting systematic investigations into how faces are perceived using controlled stimuli may turn out to be an arduous task. Virtual faces can provide a possible solution to this problem. Virtual faces are artificial faces generated by digital graphics programs either from scratch or by converting real photographs into 3D head models. Thanks to the advances in digital graphics technologies, it is now possible to generate highly realistic virtual faces that appear incredibly human-like. As a result, in recent years, virtual faces have been increasingly used as stimuli in psychological research to analyze human face perception (for a review, Dawel, Miller, Horsburgh, & Ford, 2021). Virtual faces might represent a versatile and flexible alternative to real faces within this context. These stimuli indeed offer several advantages over the photographs provided by the face databases. First, researchers can quickly generate a large number of stimuli with the characteristics of interest (e.g., age, gender, ethnicity). Second, virtual faces can be easily manipulated and standardized over several features, including, for example, facial proportions and expression intensity.

Despite their increasing realism and the many advantages over conventional photographs, virtual faces still present imperfections in their aspect (e.g., rendering, shadows, texture, etc.) or behaviour (movements, facial expressions, etc.) which could make them fall into the "*uncanny valley*". This concept refers to the Uncanny Valley (UV)

* Corresponding author. Department of Psychology, University of Milano-Bicocca, Piazza Dell'Ateneo Nuovo 1, 20126, Milan, Italy.
E-mail address: a.dinatale@campus.unimib.it (A.F. Di Natale).

https://doi.org/10.1016/j.chbr.2023.100288
Received 19 April 2022; Received in revised form 15 December 2022; Accepted 11 April 2023
Available online 14 April 2023
2451-9588/

hypothesis (Mori, 1970), according to which the more an artificial entity looks human, the more familiar and pleasant it feels until it appears so human that any imperfection makes it look creepy and unpleasant. MacDorman and Ishiguro (2006) empirically demonstrated this hypothesis for the first time. This pattern generates a "valley", the uncanny valley, in the curve describing the relationship between the entity's human-likeness and the observers' response. However, these two dimensions have been thought of and operationalized in several ways.

Regarding human-likeness, researchers are still debating whether this dimension should be conceptualized and measured from an objective or subjective perspective (Wang, Lilienfeld, & Rochat, 2015). From an objective standpoint, human-likeness can be manipulated in terms of changes in an entity's quantitative resemblance to a human being, that is, a change in realism. In recent work, Diel, Weigelt, and MacDorman (2021) have identified several techniques used by researchers to manipulate their stimuli to produce different human-likeness ranges. These included, among others, the use of distinct entities (e.g., androids, virtual characters, robots), morphing techniques, or realism manipulations (e.g., changes in texture resolution or facial proportions). Researchers can use these techniques to systematically modify their stimuli and test observers' reactions to the levels of the human-likeness range they created. However, research has indicated that observers' subjective perceptions of human-likeness are not proportional to objective variations of realism (Burleigh & Schoenherr, 2015; Cheetham, Suter, & Jäncke, 2011). One possible explanation is that observers evaluate the human-likeness of an entity categorically as either artificial or human. Research has indeed revealed that a logistic s-shape curve well represents perceived human-likeness judgments: across realism levels, an entity is judged artificial until it is perceived real enough to be classified as human (Burleigh & Schoenherr, 2015; Cheetham et al., 2011). These data suggest that quantitative manipulations of human-likeness might correspond to different responses in the observers. On these grounds, recent research (Kätsyri, de Gelder, & Takala, 2019) has suggested the importance of checking whether and how objective manipulations of realism subjectively influenced the perception of an entity's human-likeness when studying its influence on observers' reactions.

The subjective experience of an individual when observing an artificial entity has been defined by Mori and colleagues (1970; 2012) with the term "*shin-wakan*", which should be translated as "sense of affinity". However, the *shin-wakan* dimension has often been operationalized and measured using different indices (Ho & MacDorman, 2010). On one side, it is usually measured in terms of sense of eeriness and unpleasantness elicited by the entity. By contrast, affinity is often translated with familiarity. Concerning familiarity, Lay, Brace, Pike, and Pollick (2016) pointed out that this term could be interpreted both as a sense of closeness or as an absence of novelty. With respect to this ambiguity, Kätsyri, Förger, Mäkäräinen, and Takala (2015) proposed distinguishing correspondently between emotional valence and perceptual familiarity. Emotional valence is described by observers' affinity with the entity, and it can be described both as negative affect, or eeriness, and as positive affect, or familiarity per se, intended as a sense of closeness. Perceptual familiarity, instead, defines familiarity in terms of the absence of novelty and refers to the observers perceiving the digital entity as having similar properties to something they already know (Kätsyri et al., 2015). To date, the relationship between emotional valence and perceptual familiarity is largely unexplored.

The original UV hypothesis predicted that affinity grows as the human-likeness of the characters increases until the characters are so realistic that they create a negative peak, or valley, in the affinity dimension that then grows back, generating a U-shaped curve (Mori, 1970; see Fig. 1.).

So far, numerous reviews have examined the theoretical explanations for the existence of the UV effect (Kätsyri et al., 2015; Wang, Lilienfeld, & Rochat, 2015; Zhang, Li, Zhang, Du, Qi, & Liu, 2020). Explanations have been grouped into two main categories that describe the UV effect from either a cognitive or a perceptual viewpoint.

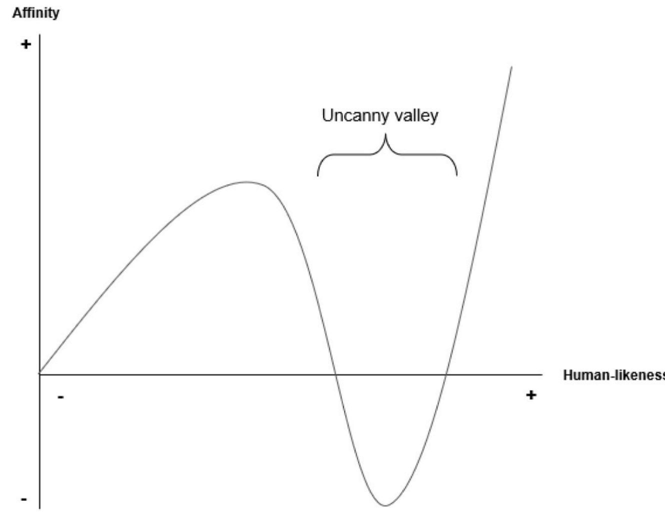

Fig 1. The Uncanny Valley Effect
An adapted graphical representation of the uncanny valley effect. According to the original theory (Mori, 1970) the feelings of familiarity and pleasantness, or *affinity* (y-axis), generated by an anthropomorphic entity increases with its increasing *human-likeness* (x-axis), until it appears so realistic that any imperfection produces a drop in observers' emotional reaction, eliciting unpleasant feelings.

Cognitive theories that consider the UV effect as a result of observers' evaluation of the stimulus and consider cognitive mechanisms, including violation of expectations and cognitive conflicts, as the primary explanations of the UV effect have been extensively reviewed in previous works (Wang, Lilienfeld, & Rochat, 2015; Zhang, Li, Zhang, Du, Qi, & Liu, 2020). Among these theories, Kätsyri et al. (2015) specifically reviewed the more recent categorization ambiguity hypothesis. According to this hypothesis, observers experience negative feelings when the task of determining whether an entity is artificial or not is difficult. This occurs when an artificial entity presents distinctive features of a given class (e.g., *human*: "has a human face"), but also specific features of another category (e.g., *robot*: "moves like a machine").

