Face Perception and Hyper-Realistic Masks

Jet Gabrielle Sanders

Doctor of Philosophy

University of York

Psychology

July 2018

Abstract

Previous research has shown that deliberate disguise deteriorates human and automatic face recognition, with consequences for person identification in criminal situations. Common forms of deliberate disguise (e.g. balaclavas or hoodies) are easy to detect. When such disguises are used, viewer can distinguish between an unmasked individual – whose identity they knowingly can observe from facial appearance – and a masked individual – whose identity they knowingly cannot. Hyper-realistic silicone masks change this. Their recent use in criminal settings suggests that they effectively disguise identity and are difficult to detect. In this thesis, I first show that viewers are strikingly poor at distinguishing hyper-realistic masks from real faces under live and photographic test conditions, and are worse in other-race conditions. I also show large individual differences in discriminating realistic masks from real faces (5%-100% accuracy), and use an image analysis to isolate information that high performers use for effective categorisation. The analysis reveals an informative region directly below the eyes, which is used by high performers but not low performers. These findings point to selection and training as routes to improved mask detection. Second, I examine the reliability of estimates made of the person beneath the mask. Demographic profiling and social character estimates are poor, and results show that recognition rates were only just above chance, even for familiar viewers. This analysis highlights a systematic bias in these estimates: demographics, traits and social characteristics of the mask were attributed to those of the wearer. This bias has theoretical and applied consequences. First, it supports the automaticity with which viewers use a face to judge a person, even when they know the face is not that of the person. Second, it suggests that predictions of the person underneath the mask, by familiar and unfamiliar viewers alike, should be treated with great caution.

Table of Content

List of Tables

List of Figures

Acknowledgements

First, I sincerely thank my supervisor Dr. Rob Jenkins. Your enthusiasm and calm ushered me through the many ventures I undertook over the course of my PhD, whether relevant to this thesis or not. No one ever gave me as much confidence that I am capable.

Second, I thank the Kokoro Research Centre for their generous welcome every time I visited armed with many masks. Your support of my research in Japan was indispensable.

I also thank the many wise at the Psychology department whom offered advice – and most of all the FaceVar Lab group. You helped me refine my thoughts and expression thereof far beyond what I could have imagined.

I am especially grateful to the nearest and dearest of you who listened to my rummaging thoughts and read my drafts from afar. Whether you are here to see the finished product or not, this is a compilation of you all.

Finally, I would like to thank all confederates who patiently wore my hyper-realistic masks for photographs or under live experimental conditions. Without you this thesis truly would not have been possible.

Financial Support

I am very grateful for the funding I received to complete this PhD by the Economic and Social Research Council 1+3 Studentship (Studentship ES/J500215/1) and the Prins Bernhard Cultuurfonds (grant E/30/30.13.0630/HVH/IE).

In addition, I want to thank the funding bodies who supported research visits to Japan: University of York International Seedcorn Award, Overseas Fieldwork Expenses by the White Rose Doctoral Training Centre (A0158430)

made out to me, and the University of York Research Priming Fund (H0022034) made out to my supervisor, Dr Rob Jenkins

Declaration

I declare that this thesis is my own work carried out under normal terms of supervision. This work has not been previously presented for an award at this, or any other University. All quotations in this thesis have been distinguished by quotation marks and they have been attributed to the original source. All sources are acknowledged as References.

Collaborations

Chapter 2. Dr. Yoshiyuki Ueda and Prof. Sakiko Yoshikawa supported the design of the outdoor experiment in Japan and facilitated the translation of instructions for all three. Dr. Eilidh Noyes and Kazusa Minemoto supported data collection.

Chapter 3. Dr. Yoshiyuki Ueda and Prof. Sakiko Yoshikawa supported data collection in Japan and translated instructions for both experiments.

Chapter 5. Dr. Yoshiyuki Ueda and Prof. Sakiko Yoshikawa supported design adjustment of the experiment in Japan.

Submitted for publication

Sanders, J. G., & Jenkins, R. (2018). Individual differences in hyper-realistic mask detection. *Cognitive Research: Principles and Implications, 3(1), 24.*

Sanders, J. G., Ueda, Y., Minemoto, K., Noyes, E., Yoshikawa, S., & Jenkins, R. (2017). Hyper-realistic face masks: a new challenge in person identification. Cognitive research: principles and implications, 2(1), 43.

Sanders, J.G., Ueda, Y., Yoshikawa, S., & Jenkins, R. (under review). *More human than human: a Turing test for synthetic faces.*

Sanders, J.G., Ueda, Y., Yoshikawa, S., & Jenkins, R. (under review). *Demographic profiling through a hyper-realistic face mask.*

Conference presentations and proceedings

Sanders, J. G., Minemoto, K., Ueda, Y., Yoshikawa, S., & Jenkins, R. (2016). Other-race effect in hyper-realistic mask detection: a new challenge for facial identification. *International Journal of Psychology, 51*, 169.

Sanders, J.G. (2016) Other-race effects in hyper-realistic mask detection: a new challenge for facial identification. *International Congress of Psychology,* August 2016, Yokohama, Japan.

Sanders, J.G. & Jenkins, R. (2015) Hyper-realistic masks: a new challenge for international security. *The Applied Face Meeting,* June 2015, York, United Kingdom.

Sanders, J.G. Byrne, A., Tominaga, A., Ueda, Y., Minemoto, K., Yoshikawa, S. & Jenkins, R. (2017) The psychological effect of masks on the wearer: a new testbed for embodied cognition, *Art & Perception, 5*, 337–426.

Sanders, J.G. Ueda, Y., Minemoto, K., Noyes, E. Yoshikawa, S. & Jenkins, R. (2017) Perception of hyper-realistic face masks, *European Conference on Visual Perception*, September 2017, Berlin, Germany.

Sanders, J.G. Ueda, Y., Minemoto, K., Noyes, E. Yoshikawa, S. & Jenkins, R. (2016) The psychological effect of masks on the wearer: a new testbed for embodied cognition, *Experimental Psychology Society,* London Meeting, January 2016, London, United Kingdom.

Sanders, J.G., (2017) Who is beneath the mask? *North East Person Perception Meeting*, October 2017, Durham, United Kingdom.

Chapter 1.

General Introduction

1.1 The face identification problem

Face identification or face recognition refers to the use of facial appearance to verify the identity of a specific individual. Face recognition is the most common means of identity verification and is crucial in security and criminal justice settings. Face recognition is also highly error prone. Although each face is different, humans are not always able to detect identity differences. I will outline this issue by considering research on identification from memory and face matching separately, by exemplifying and discussing factors that mediate face recognition ability.

Face identification from memory

The Innocence Project is a public policy organisation set up in 1992 dedicated to exonerating wrongfully convicted individuals through DNA testing (The Innocence Project, 2016). In 2012, they found that 73% of all exonerated individuals (n=175) were wrongfully convicted through eyewitness misidentification (Pezdek, 2012). These testimonies were from individuals whom directly viewed the perpetrator, yet misidentified the suspect in police questioning or line-up identification at a later time point. Twenty-five percent of these cases involved two or more eyewitnesses. This makes eyewitness misidentification the largest known contributor to false conviction of innocent individuals (Pezdek, 2012).

Memory biases are key contributors to misidentification. One of the fundamental insights from memory research is that memories are not copies of exact experience but rather reconstructed in light of previous knowledge and exposure (Bartlett, 1932; Bartlett & Burt, 1933; Hirt, McDonald & Markman, 1998). This leaves room for situational, social and cognitive bias. To give a few examples, studies confirm reduced encoding accuracy with reduced exposure

time (Ellis, Davies & Shepherd, 1977), in high-stress conditions (Deffenbacher et al., 2004), and during night time viewing (Yarmey, 1986). Recall accuracy biases include line-up administrator awareness of whom the suspect is (Phillips et al., 1999; Russano et al., 2006; Wells et al., 1998), suspect clothing in a line-up (Dysart, Lindsay & Dupuis, 2006; Lindsay, Wallbridge & Drennan, 1987) and describing the culprit's face during eyewitness questioning (RRR of Schooler & Engstler-Schooler (1990): Alogna et al., 2014).

This is a difficult problem, but lots of effort has been put into real-world measures to counteract effects of some of these biases. For example, some justice systems are slowly shifting towards using evidence-based protocols to reduce line-up biases (Wells, Steblay & Dysart, 2012). Recommendations include taking into account participant confidence levels, their viewing situations and minimising investigator biases during questioning and line-up. Note that this is a slow movement and eyewitnesses are still considered to be amongst the most solid types of evidence by judges in criminal cases (McGrath & Turvey, 2014). More importantly, resolving memory biases does not resolve the misidentification problem.

Face identification from face-matching

Identity mismatching in recognition from face matching indicates that there is also a perceptual component to misidentification. For example, the disappearance of Malaysian Airlines flight MH370 from Kuala Lumpur in 2014 highlighted two individuals who had been travelling with false European passports. One of the two, Delavar Seyed Mohammad Reza (29) travelled using the passport of Italian Luigi Maraldi (36; see Figure 1.1). Delayar passed multiple checkpoints before boarding the plane using a passport of a male with a different nationality and a 7-year age gap (Guardian, 2014). This situation illustrates that even if two individuals seem obviously different in appearance, they can be confused in a face-matching task by unfamiliar viewers. Misidentification from face-matching is still poorly understood.

Figure 1.1. Delavar Seyed Mohammad Reza (left) travelled using passport of Italian Luigi Maraldi (right). Image retrieved from https://bit.ly/2LI0oQf

Using the Glasgow Face Matching Test (GFMT) Burton, White and McNeill (2010) evidence that even with optimal viewing conditions, telling unfamiliar individuals apart or together is highly error-prone. In the GFMT viewers are presented with 168 pairs of face photographs (Figure 1.2). They simply have to decide whether the photos show the same person or two different people. They can take as long as they want to make a decision. Results show that overall error rates are at around 10%. This is strikingly high as photographs were taken roughly 15 minutes apart with two different cameras under good lighting conditions. Moreover, no deliberate effort was made for individuals to look different between photographs. This should illustrate that face matching for unfamiliar faces, even under perfect viewing conditions, is a difficult task.