Perceptual theories, instead, consider the UV effect only a stimulus-driven effect. For example, for Kätsyri et al. (2015) any perceptual mismatch in the digital entity's appearance causes the UV effect. According to the authors, perceptual mismatches can occur in two manners. First, virtual faces might show inconsistencies in the level of realism of various aspects of their appearance due to the graphical limitations of the software used for their creation (e.g., highly realistic texture combined with low-quality features, such as very artificial hairs or eyes). Alternatively, perceptual mismatches might derive from a specific sensitivity to atypical features, such as abnormalities or exaggerated facial features (e.g., enlarged eyes or oversized foreheads). Whatever the reason, perceptual mismatch theories predict that inconsistencies in the levels of human-likeness facial features are associated with negative affinity. To date, there is no consensus on whether cognitive or perceptual mechanisms better explain the UV effect, but recent works found more robust support for perceptual explanations (Diel & MacDorman, 2021; Kätsyri et al., 2015).

In addition, the original formulation of the UV, according to which the relationship between human-likeness and observers' affinity is characterized by a decrease in observers' affinity toward highly realistic human-like entities, has recently been questioned (for a review, see Kätsyri et al., 2015). Kätsyri et al. (2015) found only a few studies in support of this hypothesis (Seyama & Nagayama, 2009; Yamada, Kawabe, & Ihaya, 2013), while most studies revealed an *uncanny slope effect* (Kätsyri et al., 2019), meaning that affinity increased with increasing human-likeness, without negative peaks (Burleigh, Schoenherr, & Lacroix, 2013; Carter, Mahler, & Hodgins, 2013; Looser &

Wheatley, 2010; MacDorman, Green, Ho, & Koch, 2009; McDonnell, Breidt, & Bülthoff, 2012). Researchers are still investigating the mechanisms that could be responsible for these variations in the pattern of responses and several explanations have been suggested.

A recent meta-analysis (Diel et al., 2021) investigated the relationship between multiple methods of manipulating human-likeness and the occurrence of the UV effect. This work is particularly relevant as it provides valuable design recommendations for future research discussing the kind of human-likeness manipulations that mainly affect observers' responses to artificial entities. Differently from Diel and colleagues' work (2021), which observed the UV effect on different artificial entities, the present review focused on a particular set of stimuli: virtual faces. The goal of the current systematic review was to collect and analyze empirical evidence on the UV effect to discuss its implications when using virtual faces as proxies for real faces in psychological research. While our initial intention was to evaluate the possibility of combining results in quantitative analysis, our search resulted in a restricted pull of studies with very heterogenous experimental designs and measures. A similar pattern was observed by Diel et al. (2021), who indeed stated that this issue complicated the interpretation of their meta-analytical results. Given the specific set of stimuli, we considered our search resulted in a limited number of papers which showed a highly methodological heterogeneity. Therefore a meta-analysis was excluded. This did not prevent us from conducting a qualitative analysis of the papers, which still allowed us to evaluate the impact of the UV effect on virtual faces' suitability in psychological research. Therefore, we first explored how virtual faces' human-likeness has been objectively manipulated and subjectively measured in published studies (RQ1a). We then analyzed how affinity has been conceptualized and measured (RQ1b). Secondly, we discussed the relationship between human-likeness and affinity in association with the original formulation of the UV hypothesis (Mori, 1970) and the newer *uncanny slope hypothesis* (Kätsyri et al., 2019) (RQ2). Finally, we analyzed the theories that best explain observers' experience with virtual faces (RQ3). Results were interpreted to convey a deeper understanding of the possible implications of using virtual faces as alternatives to real faces in psychological research.

2. Method

We conducted a systematic review following the PRISMA guidelines (Page et al., 2021). The protocol for this review was pre-registered on May 17th, 2021, in the OSF registries (Di Natale, Simonetti, La Rocca, & Bricolo, 2021). The protocol was followed rigorously, with the addition of an exclusion criterion (exclusion criterion 3).

2.1. Data sources and search strategy

We used the following electronic databases for our search: Scopus, PsychInfo, Web of Science, IEEE Xplore and Pubmed. We used several databases because the topic is highly multidisciplinary, and virtual faces are used in various research contexts. The search was performed for titles, abstracts, and keywords with the following search string: (face OR replica OR agent OR character OR entity OR animation OR human OR avatar OR morph) AND (uncann* OR eeri*). Data extraction was performed on May 19th, 2021.

2.2. Selection criteria and selection process

For the goals of our review, we included: (1) studies that assessed one or more variables measuring affinity (e.g., eeriness, familiarity, affinity) either as primary or as a secondary outcome; and (2) studies that used any type of virtual faces as stimuli. Furthermore, we excluded: (1) studies that did not report empirical data on any affinity measure; (2) studies that have used digital entities other than virtual faces (e.g., robots, full-body avatars, puppets); (3) studies that did not evaluate affinity for real human faces (this criterion was added to guarantee the possibility to compare results between virtual and human faces); and (4) papers in languages other than English, reviews, meeting abstracts, notes, letters to the editor, research protocols, patents, editorials, and other editorial materials.

After the records extraction from the databases (n = 3383), we removed all duplicates semiautomatically (n = 1184). We uploaded the remaining references (n = 2199) on Rayyan (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016), a web-based collaboration platform used for the article selection process in systematic reviews, and we performed on them a two-step screening. In step 1, two researchers independently screened titles and abstracts and excluded irrelevant references that did not meet our eligibility criteria. A third researcher resolved any conflicts. Titles and abstract screening excluded 2063 records. In step 2, all the researchers independently screened the full texts of the references selected in step 1 (n = 135) against the eligibility criteria. The screening process was blind to minimize bias during the selection process. To this aim, we selected in Rayyan the option "blind status on" which makes collaborators unable to see the screening decisions of the other collaborators. Once all the researchers had completed the selection process, we changed the blinding status to off, resolving disagreements through consensus. A total of 12 articles were eligible at the end of the process.

2.3. Data collection process and quality assessment

All authors independently extracted data from all studies in the selected articles. Where necessary, we contacted study authors to resolve any uncertainty on data. In particular, we collected information about sample size, study design and methods, details on the software and procedures used to render the virtual faces and information on the outcome measures used to evaluate affinity. We then confronted the results and used these data to perform our quality assessment of individual studies.

We assessed the risk of bias of each study using a customized form (Kätsyri et al., 2015) to identify the possible methodological limitations that could have affected the robustness of the studies' conclusions. Criteria are listed in Table 1., while Table 2 shows the quality assessment for each included study. The "x" mark indicated that the study might have encountered the specified risk of bias.

Only criterion 1 was used as an exclusion criterion given that insufficient or inappropriate statistical analyses would have made it impossible to establish a cause-and-effect relationship between the measured variables.

3. Results

A total number of 12 articles, including 14 studies, were selected. However, after quality assessment (see Table 2.), we excluded from the qualitative synthesis Bagdasarian and colleagues' (2020), Green and colleagues' (2008), and MacDorman and colleagues' (2013) works as authors failed to provide adequate statistical support to draw conclusions relevant to our analysis. The flowchart of the search is represented in Fig. 2.

Eleven studies were included for qualitative synthesis (for a summary, see Table 3).

The first goal of the present review was to explore the relationship between virtual faces' human-likeness and observers' subjective experience, or affinity, with virtual faces (RQ1).