Figure 1.2. Two example trials from Glasgow Face Matching Task. The left-hand trial shows two images of different individuals. The right-hand trial shows a pair of images from the same individual. Image retrieved from Burton, White and McNeill (2010).

Face matching is often more challenging than the above situation. Research shows that performance on an image-to-person matching task has even higher error rates. Kemp et al. (1997) looked at the performance of six experienced supermarket cashiers in their decision to accept or reject store credit cards for 44 shoppers. The credit cards either contained an image of the shopper or an image of another individual, presented at the point of check out. Performance was poor, with a 67% accuracy rate and more than half of fraudulent cards accepted as true. Note that in the experiment all photographs were taken in the 6 weeks prior to the experiment. British passports are valid for 10-year periods. This suggests that image to person matching is likely even harder in reality, with higher error rates expected.

One may expect that trained professionals (e.g. police officers and border control personnel) perform better than university students, but there is no evidence for this. A study by White et al. (2014) showed in a similar but easier design that error rates between passport-issuing officers and volunteer student participants were nearly identical. In the study, 34 students (17 females) acted as live ID bearers for a mock passport application. All passport application images were taken a few days prior to the experiment. Fraudulent pairs were created subjectively by swapping images for the most similar individuals amongst the 34 volunteers. This highly limits the likelihood of a convincing foil. Nonetheless, a 14% false acceptance rate amongst passport officers (n=27) was observed and a 6% rejection of valid photographs. Interestingly, they found no improvement in performance with increased years of experience. Moreover, a follow-up study compares performance on a photo-to-photo task for the same models between these officers and a student sample two years later and finds no significant differences between groups. Similarly, a study by Burton et al. (1999) showed that a group of police officers performed not better than untrained students at matching poor quality CCTV images to facial photos. These studies suggest that professional face matching experience does not improve face-matching ability.

Also note that the above studies likely underestimate the size of the reported effects in the real world, as none of them are able to incorporate the immense variety in appearance for one person in a ten-year period (see Figure 1.3) *or* the

deteriorating performance observed for continuous security checks that officers are exposed to day-in, day-out in the real-world (Alenezi, Bindemann, Fysh, & Johnston, 2015). More importantly, these studies only consider the situations where individuals make no deliberate attempt to look different or similar to their own appearance at a different time point *or* the individual they are intended to look like.

Figure 1.3. Examples of high variability in face images between passport photos (left) staff card photos (middle) and personal photos (right) for two authors of Jenkins et al. (2011): RJ (top) and AMB (bottom) bottom row). Consider similarity by rows and columns. Image retrieved from Jenkins et al. (2011).

1.2 Effects of deliberate disguise on face identification

Disguising the face consistently impairs recognition accuracy. A meta-analysis of factors contributing to eyewitness accuracy by Shapiro and Penrod (1968) confirmed that facial transformations such as disguises are one of the key factors to reduce correct identification and increase incorrect identification.

Some studies have focused on the effect of reading glasses on face identification. Terry (1993,1994) investigated this empirically by testing for the

effect of glasses between the encoding and test phase on identification accuracy in two separate experiments. In the first experiment, participants were shown faces of 12 individuals for which they made social judgments in an encoding phase. In a recall phase six of the twelve faces were kept the same, for three faces glasses were added and for three faces glasses were removed. They found that only the removal of glasses reduced recognition performance by approximately 40%. Study two replicated these findings with the addition and removal of beards, to allow for comparison between obstruction of different face regions and found that both situations reduced recognition accuracy by approximately 30%. Replication of these results were highly consistent (Leder et al., 2011; Patterson & Baddeley, 1977; Righi et al., 2012; Terry, 1993, 1994). Hockley et al. (1999) and Vokey and Hockley (2012) also found this pattern when sunglasses were used. Kramer and Ritchie (2016) even showed that glasses deteriorated performance accuracy by 8%, when one of the images contained a pair of glasses whilst the other did not.

Righi, Peissig and Tarr (2012) aimed to investigate what causes disguises to impair recognition from memory more directly. They considered the addition of glasses and wigs to human-like images. They confirmed that a change in hairstyle and removal of glasses had more of an effect on recognition than adding glasses did and compared these results to the recognition of the same inversed images. Their findings suggest that disguises are encoded inclusive to the person's identity. In other words, that even if disguises are recognisably *not* a physical part of the face, they are still encoded as a part of that identity.

Only one study by Dhamecha, Singh, Vatsa, and Kumar (2014) considered recognition of disguised faces from face matching. Models were asked to disguise themselves using a variety of props. Participants were shown these images in pairs of match or mismatch trials. Participants were simply asked to decide whether 2 images were shown of the same or different individuals, given unlimited response time. They found that same ethnicity and familiar viewers outperformed different ethnicity, unfamiliar viewers. They also found that occlusion of the eye regions had a disproportionally high disrupting effect on performance compared to other face regions.

These studies provide an interesting starting point for understanding face matching under disguised conditions, but only apply to situations where viewers can clearly see that the culprit is disguising part of their facial appearance. For the most part, this captures real-world use of disguises. Most disguises (hats, hoodies, sunglasses, balaclavas) are easily recognised as *not* being a facial feature indicative of identity. In turn, detecting the disguise may allow viewers to process disguise-free regions independently or recognise that the disguise needs to be removed before the culprit can be profiled or identified. This is important because recent media reports have indicated the use of a new type of disguise that is going undetected, resulting in profiling and identification errors in face matching.

1.3 Hyper-realistic silicone face masks

Media reports have highlighted the arrival of a new type of disguise to the criminal scene (see appendix 1.1 for complete list of known cases, and appendix 1.2 for a selection of case descriptions), referred to as hyper-realistic masks: over-head silicone face masks produced by a small number of Asian and North-American companies (see Appendix 1.3 for mask production detail and the types of masks produced). In one case, white male Conrad Zdrierak targeted 4 banks and a pharmacy wearing a black male realistic face mask (Figure 1.4). Eyewitnesses confirmed that the perpetrator was black and some even identified a black male individual from a security photograph (Gardner, 2010; Damani, 2014). Even more, a Korean refugee boarded a flight from Hong Kong to Vancouver using an elderly male mask with a real passport (Figure 1.5). The mask wearer passed several identity checks at Hong Kong airport, only to be discovered as he took the mask off mid-flight.

Figure 1.4. Security footage of Conrad Zdrierak wearing black male face mask produced by mask company SPFX (left) and at his hearing prior to his arrest (right). Image retrieved from https://bit.ly/2LBkyvb

Figure 1.5. Korean refugee (left), wearing realistic mask (middle and right) upon arrival at the Canadian border. Image retrieved from https://cnn.it/2ofNfmi

These cases suggest that hyper-realistic masks do not just effectively hide identity, through face covering, but also manage to go undetected. Their realism distinguishes realistic masks different from other types of whole-head facial disguise. Hyper-realistic masks seem to fool the eye. Any facial disguise can provide anonymity. Anonymity imbalances the power dynamic between the mask wearer and the observer, as only the disguised can evade the consequences of their actions. Moreover, the observer cannot read the disguised person's expressions and intentions. Seeing a facial disguise would alert and heighten the observer's vigilance in most cases ('What is this person intending to do?', 'Why would someone need to hide his or her identity in this situation?'). If a disguise is realistic enough to pass for a real face, it does not trip those defences. It leaves

the beholder greatly exposed, and the mask wearer at an even greater advantage.

1.4 A framework for effects of realistic mask

Realistic masks can be used to *evade* identity and to *impersonate* specific individuals. To outline the effects realistic masks have on recognition I will place realistic masks on an existing face recognition framework.

Face space theory (Valentine, 1991) is one of the more influential theories explaining how faces are recognised. Valentine (1991) proposed that there is one multidimensional space within which all newly perceived faces or familiar faces of a new appearance (e.g. a new haircut, or weight loss) are stored and grouped according to identity. Valentine argues that each identity is stored as a face average or cluster of images. The more exposure one has had to the variability in appearance (e.g. in lighting, viewing angle, change in appearance over time) the more populated the face space (or the more accurate the stored face average in the face space) will be for that individual. The theory explains why a less populated identity cluster may lead to false attributions to similar and equally unpopulated identity clusters (e.g. two Caucasian teenage males with short brown hair). An unpopulated identity cluster also explains failure to recognise an identity under new circumstances (change in lighting, clothing, and hairstyle after first encounter).

It is difficult to reason about high-dimensional space, and intuitions about high-dimensional space are often wrong (Burton & Vokey, 1998). Nevertheless, the basic spatial metaphor can be useful. For example, it provides a framework for thinking about *within*- and *between*-person variability of images. One could argue that within the multidimensional face space, there are face spaces specific for each identity. Each cluster could be thought of as a multidimensional ball, with an averaged center and dimensions representing a different means by which a face differs (e.g. lighting, viewing angle, image quality, age, weight, facial hair etc.; see Figure 1.6a). When a new image falls into this space, the image is attributed to

that identity. If it falls out of that space, the image is *not* attributed to that identity. Everything within the space captures the within-person face variability, whereas everything outside of that space captures the between-person variability.

If individuals actively attempt to disguise their identity, in case of criminal realistic mask use, they are attempting to escape their own face space as a means to avoid recognition. This is referred to as *deliberate* (opposed to incidental) *disguise.* In the literature, deliberate disguise is discussed in two forms: the attempt to evade one's own identity (*evasion*) or the attempt to pass for a specific other person (*impersonation*; Mendoza, 2015; figure 1.1a). Evasion is successful when the individual manages to avoid identification, for example using hoodies, sunglasses, wigs or facial props (figure 1.1b). A successful impersonation relies on passing for the impersonated individual. This is a more fragile process as it relies on carefully selected props (e.g. glasses, hair colour or skin tone) that match the target individual (figure 1.1b). Recent research also showed that looking subjectively similar to the target individual significantly advantaged the success of impersonation (Noyes, 2016; Noyes & Jenkins, under review). This suggests that there is only so much that can be done to impersonate with regular disguise props. These same limitations apply to evasion. For example, effective evasion of race, gender, age or weight group changes would be immensely time consuming and likely unrealistic.

If realistic masks go undetected, they allow much quicker transformation into a drastically different person (evasion) and allow the impersonation of an individual much more distinct from oneself with much more precision than with other disguise types (see Figure 1.6b), transforming the potential of facial disguise. These masks allow change in facial structure, skin colour, and gender *in addition* to regular disguise props with a much faster turn around (see Figure 1.6d for examples).

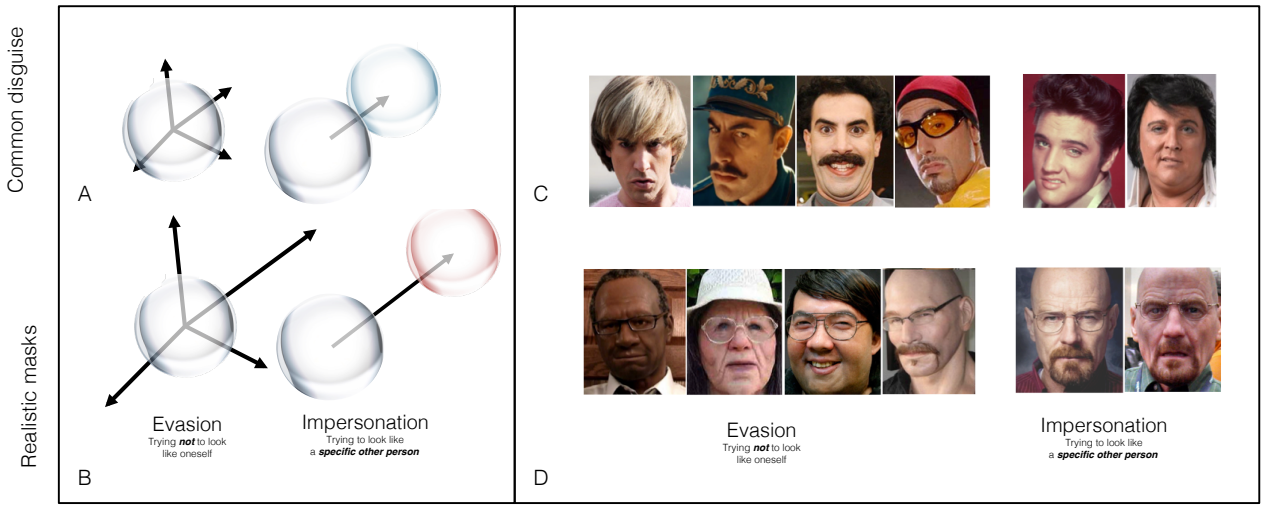

Figure 1.6. Model of evasion and impersonation disguise within face space for regular disguises (A), and realistic masks (B), accompanied by examples of these types of disguises for regular evasion and impersonation (C) and realistic mask type evasion and impersonation (D). Image adapted from Noyes (2016).

1.5 A market for realistic mask use

In terms of evasion, we expect that there is a market for realistic mask use in criminal settings. Taking US 2015 bank crime statistics alone, in 7% of cases, racial group and/or gender was impossible to determine due to use of overhead disguises (Department of Justice, 2016). We expect that in such crimes realistic masks/preserving observer ignorance could be highly valued. Appendix 1.1 list 33 cases covered by media outlets that used hyper-realistic face masks between 2009-2018 in criminal settings for evasion purposes. A steady increase in reports over time suggests that these masks are appearing increasingly on the criminal scene (figure 1.7). Amongst these criminal cases, 22 concerned (often multiple) bank robberies. One of the more extreme cases concerns the FBI-wanted bank robber, named the 'Geezer Bandit' (see Figure 1.8). This individual successfully robbed 16 banks in the last 5 years, likely wearing an old male realistic face mask, however as eyewitnesses and camera footage do not allow viewers to distinguish between the mask and a real face this remains unconfirmed. Indeed, for the first two years of his warrant, the FBI looked for a male in his 60-70's.

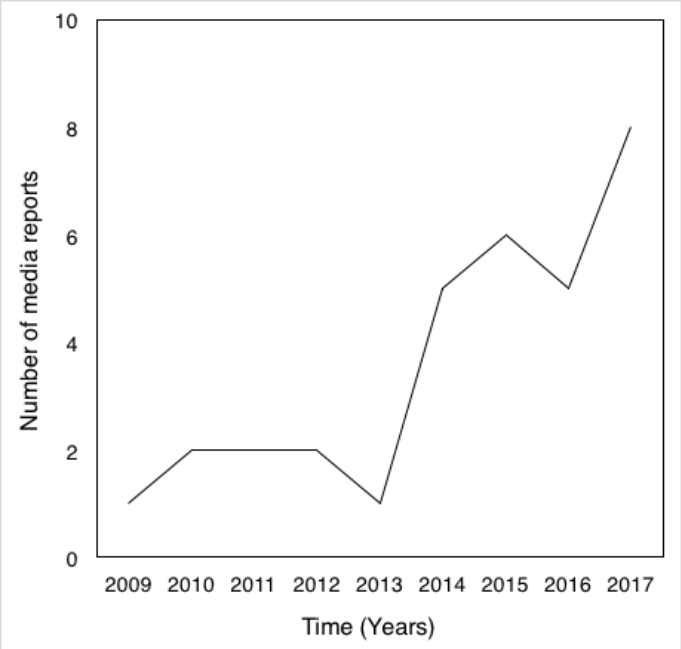

Figure 1.7. Number of independent crimes where hyper-realistic face masks were used, as reported on in the general media. See Appendix 1.1 for case details.

Figure 1.8. Example of bank robber wanted by the FBI, likely using an old male realistic face mask. For two years, they looked for a male in his 60-70's. Image retrieved from http://nbcnews.to/2dsqxUh

In terms of impersonation, we also expect that there is a market for hyper-realistic masks alongside fraudulent documentation. For example, in illegal border crossing situations. In the financial year 2013-14, of 5.7 million passport applications processed, 0.15% (over nine thousand) were detected to be fraudulent (HM Passport Office, 2014). We use the services provided by the Risk

and Airline Liaison Officer Network (RALON) overseas to estimate the market for fraudulent documents. RALON officers in 50 countries liaise with foreign governments, air, sea carriers and others identify and prevent around 8,000 people with no right to enter the UK from boarding flights or ferries to the UK alone (UK Border Agency, 2013).

Such individuals often use fraudulent documents to attempt entry. It is well known that the best fraudulent documents are real documents (often stolen or bought; Home office, 2016). Whereas forged documents can be matched to the photograph and demographics of the border crosser, real documents require the border crosser to match the person in the photograph and demographics of *another* individual – the person for which the document was rightfully produced. Until recently there was only so much an individual could do to match the appearance of the false document. For example, they could change their hairstyle and colour, change their facial hair or apply make up. Hyper-realistic mask change this situation. They allow individuals to change race, potentially gender and age category in a matter of seconds. With 3D printing, it is now even possible to produce a hyper-realistic face mask to match a photograph (e.g. hyperflesh.com). It is expected that they will increase the use of real fraudulent documents and perhaps even increase successful illegal entry. That is, if the realism of the masks pass for the person in the fraudulent document in a face-matching context *and* masks are realistic enough to pass for real faces. The prior is highly likely – based on the error rates in face matching and as masks being able to more accurately imitate an identity than any other facial props could. The latter needs to be investigated.

1.6 Understanding realistic mask detection

Media cases suggest that at least under certain circumstances, viewers are not able to distinguish realistic masks from real faces. This is striking as face perception is often thought to be a pinnacle of human vision, where we are able to detect smallest changes in emotional expression (e.g. Adolphs, 2002; Matsumoto

& Hwang, 2011; Niedenthal et al., 2001) and reliably judge gender, age, ethnicity and social traits (Sutherland et al., 2013; Sutherland et al., 2015). Moreover, detection of faces (as opposed to the detection of other stimuli) is considered to be innate (Sugita, 2009) or at least amongst the earliest abilities in human development (Jakobsen, Umstead & Simpson, 2016), and there seem to be regions of the brain specific to face processing (Kanwisher, Dermott & Chun, 1997). All this suggests that humans – as face experts - should not be fooled like the simple task of deciding whether a visual object is a real face or not. This is at odds with the reported mask crimes, where masks did in fact pass for real faces. To understand this apparent mismatch, I will discuss research on face detection, visual search and stimulus similarity and inattentional blindness.

Face detection

Faces provide important social cues, for example of person's emotional state or identity. To detect these cues, humans have to detect a face first. In turn, face detection is the required starting point of face processing. Face detection is much faster than the detection of other types of stimuli. Some studies suggest that they can actually be detected in under 100ms (Crouzet, Kirchner, & Thorpe, 2010; Crouzet & Thorpe, 2011). Moreover, under certain conditions faces have been shown to attract (Langton, Law, Burton, & Schweinberger, 2008; Theeuwes & Van der Stigchel, 2006) and retain attention (Bindemann et al., 2005) more so than other types of objects.

What information could guide face detection? Bindemann and Burton (2009) investigated the role of colour in the human advantage for detecting faces. They showed in two separate studies using face-absent and face-present trials in naturalistic scenes that removing colour or altering face colour impairs detection. They also used a half usual/half unusual colour face stimulus condition ruling out that the facial colours itself serves as a cue directing detection. Rather, they concluded that face detection must rely on combining diagnostic colour and face-shape information. The same group also provided evidence for the importance of the upper part of the face, and frontal and mid-profile poses compared to profile faces for detection (Burton & Bindemann, 2009; Bindemann & Lewis, 2013).

Realistic masks likely mimic facial colouration and key configural information essential to face detection. This would predict face detection mechanisms to respond to realistic masks and real faces equally and that distinguishing realistic masks from real faces is rather a process *following* the belief that a face has been detected. In turn, we expect spontaneous reports of realistic masks to be unlikely. Instead we expect that mask detection is more closely related to similar stimulus discrimination once attention is specifically guided to distinguishing realistic masks from real faces.

Visual search and similar stimulus discrimination

The *visual search* paradigm is amongst the most used paradigms to study visual attention, providing a highly useful lab-abstraction of the everyday task of searching a visual scene for a target object. Treisman and Gelade (1980) designed the visual search task, where they compared reaction times to stimulus detection in grids with a target and various numbers of distractor stimuli (e.g. 4, 9, 12, 18). They found highly consistent evidence for discriminating between *parallel* and *serial* searches. A parallel search requires telling apart a stimulus based on a single feature (e.g. colour, orientation, size). They found this to be an automatic *bottom-up* process, where number of additional stimuli has no effect on reaction time. Serial searches on the other hand, occur for items combining multiple features (e.g. orientation *and* size). The study found a linear increase in reaction time as distractor objects increased, meaning that in serial searches items are processed one at a time.