To do so, we first explored how the human-likeness dimension was objectively manipulated and subjectively measured in the selected studies (RQ1a).

Regarding objective manipulations of human-likeness, researchers used different techniques, including the use of distinct entities, morphing and mismatch techniques (Diel et al., 2021).

Four studies used distinct entities, a human face and a virtual face,

Table 1
Risk of bias criteria. The table shows the customized risk of bias criteria applied to identify possible flaws in study design that could have influenced studies' results.

Evaluation criteria	Code	Explanation	Threat
Poor or inadequate statistical analysis *	QA1*	The authors did not include a proper statistical test.	Impossible to verify if there is a cause-and-effect relationship between the experimental variables.
No manipulation check for human-likeness	QA2	Researchers failed to check whether their stimuli's realism manipulations corresponded to different subjective evaluations of human-likeness.	The stimuli chosen are not suitable for detecting the relationship between subjective human-likeness and affinity.
Image morphing artifacts	QA3	Stimuli presented artifacts produced by human-likeness manipulations or morphing processes.	Morphing stimuli could produce flaws, especially in the morphs in the middle of the range, that could affect observers' judgments.
Limited construct representation – 1	QA4	Researchers only included a limited number of characters.	The results could be limited to the specific stimuli used and not generalizable to the whole virtual face category.
Limited construct representation – 2	QA5	Researchers only included a limited range of human-likeness	The results could be limited to the specific human-likeness range.
Virtual experience not measured	QA6	Previous experience with digital entities has not been evaluated nor considered in analyzing observers' affinity experience with virtual faces.	Perceptual experience with virtual faces could influence observers' judgments of these stimuli.

Table 2
Quality assessment of individual studies.

First author Year	QA1[a]	QA2	QA3	QA4	QA5	QA6
Bagdasarian et al., 2020, May	x					
Chattopadhyay 2016			x	x	x	x
Cheetham 2014 (Study 2)			x		x	
Cheetham 2015 (Study 2)			x		x	
Green, MacDorman, Ho, & Vasudevan, 2008	x					
Kätsyri, 2018 (Study 2)			n.a.		x	
Kätsyri 2019 (Study 1)			x			x
Kätsyri 2019 (Study 2)			x			x
Kätsyri 2019 (Study 3)			x			x
MacDorman 2013 (Study 3)	x					
MacDorman 2016			x	x	x	x
Tinwell 2011a			n.a.	x	x	
Tinwell 2011b			n.a.	x	x	
Tinwell 2013			n.a.	x	x	x

**n.a. (not applicable).
[a] Criteria QA1 excluded papers from qualitative synthesis as they failed to provide adequate statistical analysis.

generating only two levels of human-likeness (Kätsyri, 2018; Tinwell, Grimshaw, & Nabi, 2011; Tinwell, Grimshaw, Nabi, & Williams, 2011; Tinwell, Nabi, & Charlton, 2013). Specifically, Tinwell and colleagues (2011a, 2011b, 2013) used a human actor and a character taken from a commercial video game. Kätsyri (2018), on the other side, used human faces and their virtual counterparts by taking different photographs retrieved from face databases and converting each identity into a virtual replica using computer graphics software.

Other studies used morphing techniques, first generating a virtual replica of a human face and then using the human and the virtual face as extremes of a morphing continuum, thus producing multiple levels of human-likeness. In particular, Cheetham, Suter, and Jancke (2014) created 11 morphs and 13-morphs (Cheetham, Wu, Pauli, & Jancke, 2015) continua of human likeness in which the first step represented the virtual face endpoint, and the final step represented the human endpoint. Kätsyri et al. (2019) further extended the human-likeness range to include 2D illustrations of the faces. In study 1 of their paper, they used real faces derived from face databases as source images to generate two virtual variants using several graphical programs and two painted variants, with and without shading cues. In experiment 2 of their paper, they created two separate nine-morphs continua by using painted-human or virtual-human faces as extremes. In experiment 3, they used ten-morphs painted-human continua and ten-morphs painted-virtual-human continua.

Finally, two studies (Chattopadhyay & MacDorman, 2016; MacDorman & Chattopadhyay, 2016) used a mismatch technique. In their studies, the authors used entities with three levels of anthropomorphism: objects, animals, and humans. For each entity, they created a virtual replica. They then identified two sets of features (e.g., for the human entities: feature set 1 consisted of eyes, eyelashes and mouth, and feature set 2 comprised skin, nose, and eyebrow). Next, they manipulated the degree of realism of the entity by overlaying the original picture with its replica and varying the opacity of the areas corresponding to each feature set.

Despite using various objective manipulations of human-likeness, each study included a measure of observers' perceived human-likeness. All studies except two (Cheetham et al., 2014; Cheetham et al., 2015) used self-report scales to measure perceived human-likeness, using items including, for example, "*replica-original*", "*inanimate-living*", "*artificial-human-like*", "*computer-animated-real*". The remaining studies (Cheetham et al., 2014, 2015) used a two-alternative (*avatar-human*) forced choice categorization task and analyzed responses in terms of the percentage of responses categorized as human to delineate three profiles: unambiguous avatar face, ambiguous face, and unambiguous human face. Overall, results showed that observers' judgments of human-likeness correlated with stimuli's objective manipulations of realism. All the studies except one (Kätsyri et al., 2019) entered only objective manipulations of human-likeness as their independent variable in the analyses and used subjective human-likeness solely as a manipulation check for their stimuli. Only Kätsyri et al. (2019) further explored the relationship between objective and subjective human-likeness measures and found it to be nonlinear. Specifically, subjective ratings of human-likeness do not increase proportionally to corresponding levels of manipulation defined by researchers. By contrast, results showed that the relationship is better described by an s-shape logistic curve, meaning that entities are consistently evaluated as non-human until they are subjectively perceived as realistic enough to be classified as human. Therefore, according to the authors, different objective manipulations of human-likeness (e.g., various continuum endpoints, changes in the aspect, percentages of morphing) could elicit different responses in individuals' evaluations of human-likeness. Kätsyri et al. (2019) concluded that researchers should contrast subjective evaluations of virtual faces with subjective measures rather than objective manipulations of human-likeness.

We then analyzed how observers' affinity experience with virtual faces was measured (RQ1b). As regards the measure of affinity, in all the studies, the participant's task was to observe a stimulus and rate their feelings toward it.

Most studies measured affinity in terms of eeriness (Chattopadhyay & MacDorman, 2016; Kätsyri, 2018; Kätsyri et al., 2019; MacDorman & Chattopadhyay, 2016; Tinwell et al., 2013). Eeriness was measured using either semantic differential scales (Chattopadhyay & MacDorman, 2016; Kätsyri et al., 2019; MacDorman & Chattopadhyay, 2016), Likert scales (Kätsyri, 2018; Tinwell et al., 2013) or visual analogue scales