Wolfe (1994) provided a third, intermediate type of search coined as *guided* search. He proposed that guided searching allowed top-down reconfiguration of a search once informed of the exact features that are of interest. For example, if presented with an array of horizontal and vertical green and red lines, and asked to find a green horizontal line, a subject would only serially inspect green lines. This would increase searching speed in comparison to a serial search as proposed in the original visual search studies. Only certain features can serve to guide attention. These include colour, orientation, size and motion (Wolfe &

Horowitz, 2004). More interestingly Wolfe and Horowitz (2004) also examined faces. They suggest that although we are highly attentive to the detection of faces, face processing (e.g. discriminating familiar faces, characterising emotion, social traits and racial group) is serial. As realistic masks are likely not to distinguish from real faces on any specific features, this suggests that their discrimination from real faces would require effortful, top-down and serial searching, as opposed to regular masks (e.g. Halloween, masquerade etc.) that likely differ from real faces on a number of features allowing guided searching.

Inattentional blindness

Although the visual search paradigm explains why viewers may have to put more deliberate effort into realistic mask detection than might be expected, it does not necessarily explain how realistic masks could actually go undetected. One explanation could be *inattentional blindness*, where a lack of attention can cause perceptual oblivion for stimuli much more extreme than realistic masks. One famous attention experiment by Simons and Chabris (1999) showed in a video of a six-man ball game, that a person in gorilla suit walking through the scene went largely unnoticed. This is because viewers paid attention to the ball and players, and not the other visual information in the scene. A key difference in mask detection is that we expect viewers *are* paying attention to the face of the mask wearer.

A recent study by Drew, Vo and Wolfe (2013) illustrated why despite paying close attention to a stimulus, inattentional blindness can still occur. The study showed that a gorilla hidden in a lung-nodule detection task went undetected by 83% of radiographers upon close inspection of a radiograph. This is highly similar to the situation of mask detection. The unusual stimulus presented [masks/gorillas] did not differ greatly from the usual stimulus [faces/lung nodes] and was closely inspected by expert viewers [radiographers/humans; humans are considered to be face detection experts; e.g. Yang & Huang, 1994]. Nonetheless the target stimulus went undetected, because it the gorilla was unexpected. Realistic masks are highly unexpected in everyday settings as compared to real

faces. This may in turn cause the live viewers to encode the culprit's masked appearance as a real face.

In sum, evidence from these different areas of research suggest that realistic mask detection is likely to be a difficult task from memory and perceptually. Based on the likely role of attention we predict that spontaneous detection is unlikely and that even guided detection will be effortful.

1.7 Improving realistic mask detection

It is likely that there are some physical elements to the masks that allow the discrimination between masks and real faces. The question is what can be done to direct individuals towards these differences. In this section, I will discuss three directions that could potentially improve realistic mask detection: exploring the intersection between objects and faces and trained vs untrained expertise.

Realistic masks are objects

In recognition, a distinction is generally made between the recognition of faces and non-face objects (Gauthier, Behrmann & Tarr, 1999). Realistic masks hover on the boundary, as they are technically non-face objects made to imitate the appearance of a face. There is a large body of research that shows that faces are processed differently from objects, however that processing of face-like traits in objects lies somewhere in between (Churches, Baron-Cohen, & Ring, 2009; Hadjikhani et al., 2009; Ichikawa, Kanazawa & Yamaguchi; 2011; Robertson Jenkins & Burton, 2017). Not surprising is that viewers have no difficulty telling actual faces apart from objects with face-like features. Ichikawa et al. (2011) suggest that this is due the distinctive configurational and textural information surrounding the featural face-like cues. As textural and configurational information of realistic masks are not nearly as distinctive, separating realistic masks from real faces is not as simple.

Nonetheless, studies find that the human face processing mechanism does respond to the presented non-face stimuli (like based on feature and configuration information; Churches, Baron-Cohen, & Ring, 2009; Ichikawa, Kanazawa & Yamaguchi; 2011), and moreover even allow emotional expression processing. This results in the attribution of personality and emotion to objects such as cars (Windhager et al., 2010; Windhager et al., 2008): the more human-like the features, the more likeable the object (Ichikawa, Kanazawa, & Yamaguchi; 2011; see Figure 1.9).

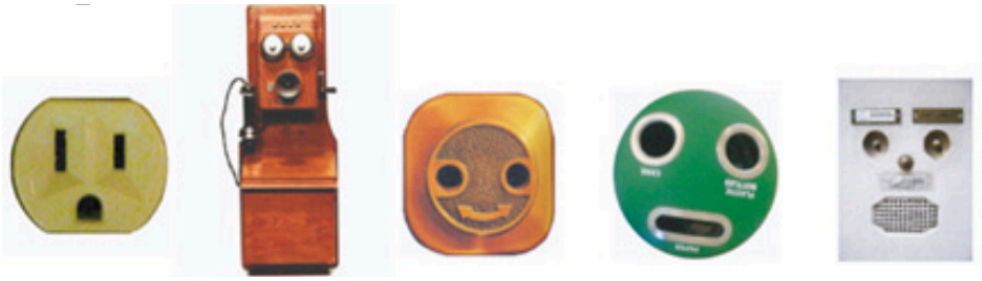

Figure 1.9. Top five objects rated to be highest in having a face as well as emotional expressions. Image taken from Ichakawa, Kanazwa & Yamaguchi (2011).

Research from artificial intelligence suggests a trajectory for the likeability of human facial features in objects, from the perspective of developing user-friendly artificial intelligence. There has long been concern for the advantages and disadvantages of human-like features of robotics. A theory by Mori (1970; see Mori, 2012 for official translation) proposed that objects or machines and even animals show an increasingly positive emotional response as they become increasingly human. This is thought to be due to a sense of familiarity between the object and the user/viewer. However, when the object or machine reaches a certain level of humanness, the users' emotional response steeply drops, as it is too similar for users to distinguish between persons and objects. This is referred to as the *uncanny valley* (Mori, 2012; official translation from Mori, 1970; see Figure 1.10). Realistic masks may be in the uncanny valley, where despite their face likeness they may be causing viewers unease (Seyama & Nagayama, 2007). In turn, it may be that this unease serves as a sign that allows viewers to distinguish between real faces and the masks.

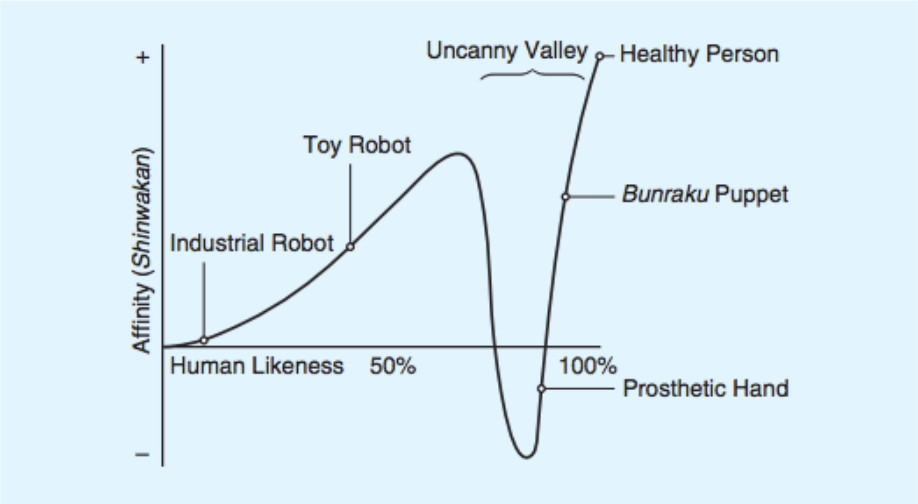

Figure 1.10. The uncanny valley depicts the relationship between human likeness of an object/entity (x-axis) and the perceivers' emotional response to this object/entity (y-axis). Bunraku refers to a traditional Japanese puppet used in musical theatre. Image retrieved from Mori et al. (2012).

It should be noted that empirical evidence for the uncanny valley is limited and that there may be alternative reasons to explain the same observation (see Pollick, 2009 and Złotowski et al., 2018 for a review). For example, it may be that realistic masks are merely *portraying* unpleasant real people, rather than that realistic masks generally have an effect, hence it is important that any unease effects are interpreted with caution.

Trained expertise

There is some research suggesting that you could train discrimination between real faces and realistic masks. A study by Miles-Worsley, Johnston & Simons (1988) compared memory performance of X-ray specialists for X-rays and faces and found in four different studies that memory for abnormal X-ray films increased with radiological experience. Moreover, they found that for the

experienced radiologists' memory for X-rays and faces was equivalent. In sum, there is evidence to suggest that there is an advantage for expertise, which could possibly translate to the special case of mask detection.

It should be noted that expertise only improves performance to a point. In real-world search tasks such as nodule detection, even high performing radiologists still average on a 20-30% miss rate (Drew, Vo, and Wolfe, 2013). In simulation cases of airport security checks miss rates are even higher (Clark et al., 2014). The above study by Drew, Vo, and Wolfe (2013) used the gorilla paradigm to evidence that despite increased performance and expertise in a certain task (e.g. direction nodules) experts are still bounded by the same attributes that guide detection as non-experts. Opposed to inventing new search strategies, they merely learned to use the same attributes more effectively. This means that, although we would expect improvement for mask detection for mask experts, they may not perform perfectly.

Untrained expertise

The above suggestions assume that indeed realistic masks can be distinguished from real faces through physical markers available to viewers for training. It is also possible that realistic masks are so face-like that masks need to be recognised rather than detected. Research from face recognition shows that there are some systematic person characteristics that predict face-matching accuracy that may translate to the situation of mask recognition.

One of the most consistent factors to mediate recognition accuracy is viewer familiarity. The above studies all concern *unfamiliar* viewers, however recognition drastically improves for *familiar* viewers. A study by Jenkins et al. (2011) explored the differences in performance for familiar and unfamiliar viewers using a card-sorting task and sets of celebrities from different countries. Dutch and British participants sorted 40 images of two celebrities by identity (see Figure 1.11 for the stimuli). Once they sorted familiar celebrities and once they sorted unfamiliar celebrities (Dutch and British respectively). Results showed that unfamiliar

viewers generally perceived 9 identities in the set, whereas familiar viewers accurately nearly always selected just two. In sum, face matching becomes a much easier task when you are familiar with the identity than when you are unfamiliar.

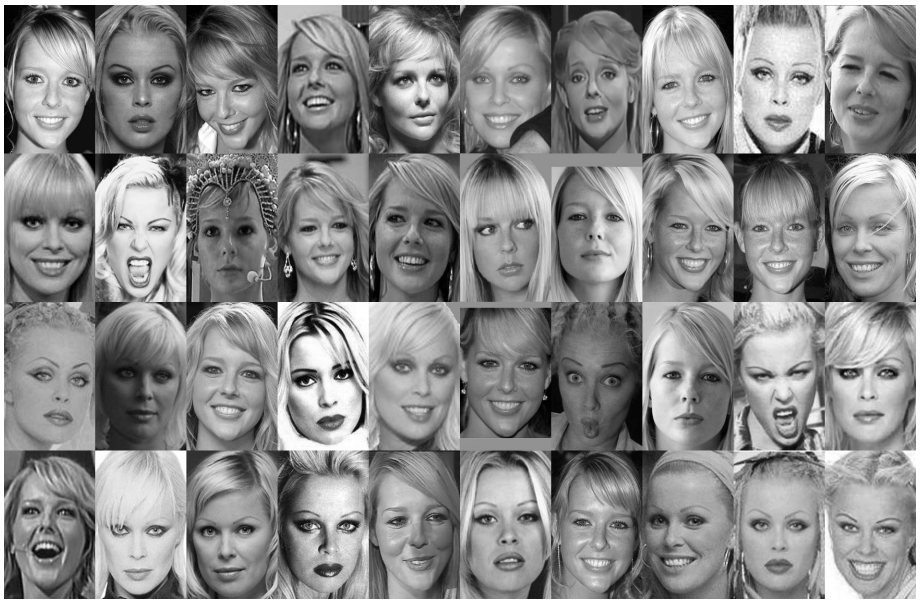

Figure 1.11. Card sorting task displaying two Dutch celebrities. Unfamiliar viewers struggle to sort these by identity, whilst it is easy for familiar viewers. Image retrieved from author of Jenkins et al. (2011).

Ideally, you would be able to train professionals how to become familiar with faces quickly. There is little evidence to support that face recognition ability can be trained, however we do see large individual differences in performance. For example, ability on a live-to-photo face-matching task by White et al. (2014) found performance to vary from 70% to 100% accuracy in only forty-nine individuals. Moreover, face recognition ability is thought to be on a spectrum, from individuals that completely lack the ability to recognise faces (*congenital prosopagnosics;* Behrmann & Avadin, 2005) to highly skilled face recognisers on a variety of tasks (*super-recognisers*; Russell, Duchaine & Nakayama, 2009). A recent study on face-matching classed super-recognisers (Robertson et al., 2016; n=4) working

for the London Metropolitan police force showed that they consistently performed above normal levels as measured in police trainees on the GFMT (n =194), and a student sample on the Model Face Matching Test (similar, but more difficult version of GFMT, n = 64) and Pixelated Lookalike Test (face matching for pixelated images of famous individuals and their lookalikes, n = 30). Although training individuals is unlikely to be a solution to the face recognition problem, it is possible to recruit individuals that are naturally good at face recognition. Similarly, if training of mask detection fails, it may be possible to turn to naturally able mask detectors.

Face familiarity is essentially a very narrow pocket of expertise for one specific identity, and only advantages identification performance of that identity. Beyond familiarity, there are other nested levels of expertise with a broader scope, but smaller effect size. For example, one of the most consistent effects in face perception is the *own-race bias*, where viewers are better at recognising faces from their own vs. other racial groups. A meta-analytic review by Meissner and Brigham (2001) showed that across 39 independent studies (N=5000) own-race recognition was more than twice as likely than other-race recognition. More recent work by Megreya, White, and Burton (2011) found that this even occured in a face-matching context. Their study used 240 target present and target absent trials for an own- and other-race line-up task. They found lower hit rates, misidentification and false alarms for own- vs. other-race face matching. Perceptual other-race differences in accuracy levels have also been reported using a card sorting task (Yan et al., 2016) and a photograph-to-passport matching task (Meissner, Susa & Rosa, 2013). They attributed the observed effect to the limitations of perceptual encoding of unfamiliar faces.

Similarly, meta-analyses confirmed an own-age bias (Anastasi & Rhodes, 2005; Rhodes & Anastasi, 2012; Wiese, Komes & Schweinberger, 2013; Neil et al., 2016) and own-gender bias in females (Herlitz & Loven, 2013). More generally, there is also discussion of a *cross-category effect* in face recognition (e.g. Bernstein, Young & Hugenberg, 2007), where they even found reduced face recognition for members from a different university as manipulation in the experiment. To the best of my knowledge, there is no evidence from face-

matching studies on the above effects, but based on face-matching effects for the other-race manipulation we would expect similar results.

These effects all assume that *my* race, gender or racial group are a proxy for either my exposure or affiliation with that group, leading to identification expertise. Some interesting studies have shown that this expertise can cultivate with exposure. For example, where one grows up can determine the other-race advantage (e.g. Asian person in the Caucasian environment; Tham, Bremner & Hay, 2017), and where one works can determine the other-age advantage (e.g. nursing home workers; Wiese, Wolf, Steffens & Schweinsberger, 2013) or even cross-species (e.g. a farmer with expertise of sheep faces; McNeil & Warrington, 1993; McKone, Kanwisher & Duchaine, 2007) or object advantages (e.g. a car dealer with car expertise; Sergent & Signoret, 1992; Gauthier et al., 2000). This suggests that variance in expertise of realistic masks could determine discrimination performance too. Hence, even if we are not clear on the category which provides advantage yet, we expect that certain individuals may be better than others at this categorisation task.

In addition, the demographic cues carried by hyper-realistic face masks could align or misalign with other expertise clusters (see Figure 1.12). For example, someone with Asian face expertise may be more likely to notice that an Asian face mask is not realistic than a person with western face expertise. Hence, we also expect that race, gender and age group advantages to identification could affect realistic mask detection performance.

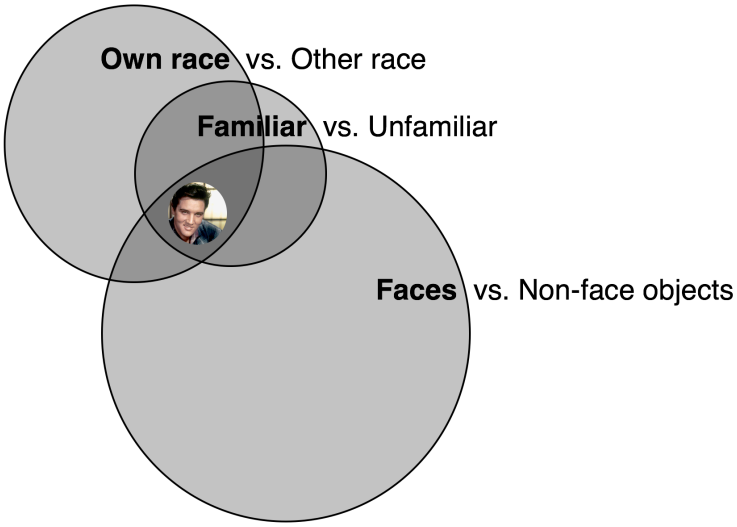

Figure 1.12. Illustration of clusters of expertise having overlapping benefit to recognition accuracy.

1.8 Overview of current work

In this thesis, I will investigate the effects of realistic masks in two areas of face perception using behavioural experiments. First, I will investigate under which circumstances realistic masks are mistaken for real faces (chapter 2-4). Second, I will investigate which information is preserved of the wearer beneath the mask (chapter 5-7). In both parts I examine these effects in live viewing and photographic conditions.

Chapter 2 considers mask detection in a situation similar to border control, where a line up of faces is inspected one at a time, with one mask amongst the faces. Experiment 1 examines participant awareness of the presence of the mask using graded detection questions, with a memory-based and perceptual detection component. Experiment 2 replicates this study in Japan to illustrate that the same effects are observed across largely different samples. To confirm that this is not just a lab-based artefact, Experiment 3 investigates an adaptation of the same design in an outdoor, real world situation both in Japan and in the UK, comparing high-realism masks to a low-realism and no-mask condition.

Chapter 3 uses a more stringent test of realism. This chapter uses a computer-based two-alternative forced-choice (2AFC) paradigm by means of Turing test for synthetic faces, where participants have to decide which one of two images on the screen is the mask. Participants are shown pairings of a real face with a low or a high realism mask. Participants are asked to respond as quickly as possible. The task measures differences in reaction time and response accuracy. We expect that low realism masks are slower to detect than high realism masks. To follow up on the other-race effect performance by British and Japanese participants are compared for Western and Asian image sets. To ensure performance would not be at ceiling Experiment 4, allows just 500ms exposure time to the view the images. Experiment 5 repeats the same task, but with unlimited image exposure time.

Chapter 4 uses the Western and Asian image set from Chapter 3 to investigate the individual differences in detection rates of participants and mask images in British participants. Here, we use an adapted paradigm, where participants decide for each image whether it contains a mask or not with unlimited time to decide. Experiment 6 compares performance accuracy for discriminating Low and High realism face masks from real faces. Experiment 7 looks only at discriminating High realism masks from real faces, then followed up with an Image Analysis. We use an image analysis to separate high and low performing individuals to determine which cues high performing individuals use to separate high realism masks from real faces.

In Chapter 5 (Experiment 8) we switch to our second question and considers which visual information is preserved through the realistic mask. We start by investigating whether demographic cues of the wearer are visible through the mask. We compare whether viewers can detect the gender, age and race of the mask wearer, comparing a head only to a head and body condition. We also use this data to test for a fundamental attribution error in the attribution of the mask to the wearer.