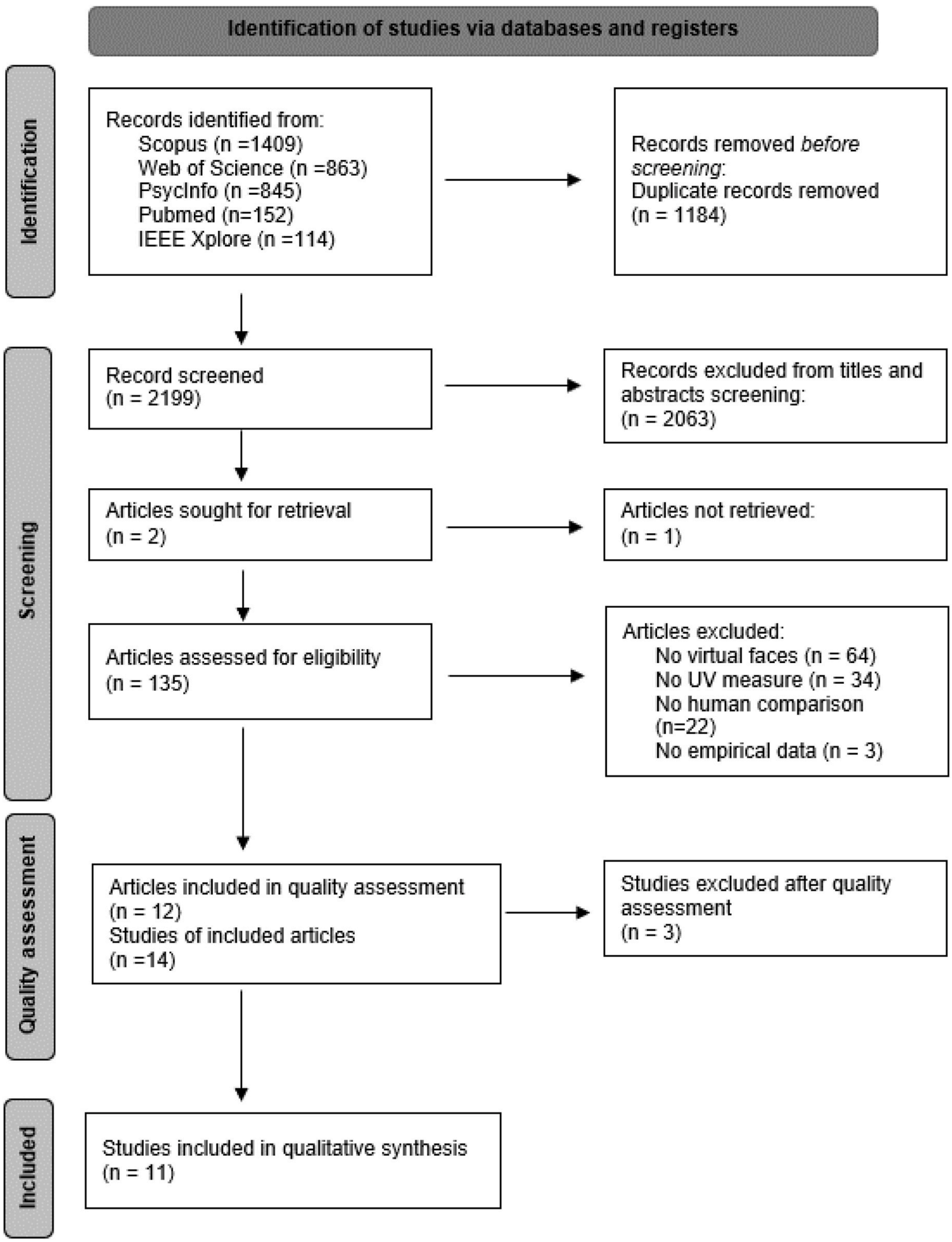

Fig 2. Selection process flowchart
The figure shows the selection process conducted following the PRISMA guidelines. *Records* refer to the title and abstract of a report indexed in a database. *Articles* refer to paper or electronic documents that provide information about one or more studies. *Studies* refer to single experiments conducted in each article.

(Kätsyri et al., 2019). Examples of the semantic differential scales anchors were "*ordinary-creepy*", "*plain-weird*", and "*predictable-eerie*". Kätsyri (2018) asked participants to evaluate how eerie virtual faces were using a 7-point Likert scale ranging from "completely disagree" to "completely agree". The visual analogue scales used by Kätsyri et al. (2019) comprised either a scale ranging from −100 ("*extremely unpleasant and creepy*") to 100 ("*extremely pleasant and not at all creepy*") or a scale ranging from −100 ("*quite creepy*") to 100 ("*quite nice*"). Other studies analyzed affinity in terms of familiarity (Chattopadhyay & MacDorman, 2016; Cheetham et al., 2014, 2015; Tinwell et al., 2011a, 2011b). To assess familiarity, researchers asked participants to rate the stimuli using Likert scales ranging, for example, from "*very strange*" to

Table 3
Summary of the studies included in the qualitative synthesis.

First author Year	Human-likeness technique	Affinity measure	UV hypothesis	Major results
Chattopadhyay 2016	Mismatch	Eeriness Familiarity	Perceptual mismatch	Reducing realism consistency decreases observers' sense of familiarity with anthropomorphic entities while increasing uncanny feelings. They also found that eerie feelings occur only when perceived familiarity is reduced.
Cheetham 2014 (Study 2)	Morphing	Familiarity	against Categorization Ambiguity	Familiarity ratings varied linearly from virtual characters to humans, with virtual faces judged less familiar than human faces.
Cheetham 2015 (Study 2)	Morphing	Familiarity	against Categorization Ambiguity	Familiarity ratings decrease when morphs move away from the human endpoint of the continua.
Kätsyri, 2018 (Study 2)	Distinct entities	Eeriness	Perceptual mismatch	Virtual faces are evaluated eerier than real faces.
Kätsyri 2019 (Study 1)	Morphing	Eeriness	Perceptual mismatch	Virtual faces elicited more negative reactions than human faces but more positive reactions than stimuli further from the human endpoint of the human-likeness continua, such as painted faces.
Kätsyri 2019 (Study 2)	Morphing	Eeriness	Perceptual mismatch	Observers' feelings of eeriness increase linearly with faces moving away from the human endpoint of the painted- virtual -human continua. However, they also found that virtual faces elicited more negative reactions in the observers than their painted and human near stimuli, suggesting that these stimuli look particularly uncanny.
Kätsyri 2019 (Study 3)	Morphing	Eeriness	Perceptual mismatch	Eerie feelings increased when human-likeness decreased, with virtual faces evaluated slightly more eerie than their painted and human near stimuli.
MacDorman 2016	Mismatch	Eeriness	Perceptual mismatch	Reducing realism consistency resulted in higher feelings of eeriness.
Tinwell 2011a	Different entities	Familiarity	Perceptual mismatch	Virtual faces are evaluated less familiar than humans and that a lack of facial expression in the upper parts of the face strengthens the uncanny valley effect for male characters.
Tinwell 2011b	Distinct entities	Familiarity	Perceptual mismatch	Exaggerating mouth movements when characters communicate emotions influences the perception of familiarity of virtual faces.
Tinwell 2013	Distinct entities	Eeriness	Perceptual mismatch	Virtual faces are evaluated less familiar than humans and that a lack of facial expression in the upper parts of the face strengthens the uncanny valley effect for both female and male characters and that perception of personality traits associated with psychopathy predict self-reported feelings of uncanny.

"*very familiar*" (Cheetham et al., 2014, 2015; Tinwell et al. 2011a, 2011b). One study further measured familiarity in terms of perceptual familiarity using semantic differential scales with the following anchors: "*rarely seen-common*", "*unfamiliar-recognizable*", and "*unique-familiar*" (Chattopadhyay & MacDorman, 2016).

These studies revealed that virtual faces are evaluated eerier (Chattopadhyay & MacDorman, 2016; Kätsyri, 2018; Kätsyri et al., 2019; MacDorman & Chattopadhyay, 2016; Tinwell et al., 2013) and less familiar (Cheetham et al., 2014, 2015; Tinwell et al., 2011a, 2011b) than real faces. Furthermore, the results of Chattopadhyay and MacDorman (2016) showed that ratings of perceptual familiarity were lower when realism inconsistency was higher.