Chapter 6 (Experiment 9) follows by investigating recognition of two confederates wearing three different masks, using a simple computer-based

2AFC line up task. As a possible route to improving recognition performance in the real world, we compare performance for participants who are familiar with the confederates to those who are not. For the same reason, we consider individual differences in performance on this task. In addition, we use data as a test bed for the attribution error of the mask to the wearer.

As a final experimental chapter, Chapter 7 follows up on the attribution error of the mask to the wearer and aims to isolate the effect. Using the same image set as Chapter 6, Experiment 10 considers the preservation of character traits estimates in these variable face images comparing masked and unmasked face images. Experiment 11 uses the same design but looks at the effects of the mask on personality estimates of the wearer.

To close, Chapter 8 discusses where realistic masks may be of interest, in light of the research findings. In relation to security, I discuss the lack of mask detection, and the lack of preserved facial information underneath the mask from a theoretical and an applied perspective. Finally, I discuss the practical and future uses of realistic masks in wider research contexts.

Chapter 2.

Detecting hyper-realistic face masks

2.1 Summary

We often identify people using face images. This is true in occupational settings such as passport control, and also in everyday social environments. Mapping between images and identities assumes that facial appearance is stable within certain bounds. For example, a person's apparent age, gender, and ethnicity change slowly if at all. It also assumes that deliberate changes beyond these bounds (i.e. disguises) would be easy to spot. Hyper-realistic face masks overturn these assumptions by allowing the wearer to look like an entirely different person. If unnoticed, these masks break the link between facial appearance and personal identity, with clear implications for applied face recognition. To date however, no one has assessed the realism of these masks, or specified conditions under which they may be accepted as real faces. Here we examined incidental detection of unexpected but attended hyper-realistic masks in both photographic and live presentations. Experiment 1 (UK; N = 60) revealed no evidence for overt detection of hyper-realistic masks among real face photos, and little evidence of covert detection. Experiment 2 (Japan; N = 60) extended these findings to different masks, mask-wearers, and participant pools. In Experiment 3 (UK and Japan; N = 407), passers-by failed to notice that a live confederate was wearing a hyper-realistic mask, and showed limited evidence of covert detection, even at close viewing distance (5 m vs 20 m). Across all of these studies, viewers accepted hyper-realistic masks as real faces. Specific countermeasures will be required if detection rates are to be improved.

2.2 Introduction

Face recognition is a common means of identifying people, and an important

component of security and crime prevention internationally. For example, passport issuance (White et al., 2014) and passport control (McCaffery & Burton, 2016) both involve facial image comparison. Conviction of criminal suspects can sometimes hinge on eyewitness testimony (Wells & Olson, 2003; Bruce, 1988; https://www.innocenceproject.org) or CCTV footage (Burton et al., 1999; Davis & Valentine, 2009). In many countries, photo-ID is required for the purchase of age-restricted goods (Gosselt et al., 2007; Vestlund et al., 2009). Because face identification carries such weight in these situations, it is also a major focus for identity fraud and deception (Robertson, Kramer, & Burton, 2017). In particular, individuals may wish to impersonate someone else or to avoid being recognised themselves (Dhamecha et al., 2014).

One way to conceal identity is simply to cover the face, for example, using fabric or a mask (Fecher & Watt, 2013). Covering the face is generally effective in obscuring identity (Burton et al., 1999), but it is also visually and socially salient, and likely to arouse the suspicion of onlookers (Zajonc, 1968). Over the past decade, this limitation has been challenged by the emergence of hyper-realistic masks (Figure 2.1). These hand-painted silicone masks were originally developed in the special effects industry as an alternative to multi-hour make-up sessions. The flexibility and strength of silicone confer several advantages in this situation. Unlike traditional masks that cover the face only, a silicone mask may cover the whole head and neck, so that it extends below the collar without any joins. This seamless construction creates the impression that the visible face is part of a continuous body surface rather than being a separate overlay (Anderson, Singh, & Fleming, 2002). Realism is further enhanced by transmission of non-rigid movement (e.g. rotation of the head relative to the body; opening and closing of the mouth; gross changes in facial expression) from the surface of the face to the surface of the mask. Importantly, the wearer's real eyes, nostrils, and mouth cavity are all visible through the mask via close-fitting holes that match the topology of the face beneath. Several manufacturers offer hand-punched human hair and stubble as optional extras.

Figure 2.1. Hyper-realistic silicone masks. Images show (from left to right) Young Male Mask (YMM), followed by Young Male Mask (YMM), Old Female Mask (OFM), and Old Male Mask (OMM) worn by Rob Jenkins.

These advances in mask fabrication raise the question of how realistic a mask can be. For present purposes, we adopt a pragmatic definition of realism: *a mask is realistic if it is perceived as a real face*. This criterion has the advantage of being testable, and can be applied across different viewers and viewing conditions. It also gets to the heart of the practical problem. If covering one's face arouses suspicion, the ability to cover one's face without arousing suspicion would seem to favour the deceiver.

There are reasons to doubt that this level of realism can be achieved in practice. For one, the visual system is highly attuned to face stimuli, including subtleties of skin tone (Fink, Grammer, & Matts, 2006; Frost, 1988; Bindemann & Burton, 2009) and face shape (Ekman, 2003; Oosterhof & Todorov, 2008). Thus, it seems plausible that even minor departures from authentic appearance at the physical level could loom large at the perceptual level. Paradoxically, some demands of the perceptual system may become harder to satisfy as authenticity increases. The *uncanny valley* refers to the phenomenon whereby human response to humanoid artifacts (e.g. robots, dolls, puppets), shifts from empathy to revulsion as the humanoid approaches, but fails to attain, lifelike appearance (Mori, 1970; see Mori, MacDorman & Kageki, 2012, for an English language

translation). Given humans' particular sensitivity to face stimuli, one might expect the uncanny valley to pose a particular challenge for masks (Seyama & Nagayama, 2007). A sense of eeriness could undermine an otherwise compelling overall impression of realism.

Theoretical concerns aside, the important question is whether these masks actually fool anyone. There is now a good deal of anecdotal evidence that hyper-realistic masks can pass for real faces in everyday life. In one incident, a white bank robber used a silicone mask to disguise himself as a black man for a string of robberies in the USA. Six out of seven bank tellers wrongly identified a black man as the culprit in a photo line-up. Only when the robber's girlfriend intervened was the black suspect released from jail (Bernstein, 2010). In another case, a young Asian man disguised himself as an elderly white man using a silicone mask, and boarded a flight from Hong Kong to Canada (Zamost, 2010). The deception was only detected when the passenger removed the mask midflight, and a fellow traveller brought the change in appearance to the attention of the crew. These examples imply that realistic masks can be mistaken for real faces, even when the viewer's attention is focused on facial appearance (as is the case in police line-ups and passport checks). Surprisingly however, there has been no experimental research into hyper-realistic masks and the conditions under which they can be detected.

Here, we address these questions in three experiments. We examine mask detection from static photographs (Experiment 1 and 2) and in live viewing (Experiment 3), to assess performance in these two modes of face identification. We had the opportunity to collect data from both British and Japanese participants, allowing us to compare performance for own-race and other-race faces. A large body of research on the other-race effect has shown that identification performance is more reliable for own-race faces than for other-race faces (Meissner & Brigham, 2001). Our question here is whether a similar bias operates when distinguishing hyper-realistic masks from real faces.

2.3 Experiment 1:

Detection from photographs with British participants

In Experiment 1 we secretly embedded photos of hyper-realistic masks among photos of real faces. Participants worked through these photos sequentially, rating the person in each photo on a series of social dimensions. This task ensured that participants processed the images, but did not draw attention to the distinction between real faces and masks. We then asked a series of graded questions to determine whether or not they had noticed any masks among the faces. After explaining the manipulation, we showed the stimuli again, and asked participants to pick out any photos that contained masks. We predicted that when participants were not expecting to see masks (i.e. during the rating phase), realistic masks might not be detected, resulting in few spontaneous reports of masks in post-test questioning. However, when participants are expecting to see masks (i.e. after the manipulation has been explained), they should be able to distinguish realistic masks from real faces, merely by inspecting the photographs.

Method

Ethics statement. Ethical approval was granted by the departmental ethics committee at the University of York.

Participants. Sixty undergraduate and postgraduate members of the volunteer panel at the University of York (10 males; mean age = 21, age range 18–39 years) took part in exchange for a small payment or course credit.

Stimuli and Design. We used three different models of mask from Realflesh Masks, Quebec, Canada: *The Pensioner* (Old Male Mask; OMM), *The Fighter* (Young Male Mask; YMM), and *The Grandma* (Old Female Mask; OFM). The company offers a range of hair options for its masks. We opted for punched human hair eyebrows on all three, and a full head of hair on *The Grandma*.

To generate mask images, we took multiple photographs of the same volunteer model wearing each of the three masks. We took photos indoors and outdoors under different viewing conditions to approximate the range of variability seen in natural face images (Jenkins et al., 2011). For each mask, we selected two different photos that depicted the mask in frontal view with no occlusions (6 mask images in total).

To generate real face images, we entered the terms ‘young male’, ‘old male’, ‘young female’, and ‘old female’ into Google Image search. For each of these four face types, we selected the first five colour photos of unfamiliar Caucasian faces that (i) exceeded 200 pixels in height, (ii) showed the face in roughly frontal aspect, and (iii) were free from occlusions (20 real face images in total). All photos (masks and real faces) were cropped to show the head region only and resized to 540 pixels high x 385 pixels wide for presentation.

Starting with the 20 real face photos, we created different stimulus sets by substituting one mask for one real face of the same type (young male, old male, or old female). This resulted in six variant image sets, each consisting of 1 mask photo embedded in 19 real face photos. Ten participants saw each variant.

Procedure. Participants viewed 20 photographs (19 real faces + 1 hyper-realistic mask) one at a time, in a random order. To encourage deep processing of facial appearance, we asked participants to estimate the age of the person in each photo, and to rate the person for *Trustworthiness*, *Dominance*, and *Attractiveness*, using a 7-point Likert scale (for ratings, see appendix 2.3). There was no time limit for this task, and photos remained on screen until all responses were made. This rating task was followed by a series of graded questions to assess detection of the mask. Question 1, ‘What did you think of the faces you saw?’, was deliberately open, and was intended to capture spontaneous, overt detection of the mask. Question 2, ‘Did you notice anything unusual about any of the faces?’, encouraged participants to report any suspicions that they may have had during the task (i.e. more covert detection). Both of these questions invited typed responses. Question 3 led to a two-alternative forced choice (2AFC), which was intended to provide a more sensitive measure: ‘In this experiment, half of the

participants are in the *Mask* group (where at least one of the photos contains a mask). The other half are in the *No mask* group (where none of the photos contained a mask). Which group do you think you were in (*Mask* vs *No mask*)?'. After responding, participants were informed that they were in the *Mask* group. They were then presented with all 20 of the photos they had rated (19 real faces and 1 mask) in a randomly-ordered 5 x 4 array, and asked to indicate any photo that contains a mask (Question 4; see Figure 2.2). At the end of the experiment, participants were debriefed and asked to indicate whether or not they had prior knowledge of realistic silicone masks before the start of the experiment.

Figure 2.2. Example array challenge from Experiment 1. Participants were asked to indicate any photos that show a mask. The array always contained 19 real faces photos and 1 mask photo. In this example, image 9 shows Rob Jenkins in the old male mask (OMM).

Results

Mask Detection. We first tested for overt detection of the masks by analysing the content of typed responses to Question 1 ('What did you think of the faces you saw?') and Question 2 ('Did you notice anything unusual about any of the faces?'). To avoid imposing our own interpretations on these responses, we simply coded for the presence (1) or absence (0) of the word 'mask' in the text. As it turned out, none of the sixty participants included the word 'mask' in either response. That is, there were no cases of overt detection (see appendix 2.1 for raw data). For the 2AFC item (Question 3), only 21.7% of participants guessed that they were in the *Mask* group, significantly lower than the chance level of 50% [$t(59) = 5.28$, $p < .001$, $d = -17$]. Finally, in the array challenge (Question 4), 70% of participants correctly picked out the mask. However, participants also picked out an average of 2.5 (range: 0-10) real faces (see Figure 2.3, left). In fact, all but one of the real faces (YM1) was reported as a mask at least once. Chi-square analysis revealed no significant differences in detection performance across mask types [2AFC: X^2 (3, N=60) = .79, p = .680, Cramer's v = .13; Array challenge: X^2 (3, N=60) = 1.43, p = .490, v = .12].

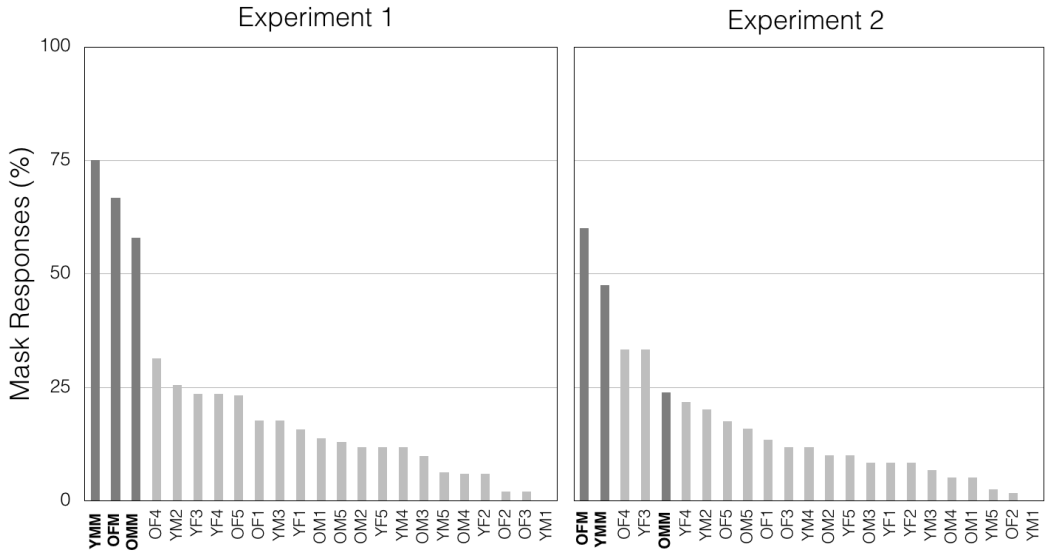

Figure 2.3. Responses to the array challenge in Experiment 1 (left) and Experiment 2 (right). Bars show, for each image in the array, the percentage of participants who reported it as a mask, and are ordered by frequency. Dark bars represent mask images (YMM, OFM, OMM). Light bars represent real face images (YM, Young Male; OM, Old Male; YF, Young Female; OF, Old Female).

Mask Knowledge. 38 of the 60 participants declared prior knowledge of hyper-realistic masks. Chi-square analyses revealed no significant difference in 2AFC performance between *Knowledge* (N = 38; 21.1%) and *No Knowledge* (N = 22; 22.7%) subgroups [X^2 (2, N = 60) = .02, p = .807, *v* = .03]. However, prior knowledge conferred a significant advantage in the array challenge [*Knowledge*: 78.9%; *No Knowledge*: 54.5%; X^2 (3, N = 60) = 3.95, p = .046, *v* = .28).

Discussion

We find it quite striking that not a single participant volunteered that they had seen a mask. Even under 2AFC questioning, only 22% thought that a mask might have been presented. These findings suggest that, at least in the context of viewing photos, participants need to both (i) be informed that a mask may be

present, and (ii) have the images available for inspection, if they are to distinguish hyper-realistic masks from real faces. Even when these conditions were met (in the array challenge), 30% of participants missed the mask, and 78% picked out at least one real face. The message from this experiment is that detecting hyper-realistic masks is hard, even when the test conditions are highly favourable. We next consider a situation in which the test conditions may be less favourable: viewing other-race faces.

2.4 Experiment 2:

Detection from photographs with Japanese participants

Viewers are generally poor at identifying other-race faces compared with own-race faces. This is true for tasks involving recognition memory (Meissner & Brigham, 2001) and also for tasks involving perceptual comparison of face photographs (e.g. Megreya, White & Burton, 2011). The perceptual explanation of this own-race bias is that the ability to distinguish individuals is refined by experience: viewers become attuned to the variability that surrounds them, and remain relatively insensitive to variability outside of this range (O'Toole et al., 1994). This differential sensitivity supports finer perceptual discriminations for own-race faces than for other-race faces. In the case of hyper-realistic masks, distinguishing a mask from a real face also requires fine perceptual discriminations, perhaps akin to distinguishing one person from another. If so, the task of hyper-realistic mask detection may also be susceptible to own-race bias. In Experiment 2, we had the opportunity to replicate Experiment 1 in Japan, using the same stimuli and procedure as before, but now with Japanese participants. Given that all of our stimuli showed Western (Caucasian) faces and masks, our main interest was whether hyper-realistic masks would be more readily accepted by Japanese participants compared with the UK participants in Experiment 1.

Method

Ethics statement. Ethical approval was obtained from the Kokoro Research Center ethics committee at Kyoto University.

Participants. Sixty undergraduate and postgraduate members of the volunteer panel at Kyoto University (36 males; mean age = 22, age range 19–36 years) took part in exchange for a small payment.

Stimuli and procedure. The stimuli, design, and procedure were exactly as for Experiment 1, except that the task instructions were now translated into Japanese. Two experienced translators provided translations independently. The best translation was selected and verified for functional similarity with the English version by a third, bilingual English-Japanese speaker. For social charactertistic ratings, see appendix 2.3.

Results

Mask Detection. Consistent with Experiment 1, none of the sixty participants mentioned the Japanese word for 'mask' in response to Question 1 or Question 2 (see appendix 2.2 for raw data). For the 2AFC item (Question 3), 33.3% of participants guessed that they were in the *Mask* group, significantly below chance [$t(59) = 2.72$, $p = .009$, $d = -10$]. Finally, in the array challenge (Question 4), just 45% of participants correctly picked out the mask. Participants picked out an average of 2.3 (range 0-11) real faces (see Figure 2.3, right). As in Experiment 1, all but one of the real faces (YM1) was identified as a mask at least once. Again, there were no significant differences in detection performance across mask types [2AFC: X^2 (3, N = 60) = 2.17, p = .338, v = .103; Array challenge: X^2 (3, N = 60) = 3.75, $p = .074$, $v = .27$].

Mask Knowledge. Only 3 participants in the Japanese sample reported prior, knowledge of hyper-realistic masks. Of the 57 participants who had no prior knowledge of masks, 32.2% guessed that they were in the mask group (Question

3), and 47% picked the mask in the array challenge Question 4). Of the three participants who reported prior knowledge, one picked the mask group (Question 3) and two picked the mask out of the array correctly (Question 4).

Comparison of UK and Japan samples. None of the 120 participants (60 UK, 60 Japan) mentioned masks spontaneously (Question 1) or when prompted (Question 2). For the 2AFC item (Question 3), the proportion of 'mask' responses was higher for Japanese participants (33.3%) than for UK participants (21.7%), though this difference was not significant [X^2 (1, N = 120) = 2.05, p = .152, v = .14]. However, in the array challenge (Question 4), Japanese participants picked out the actual mask significantly less often (46.7%) than the UK participants (70%) [X^2 (1, N = 120) = 6.72, p = .010, *v* = .24].

Discussion

Overall, the results are very similar to those seen in Experiment 1. Like the UK viewers, Japanese viewers did not spontaneously report seeing a mask despite two opportunities to do so (Questions 1 & 2). A low proportion of viewers believed that they were in the *Mask* condition (Question 3), and a low proportion picked the mask out from an array of real face photos (Question 4). Accuracy on this array challenge was reliably lower in Experiment 2 (Japanese viewers) than in Experiment 1 (UK viewers), possibly reflecting an other-race effect, although there are many other possible explanations for this difference.

To follow up these findings, we expanded to a fully crossed design in which both British and Japanese participants viewed both Asian and Western faces. More importantly, we also progressed from viewing photographs on a computer screen to viewing live faces outdoors.

2.5 Experiment 3:

Live detection with British and Japanese participants

Mask detection rates in the preceding experiments were consistently low. There are several reasons to be cautious in interpreting this finding. One reason is that all of the stimuli in Experiments 1 and 2 were photographic images. Single, static photos present much less information than dynamic, live faces (Jenkins & Burton, 2011). It is possible that under live viewing conditions, detection rates could be much higher. On the other hand, all of the participants knew that they were taking part in a psychology experiment, and this setting may have made them especially vigilant. On that basis, it is possible that under live viewing conditions, detection rates could be even lower.