Therefore, results showed that, overall, virtual faces were judged eerier and less familiar. Nevertheless, the fact that virtual faces elicit an adverse reaction in the observers does not necessarily corroborate the UV hypothesis. To explore the relationship between human-likeness and observers' affinity and to discuss its association with the original UV hypothesis (RQ2), we only considered those studies that included more than two levels of human-likeness (Chattopadhyay & MacDorman, 2016; Cheetham et al., 2014, 2015; Kätsyri et al., 2019; MacDorman & Chattopadhyay, 2016). It would be indeed impossible to discuss the shape of the curve describing the relationship between these two dimensions considering only two levels of human-likeness.

Most of the studies showed an increase in affinity as human-likeness increases without any negative peak as predicted by the UV hypothesis (Chattopadhyay & MacDorman, 2016; Cheetham et al., 2014, 2015; MacDorman & Chattopadhyay, 2016): the more human-like the face is, the more familiar and the less eerie it is perceived. According to these results, the relationship between human-likeness and affinity seems to be best represented by a slope that increases continuously, rather than by a curve with a negative peak, as initially hypothesized (Mori, 1970). However, since these studies manipulated the human-likeness continuum by using real faces and their virtual replica as extremes, this could have restricted the range of human-likeness and the resulting conclusions on observers' affinity. Only one study has attempted to extend the range of human-likeness by including not only virtual faces but also 2D illustrations, or painted faces (Kätsyri et al., 2019). In line with the studies cited above, the authors observed that, in general, affinity increased as human-likeness increased. However, they also found that virtual faces cause a small drop in affinity levels compared to painted and real faces. Based on this evidence, Kätsyri et al. (2019) formulated a possible alternative to the original UV hypothesis, namely the *weak UV hypothesis*. According to this hypothesis, the drop observed in affinity levels would be less remarkable than that originally proposed by Mori (1970), renamed as the *strong UV hypothesis*. This result is important as it shows that authors should carefully consider the range of human-likeness investigated in their studies before drawing conclusions on the relationship between this dimension and observers' affinity.

Several theoretical accounts have been proposed for explaining the reported effects of human likeness levels on observers' affinity (RQ3). Most researchers explained their results with perceptual mismatch theories (Chattopadhyay & MacDorman, 2016; Kätsyri, 2018; Kätsyri et al., 2019; MacDorman & Chattopadhyay, 2016; Tinwell et al., 2011a, 2011b). As suggested by Kätsyri et al. (2015), perceptual mismatch could refer either to realism inconsistency or to sensitivity to atypical features.

According to realism inconsistency theories, the UV effect is caused by a discrepancy between the level of realism of different facial features in a virtual face. An example could be a face with fine-grained skin details but very artificial hairs or eyes or vice versa. Five studies provided evidence supporting this hypothesis, according to their authors (Chattopadhyay & MacDorman, 2016; Kätsyri, 2018; all three experiments in Kätsyri et al., 2019; MacDorman & Chattopadhyay, 2016). As discussed previously, these studies used various techniques to manipulate the degree of realism of their virtual faces. Kätsyri et al. (2019) asked participants to rate a set of stimuli distributed onto a morph continuum comprising painted, virtual, and real faces of the same identity. They observed a linear relationship between human-likeness and affective response, with less human-like faces eliciting a more negative experience. The researchers speculated that this is because virtual faces present greater realism inconsistency since their facial characteristics unavoidably have several levels of realism (Chattopadhyay & MacDorman, 2017). The same explanation was given for the results of another study (Kätsyri, 2018). However, in neither study

(Kätsyri, 2018; Kätsyri et al., 2019), realism inconsistency was directly measured. On the contrary, Chattopadhyay and MacDorman (2016) directly manipulated the level of realism inconsistency of their stimuli. In their study, the authors compared the affective reactions toward entities with varying degrees of anthropomorphism (objects, animals, and humans) in their real and virtual forms. Additionally, in each stimulus, the level of realism of specific features (e.g., for the human faces: eyes, eyelashes, and mouth or skin, nose, and eyebrows) was progressively changed to generate realism inconsistencies. They then asked participants to categorize each face as either human or virtual and rate the stimuli for eeriness. Chattopadhyay and MacDorman (2016) observed that reducing realism consistency resulted in higher feelings of eeriness. Similar results using the same study design were obtained by MacDorman and Chattopadhyay (2016), who found that virtual faces were evaluated eerier than real faces.

According to the sensitivity to atypical features hypothesis, on the other end, the uncanny feelings are elicited by the presence of atypical features either in the appearance or in the character's behaviour. Three studies interpreted their results according to this hypothesis (Tinwell et al., 2011a, 2011b, 2013). In two studies Tinwell and colleagues (2011a; 2013) examined observers' affinity with real and virtual faces. Virtual faces could be either fully animated or lack emotional expressivity in the upper part of the face (eyelids). The authors found lower ratings of familiarity (Tinwell, Grimshaw, Nabi, et al., 2011) and higher ratings of eeriness (Tinwell et al., 2013) for virtual faces than real faces and further showed that virtual faces without upper facial animation were rated the least familiar (Tinwell, Grimshaw, Nabi, et al., 2011) and the uncanniest (Tinwell et al., 2013). A further study (Tinwell, Grimshaw, & Nabi, 2011) showed that exaggerated facial expressions, such as an intense movement of the mouth during a speech to portray distinct emotions, might alter the uncanny sensations elicited by the virtual characters. Tinwell, Grimshaw, and Nabi (2011) concluded that this effect is due to the fact that the behavioural fidelity of the character did not match its human-like appearance. Nevertheless, the results found with moving characters are not necessarily comparable we those found with static images. Indeed, according to the original formulation of the UV hypothesis (Mori, 1970), eerie feelings would be intensified by a character's movement (Rosenthal-Von Der Pütten & Krämer, 2014; Thompson, Trafton, & McKnight, 2011).

Only two studies (Cheetham et al., 2014, 2015) considered other possible interpretations. Specifically, the authors analyzed the categorization ambiguity (Burleigh et al., 2013; Kätsyri et al., 2015; Yamada et al., 2013) theory and found support against it. According to this theory, ambiguous virtual characters at the turn of the category boundary between artificial and human should be considered eerier and less familiar due to the difficulty in deciding whether they are human or not. Cheetham et al. (2014; 2015) tested this hypothesis, examining whether ambiguous faces were harder to categorize and discriminate and whether this caused a drop in observers' affinity ratings. The authors failed to find support for the category ambiguity theory. Specifically, they found greater discrimination capacity for more ambiguous face stimuli and did not find ambiguous faces to be considered the strangest and less familiar. Instead, they found a linear relationship between familiarity and morphing continua from virtual to real faces, suggesting that category ambiguity might not be sufficient to explain observers' negative affinity.

Furthermore, to better explain the reported effects of human likeness levels on observers' affinity, some authors suggested that differences in affinity responses to virtual faces might be modulated by observers' perceptual expertise with the various types of faces, virtual or real (Chattopadhyay & MacDorman, 2016; Kätsyri, 2018; Kätsyri et al., 2019; MacDorman & Chattopadhyay, 2016). In particular, these authors hypothesized that the negative responses of individuals toward virtual faces might be explained by perceptual narrowing. Perceptual narrowing refers to an increased perceptual sensitivity to frequently encountered stimuli (Nelson, 2001). This phenomenon characterizes the development of different abilities, such as the ability to discriminate faces to which people are more exposed during their lives. Undoubtedly, most people have less perceptual expertise with virtual than real faces. In support of this hypothesis, Kätsyri (2018) presented participants with real and virtual faces and asked them to complete a recognition memory task (experiment 2 in the paper). The study demonstrated that virtual faces were less efficiently recognized and were rated eerier than real faces. The authors interpreted the poorer performances for virtual faces on the memory task as an indicator of lower perceptual expertise with these stimuli and suggested that this may lead to the discomfort and eerie feelings observed via self-ratings. Furthermore, the results of Chattopadhyay and MacDorman (2016) and MacDorman and Chattopadhyay (2016) showed that the effect of realism inconsistency was more remarkable for anthropomorphic entities than for animals or objects. The authors suggested that since perceptual narrowing increases the ability to discriminate between faces, any deviation from the norm, that is, realism inconsistencies, appear more evident and make a face look more unfamiliar.

4. Discussion

The review revealed that virtual faces are often judged eerier and less familiar than real faces. Nevertheless, we noticed that, as suggested by Lay and colleagues (Lay et al., 2016), familiarity has been employed both in terms of affinity and perceptual familiarity. The items used to measure it (e.g., "*very strange-familiar*", or "*rarely seen-common*") indeed represent an excellent example of how familiarity can be interpreted in several ways. In the first case (Cheetham et al., 2014; Cheetham et al., 2015; Tinwell, Grimshaw, Nabi, et al., 2011; Tinwell, 2011b), familiarity seems to describe an experience of strangeness, while in the second (Chattopadhyay & MacDorman, 2016) appears to contrast feelings of familiarity with perceptions of novelty. Therefore, these indices seem to measure different aspects, namely emotional valence and perceptual familiarity, whose implications for the use of virtual faces in psychological research will be discussed separately.

First, it has been widely demonstrated that the perception of uncanniness in human-like entities is often associated with negative emotions, such as fear, disgust and anxiety and also with avoidant behaviours (Ho, MacDorman, & Pramono, 2008; MacDorman, 2006; Mathur & Reichling, 2016; Strait, Vujovic, Floerke, Scheutz, & Urry, 2015). Eeriness, in particular, has been associated with the tendency to avoid potentially threatening entities as a result of an alarm system that recognizes in anthropomorphic entities a potential danger (Sasaki, Ihaya, & Yamada, 2017). Furthermore, some authors (Balas, Tupa, & Pacella, 2018) revealed that, while virtual faces' evaluations (e.g., trustworthiness, attractiveness, dominance, confidence, etc.) are made within the same social face space as real faces (Oosterhof & Todorov, 2008), they are located in different positions, meaning that evaluations were dissimilar in virtual and real faces. In particular, some studies found that near human-like entities not only elicit eeriness feelings but also disrupt trustworthiness, attractiveness, and credibility perceptions (Balas & Pacella, 2017; Groom et al., 2009; Ho et al., 2008; Stein & Ohler, 2018; Weisman & Peña, 2021). Given that our results showed that virtual faces elicit greater sensations of eeriness and unfamiliarity than real faces and that some other face evaluations, including trustworthiness and attractiveness, might be disrupted, researchers should be cautious when using virtual faces as proxies for real faces in experimental psychological research as they might trigger feelings of discomfort and uneasiness, possibly affecting observers' judgments on some crucial facial traits considered in psychological research.

Second, one study in the present review (Chattopadhyay & MacDorman, 2016) attempted to test empirically the relationship between perceived familiarity and emotional valence. The authors fitted a structural equation model to investigate the relationship between physical (realism inconsistency), perceptual (perceived realism and familiarity), and affective (eeriness and warmth) aspects of the UV effect.

The model showed that realism inconsistencies on anthropomorphic entities lead the observers to perceive an entity as less familiar and that this impression of unfamiliarity resulted in eerie feelings. It would be interesting to investigate how these two aspects are related further. This contribution indeed showed that most of the results were interpreted from a perceptual perspective. Specifically, the authors argued that lower affinity is often associated with perceptual mismatches applied to virtual faces (e.g., changes in facial proportion, realism inconsistencies, behavioural infidelity, etc.). None of the studies considered possible alternative explanations, except for Cheetham et al. (2014; 2015), who failed to find support for the category ambiguity theory. This is in line with previous findings that found scant support for this theory but great support for perceptual mismatch theories (Diel & MacDorman, 2021; Kätsyri et al., 2015). In particular, recent work tested perceptual (configural processing, atypicality and perceptual mismatch), cognitive (category uncertainty and novelty avoidance), and evolutionary (mate selection, psychopathy avoidance, threat avoidance, empathy), and found more support for perceptual theories than cognitive and evolutionary theories.

Recent work specifically analyzed the role of perceptual familiarity on uncanniness sensitivity on real faces (Diel & Lewis, 2022a). In their experiments, Diel and Lewis (2022a) found a positive association between ratings of uncanniness and the distortion of faces with a stronger effect for familiar (famous people) than non-familiar faces. Similar results were found for greebles faces study (Diel & Lewis, 2022b) found a similar effect with greebles, (that is, virtual objects often used as stimuli in face perception research), meaning that distorted greebles were considered more uncanny only by trained participants who had previously gained familiarity with them. In this respect, it has been further suggested that observers' perceptual expertise might also modulate lower affinity with virtual faces with these stimuli (Chattopadhyay & MacDorman, 2016; Kätsyri, 2018; Kätsyri et al., 2019). Since early life, humans are exposed to a multitude of faces, gaining exceptional abilities in discriminating between identities and accurately recognizing faces (Scott, Pascalis, & Nelson, 2007). Although virtual faces are becoming increasingly realistic, they may still be distinguishable enough for the perceptual system to consider them as a separate category of stimulus. As a result, virtual faces might be processed differently from, and/or less efficiently than, real faces. In this respect, researchers (Balas & Pacella, 2015; Kätsyri, 2018) found virtual faces to be susceptible to the inversion effect to the same extent as real faces. The authors interpreted this result as an indicator that virtual faces are processed as face stimuli. The inversion effect, or the decrease in face recognition performances when faces are presented upside-down, is indeed considered a hallmark of holistic processing specific to faces (Yin, 1969). However, additional data analyzing individual recognition (Balas & Pacella, 2015; Crookes et al., 2015; Kätsyri, 2018) and discrimination (Balas & Pacella, 2015; Crookes et al., 2015) accuracy of faces revealed poorer performances for virtual faces than real faces. The authors concluded that although virtual faces are processed as faces, as indicated by the data on the inversion effect, they are processed less efficiently. Specifically, although virtual faces have a high degree of realism, they are still too different from real ones to benefit from face expertise acquired with real faces and therefore are processed less efficiently. All these results have been explained as differences in individuals' perceptual expertise with virtual faces, but without explicitly testing the potential effect of expertise. Given the increasing use of virtual faces in psychological studies as proxies for human faces and since the differences in processing have been attributed mainly to the difference in expertise with the stimulus, future experiments should directly explore the influence of perceptual expertise on virtual face elaboration.

Therefore both negative affect (eeriness and familiarity) and lower perceptual familiarity could raise challenges when using virtual faces instead of real faces in psychological research.

As a final consideration, the papers analyzed in the present review suggest that affinity increases with increasing human-likeness, without any negative peak as predicted by the original UV hypothesis (Mori, 1970). This seems to support the *uncanny slope effect* (Kätsyri et al., 2019). However, the limited human-likeness ranges used in the analyzed studies prevented us from clearly understanding the curve's shape describing the relationship between human-likeness and affinity. Our quality assessment showed that most studies measured the observers' reactions using a small range of face types, which therefore do not cover a wide range of human-likeness. The most frequently used stimulus sets included images morphed between a virtual and a real face (Chattopadhyay & MacDorman, 2016; Cheetham et al., 2014, 2015; MacDorman & Chattopadhyay, 2016) or measured observers' subjective experience using only single images of virtual and real faces (Kätsyri, 2018; Tinwell et al., 2011a, 2011b, 2013). Only one study has attempted to expand the human-likeness range by including 2D illustrations (Kätsyri et al., 2019). However, using only virtual and real faces, either singularly or as extremes of a morphing, restricts the range of human-likeness as these stimuli might represent two very close points in the continuum. Consequently, as Kätsyri et al. (2019) suggested, it is difficult to understand where virtual faces are located in the original UV curve and establish where in human-likeness continuum virtual faces fall. Indeed, if the endpoint images are too similar, the x-axis (human-likeness) may not include the uncanny valley.

Also, some authors (Ho & MacDorman, 2010, 2017) suggested that the subjective measures of human-likeness and affinity usually found in the literature may be biased as human-like characters are perceived categorically (Chattopadhyay & MacDorman, 2016; Cheetham et al., 2011; Cheetham, Pavlovic, Jordan, Suter, & Jancke, 2013; MacDorman & Chattopadhyay, 2016; MacDorman, Vasudevan, & Ho, 2009). In this regard, the authors (Ho & MacDorman, 2010, 2017) developed an alternative set of indices (humanness, warmth, eeriness and attractiveness) that could be used in future research to disentangle this bias.

4.1. Limitations

The studies included in the present review present some limitations. First, the studies' conclusions result from the analysis of very heterogeneous stimuli. Some studies used morphed stimuli from virtual faces to real faces (Cheetham et al., 2014, 2015; Kätsyri, 2018), or from painted to real faces (Kätsyri et al., 2019). Further studies used faces with inconsistencies in the level of realism of different features (Chattopadhyay & MacDorman, 2016; MacDorman & Chattopadhyay, 2016), and three others used single virtual faces compared against real faces (Tinwell et al., 2011a, 2011b, 2013). Stimuli manipulations also differed in technical features, including cropping and rendering. Some of them were presented with concealed hairs and ears (Chattopadhyay & MacDorman, 2016; Cheetham et al., 2014, 2015; Kätsyri, 2018; Kätsyri et al., 2019; MacDorman & Chattopadhyay, 2016) while others not (Tinwell et al., 2011a, 2011b, 2013). Thus, both the creation and manipulation processes did not follow a consistent method, making it impossible to compute equivalent levels of human-likeness across studies. While this limitation does not allow direct comparison across studies, it has to be acknowledged that stimuli heterogeneity can lead to a larger external validity.

Second, most studies have used self-report measures to assess observers' affinity for virtual faces. These measures often include a series of questions measured on Likert scales, visual analogue scales, or semantic differential scales to evaluate observers' affinity with virtual characters. However, self-reported measures may not be valid measures and are often largely heterogeneous across studies (Diel et al., 2021; Kätsyri et al., 2015). Therefore, it would be desirable to use additional measures, for example, implicit or physiological measures. The UV effect has been argued to be an automatic perceptual judgment that arises as soon as the observer perceives the stimulus (MacDorman, Green, et al., 2009). It could be interesting to use implicit measures to study first impressions occurring automatically within a short exposure time. Physiological measures as well could be interesting in analyzing the UV phenomenon. Cheetham et al. (2015), for example, used late positive potentials and electromyography to measure participants' affective states while observing virtual faces and found that these measures converged with self-report measures indicating a general negative affective state with virtual faces.

Third, even though some studies suggested an important role of perceptual expertise in virtual face evaluation, none adequately measured participants' experience with virtual faces. Only a few studies reported controlling for participant experience by asking participants to confirm they had little or no experience with video games virtual characters or with generating and designing computer-generated models (Cheetham et al., 2014, 2015; Kätsyri, 2018). Other studies (Tinwell et al., 2011a, 2011b) tried to control potential differences in participants' responses by including video designer students supposed to have almost the same knowledge of virtual faces. No study explicitly analyzed the influence of experience on observers' affinity judgements.

Finally, given the rapid advances in computer technology, new graphics programs are now available on the market, capable of generating hyper-realistic virtual faces, which are becoming even more difficult to distinguish from real humans. For example, Dawel et al. (2021) identified a new kind of unexplored stimuli, that is, virtual faces created with Generative Adversarial Networks (GANs). These faces are generated by an artificial intelligence system trained on human photographs that creates incredibly realistic virtual human faces. While it is argued that realism inconsistencies are unavoidable in virtual faces created with graphical software, it could be possible that this more advanced technology could overcome this problem.

5. Conclusions

The purpose of this study was to review research studies exploring observers' reactions to virtual faces. In particular, we were interested in investigating whether virtual faces are judged eerier and less familiar than real faces.

The present review focused on virtual faces, thus considering stimuli from the top of the head to the shoulders. Therefore, our conclusions refer specifically to these stimuli. We chose to investigate exclusively faces because we wanted to specifically examine the feasibility of using these stimuli in research experiments studying human face processing mechanisms. However, it would be interesting in the future to expand this investigation to full-body virtual humans as they are increasingly used in other experimental psychology settings.

Unfortunately, the stimuli and measures used in the included studies were highly heterogeneous, preventing us from aggregating results and conducting a quantitative synthesis of the data. For this reason, we encourage researchers to agree on and follow the proposed design principles and recommendations suggested in previous work (Diel et al., 2021; Lay et al., 2016). In particular, there is the need to standardize stimuli creation techniques and to align the measures used to evaluate the different aspects of the observers' experience with virtual faces to make the results comparable.

From a practical viewpoint, we suggest that researchers carefully consider their research goals before using virtual faces in their studies. Considerations about the level of realism achievable from available programs and the possible hindrances resulting from the use of these stimuli should be done rigorously before including virtual faces in experimental settings and investigating human face perception and evaluation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Bagdasarian, M. T., Hilsmann, A., Eisert, P., Curio, G., Müller, K. R., Wiegand, T., et al. (2020). EEG-based assessment of perceived realness in stylized face images. May. In *2020 twelfth international Conference on Quality of multimedia experience (QoMEX)* (pp. 1–4). IEEE.

Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in Human Behavior, 52*, 331–337.

Balas, B., & Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior, 77*, 240–248.

Balas, B., Tupa, L., & Pacella, J. (2018). Measuring social variables in real and artificial faces. *Computers in Human Behavior, 88*, 236–243.

Burleigh, T. J., & Schoenherr, J. R. (2015). A reappraisal of the uncanny valley: Categorical perception or frequency-based sensitization? *Frontiers in Psychology, 5*, 1488.

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior, 29*(3), 759–771.

Carter, E. J., Mahler, M., & Hodgins, J. K. (2013). Unpleasantness of animated characters corresponds to increased viewer attention to faces. August. In *Proceedings of the acm symposium on applied perception* (pp. 35–40).

Chattopadhyay, D., & MacDorman, K. F. (2016). Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *Journal of Vision, 16* (11), 7-7.

Cheetham, M., Pavlovic, I., Jordan, N., Suter, P., & Jancke, L. (2013). Category processing and the human likeness dimension of the uncanny valley hypothesis: Eye-tracking data. *Frontiers in Psychology, 4*, 108.

Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": Behavioral and functional MRI findings. *Frontiers in Human Neuroscience, 5*, 126.

Cheetham, M., Suter, P., & Jancke, L. (2014). Perceptual discrimination difficulty and familiarity in the uncanny valley: More like a "happy valley". *Frontiers in Psychology, 5*, 1219.

Cheetham, M., Wu, L., Pauli, P., & Jancke, L. (2015). Arousal, valence, and the uncanny valley: Psychophysiological and self-report findings. *Frontiers in Psychology, 6*, 981.

Crookes, K., Ewing, L., Gildenhuys, J. D., Kloth, N., Hayward, W. G., Oxner, M., et al. (2015). How well do computer-generated faces tap face expertise? *PLoS One, 10*(11), Article e0141353.

Dawel, A., Miller, E. J., Horsburgh, A., & Ford, P. (2021). A systematic survey of face stimuli used in psychological research 2000–2020. *Behavior Research Methods*, 1–13.

Di Natale, A. F., Simonetti, M. E., La Rocca, S., & Bricolo, E. (2021). *The uncanny valley effect on computer-generated faces: A protocol for a systematic review*. May 17. https://doi.org/10.17605/OSF.IO/4EXS8.

Diel, A., & Lewis, M. (2022a). Familiarity, orientation, and realism increase face uncanniness by sensitizing to facial distortions. *Journal of Vision, 22*(4), 14-14.

Diel, A., & Lewis, M. (2022b). The deviation-from-familiarity effect: Expertise increases uncanniness of deviating exemplars. *PLoS One, 17*(9), Article e0273861.

Diel, A., & MacDorman, K. F. (2021). Creepy cats and strange high houses: Support for configural processing in testing predictions of nine uncanny valley theories. *Journal of Vision, 21*(4), 1-1.

Diel, A., Weigelt, S., & MacDorman, K. F. (2021). A meta-analysis of the Uncanny Valley's independent and dependent variables. *ACM Transactions on Human-Robot Interaction (THRI), 11*(1), 1–33.

Green, R. D., MacDorman, K. F., Ho, C. C., & Vasudevan, S. (2008). Sensitivity to the proportions of faces that vary in human likeness. *Computers in Human Behavior, 24* (5), 2456–2474.

Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J. K., & Robles, E. (2009). Evaluating the effects of behavioral realism in embodied agents. *International Journal of Human-Computer Studies, 67*(10), 842–849.

Gross, R. (2005). Face databases. In *Handbook of face recognition* (pp. 301–327). New York, NY: Springer.

Ho, C. C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior, 26*(6), 1508–1518.

Ho, C. C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect. *International Journal of Social Robotics, 9*(1), 129–139.

Ho, C. C., MacDorman, K. F., & Pramono, Z. D. (2008). Human emotion and the uncanny valley: A GLM, MDS, and isomap analysis of robot video ratings. March. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction* (pp. 169–176).

Kätsyri, J. (2018). Those virtual people all look the same to me: Computer-rendered faces elicit a higher false alarm rate than real human faces in a recognition memory task. *Frontiers in Psychology, 9*, 1362.

Kätsyri, J., de Gelder, B., & Takala, T. (2019). Virtual faces evoke only a weak uncanny valley effect: An empirical investigation with controlled virtual face images. *Perception, 48*(10), 968–991.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology, 6*, 390.

Lay, S., Brace, N., Pike, G., & Pollick, F. (2016). Circling around the uncanny valley: Design principles for research into the relation between human likeness and eeriness. *I-Perception, 7*(6), Article 2041669516681309.
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.
MacDorman, K. F. (2006). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. July. In *ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science, 4*.
MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition, 146*, 190–205.
MacDorman, K. F., Green, R. D., Ho, C. C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior, 25* (3), 695–710.
MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*(3), 297–337.
MacDorman, K. F., Srinivas, P., & Patel, H. (2013). The uncanny valley does not interfere with level 1 visual perspective taking. *Computers in Human Behavior, 29*(4), 1671–1685.
MacDorman, K. F., Vasudevan, S. K., & Ho, C. C. (2009). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & Society, 23*(4), 485–510.
Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition, 146*, 22–32.
McDonnell, R., Breidt, M., & Bülthoff, H. H. (2012). Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics, 31*(4), 1–11.
McKone, E., & Robbins, R. (2011). *Are faces special* (pp. 149–176). Oxford handbook of face perception.
Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy, 7*, 33–35.
Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics and Automation Magazine, 19*(2), 98–100.
Nelson, C. A. (2001). The development and neural bases of face recognition. *Infant and Child Development: An International Journal of Research and Practice, 10*(1-2), 3–18.
Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087–11092.
Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews, 5*(1), 1–10.
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372.
Rosenthal-Von Der Pütten, A. M., & Krämer, N. C. (2014). How design characteristics of robots determine evaluation and uncanny valley related responses. *Computers in Human Behavior, 36*, 422–439.
Sasaki, K., Ihaya, K., & Yamada, Y. (2017). Avoidance of novelty contributes to the uncanny valley. *Frontiers in Psychology, 8*, 1792.
Scott, L. S., Pascalis, O., & Nelson, C. A. (2007). A domain-general theory of the development of perceptual discrimination. *Current Directions in Psychological Science, 16*(4), 197–201.
Seyama, J. I., & Nagayama, R. S. (2009). Probing the uncanny valley with the eye size aftereffect. *Presence: Teleoperators and Virtual Environments, 18*(5), 321–339.
Stein, J. P., & Ohler, P. (2018). Uncanny… but convincing? Inconsistency between a virtual agent's facial proportions and vocal realism reduces its credibility and attractiveness, but not its persuasive success. *Interacting with Computers, 30*(6), 480–491.
Strait, M., Vujovic, L., Floerke, V., Scheutz, M., & Urry, H. (2015). Too much humanness for human-robot interaction: Exposure to highly human-like robots elicits aversive responding in observers. April. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3593–3602).
Thompson, J. C., Trafton, J. G., & McKnight, P. (2011). The perception of humanness from the movements of synthetic agents. *Perception, 40*(6), 695–704.
Tinwell, A., Grimshaw, M., & Nabi, D. (2011). Effect of emotion and articulation of speech on the uncanny valley in virtual characters. October. In *International Conference on affective Computing and intelligent interaction* (pp. 557–566). Berlin, Heidelberg: Springer.
Tinwell, A., Grimshaw, M., Nabi, D. A., & Williams, A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior, 27*(2), 741–749.
Tinwell, A., Nabi, D. A., & Charlton, J. P. (2013). Perception of psychopathy and the Uncanny Valley in virtual characters. *Computers in Human Behavior, 29*(4), 1617–1625.
Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology, 19*(4), 393–407.
Weisman, W. D., & Peña, J. F. (2021). Face the uncanny: The effects of doppelganger talking head avatars on affect-based trust toward artificial intelligence technology are mediated by uncanny valley perceptions. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 182–187.
Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. *Japanese Psychological Research, 55*(1), 20–32.
Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81* (1), 141.
Zhang, J., Li, S., Zhang, J. Y., Du, F., Qi, Y., & Liu, X. (2020). A literature review of the research on the uncanny valley. July. In *International conference on human-computer interaction* (pp. 255–268). Cham: Springer.