To avoid these limitations, we adapted the mask detection measures from Experiments 1 and 2 to a very different situation. Instead of recruiting participants to a laboratory-based experiment, we recruited passers-by in an outdoor area of the University. And instead of asking these volunteers to rate onscreen photographs, we asked them about a live confederate. In one condition, the confederate wore a hyper-realistic mask. As in the previous experiments, our main interest was whether viewers noticed the mask or accepted it as a real face (*High-realism mask* condition). To establish a false alarm rate, we included a condition in which the confederate did not wear a mask (*Real face* condition). To establish the rate of miss errors (due to inattention, misunderstanding task instructions, etc.), we also included a condition in which the confederate wore a highly salient party mask (*Low-realism mask* condition). This allowed us to assess the detection rate for hyper-realistic masks relative to these base-rates.

To test for other-race effects in this task, we recruited participants in both Japan and the UK to view both Asian and Western masks. An other-race effect should result in poorer detection of hyper-realistic masks for *Other-race* trials (Japanese participants viewing Western masks and British participants viewing Asian masks), compared with *Own-race* trials (Japanese participants viewing Asian masks and British participants viewing Western masks). Finally, we

examined effects of viewing distance by comparing performance in *Near* (5 m) and *Far* (20 m) conditions. We expected improved detection of high-realism masks at the closer viewing distance, where more detail is visible.

Method

Ethics statement. Ethical approval was granted by the Kokoro Research Center ethics committee at Kyoto University and the departmental ethics committee at the University of York.

Participants. Four hundred and seven volunteers participated in the study. All participants were undergraduate or postgraduate students at the University of York, UK (N = 199; 107 males; mean age = 20, age range 18–44 years) or Kyoto University, Japan (N = 208; 134 males; mean age = 21 years, age range 18-38 years).

Stimuli and Design. Four male confederates were briefed on the aims of the study. For the *High-realism mask* condition, we used four masks in total. Three of these were produced by Realflesh Masks, Quebec, Canada: *The Pensioner* (Western Old Male Mask), *The Fighter* (Western Young Male Mask), and *The Asian* (Asian Old Male Mask). The remaining mask was the *Jae* model (Asian Young Male Mask), by Composite Effects (CFX), Los Angeles, United States. We ordered punched human hair eyebrows on all four masks, a goatee beard and horseshoe hair on *The Asian*, and a full head of hair on the *Jae*. To avoid overcomplicating the design, confederates wore own-race masks only. For the *Low-realism mask* condition, we used two visually salient masks that covered the face only, rather than the whole head. These were a plain green Halloween-style mask (see Figure 2.4) and a black butterfly-shaped masquerade mask. Note that the distinction between *Own-race* and *Other-race* applies to the *High-realism mask* condition and the *Real face* condition, but does not apply to the *Low-realism mask* condition.

Combining each of these presentations with *Near* and *Far* viewing

distances resulted in 10 conditions in total. Each participant saw one condition only (between-subjects design). As in the preceding experiments, each participant responded to an open question, a prompted question, and a 2AFC question.

Procedure. Testing took place in campus courtyards at the University of York and Kyoto University between 11:00 and 14:00 on different dry weather days between November 2014 and October 2016. For the duration of the testing session, the confederate remained seated at a bench in a university courtyard with reliable foot traffic. The two experimenters recruited viewers at approximately 5 m (*Near* condition) and 20 m (*Far* condition, see Figure 2.4, right panel) from the confederate by pointing out the confederate to individual passers-by, and asking whether they would mind answering a few questions about him. To encourage deep processing of facial appearance, the participant was first asked to rate the confederate for *Trustworthiness*, *Dominance*, and *Attractiveness*, using a 7-point Likert scale. After responding, the participant was asked to turn to the experimenter so that the confederate was no longer in view. The experimenter then asked graded mask detection questions that were adapted from the preceding experiments: 'What did you think of that person?' (Open question), 'Did you notice anything unusual about the person?' (Prompted question), and 'There are two conditions in this experiment, one where the person is wearing a mask and one where he is not wearing a mask. Which condition are you in?' (2AFC question). Data were recorded by the experimenters using prepared response sheets. The entire procedure lasted approximately two minutes for each participant.

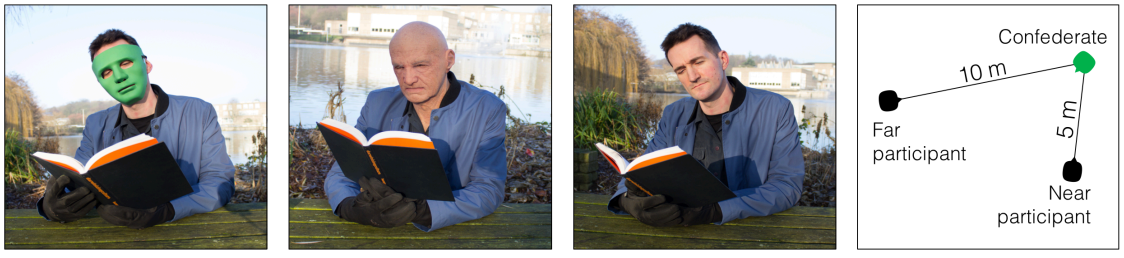

Figure 2.4. Illustration showing (from left to right) Rob Jenkins in the *Low-realism mask*, *High-realism mask*, and *Real face* conditions of Experiment 3, and the spatial arrangement of confederate and participants.

Results

Descriptives. Table 2.1 summarises the distribution of participants across conditions.

Test Location	Viewing distance	Low-realism mask	High-realism mask		Real face	
			Own-race	Other-race	Own-race	Other-dace
Japan	Near (5 m)	24	20	20	20	20
	Far (20 m)	23	20	20	20	21
UK	Near (5 m)	20	20	22	20	20
	Far (20 m)	18	20	18	20	21

Table 2.1. Number of participants tested in each of the 10 different conditions in Experiment 3, shown separately for testing in UK and Japan. Note that the Own-race / Other-race distinction does not apply to the Low-realism mask condition.

Mask Detection. To ensure consistency across experiments, we coded responses to Questions 1 and 2 according to the presence or absence of the word 'mask' in the response. As expected, detection rates in the *Low-realism mask* group were high overall (see Figure 2.5), indicating good engagement with the task. For the Open question (Question 1), 49.2% of *Near* participants and 42.1% of *Far* participants included the word 'mask' in their responses. For the Prompted question (Question 2), these proportions rose to 67.2% (*Near*) and 82.3% (*Far*). Finally, for the 2AFC item (Question 3), almost all participants guessed that they were in the *Mask* group (*Near* 95.0%, *Far* 97.9%). In sum, *Low-realism masks* were rarely missed.

Complementing this pattern, performance in the *Real face* group shows a low false alarm rate. None of the participants in this group used the word 'mask' in their responses to either the Open question (Question 1) or the Prompted question (Question 2). For the 2AFC item (Question 3), participants in the *Own-race* condition almost never guessed that they were in the 'mask' group (*Near*

2.5%, *Far* 2.5%). Interestingly, participants in the *Other-race* group occasionally picked the 'mask' group, especially those at the closer viewing distance (*Near* 22.5%, *Far* 7.5%). This observation may be important for interpreting the pattern of results. For now, the main message is that real faces were rarely mistaken for masks.

The critical issue is the performance of the *High-realism mask* group relative to the two comparison groups. Of the 160 participants in this group, only two (1.3%) used the word 'mask' in their responses to the Open question. For the Prompted question, this number rose to five (3.1%). All five of these participants were in the *Near* condition. Given this very low rate of spontaneous detection, the rest of the analysis focuses on responses to the 2AFC item (Question 3).

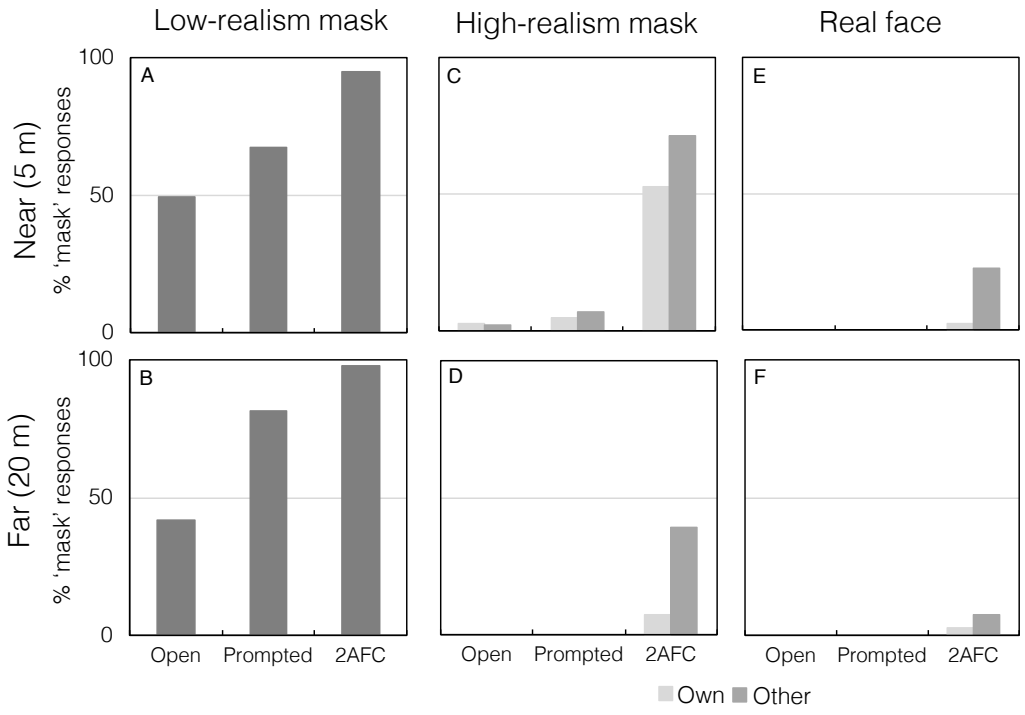

Figure 2.5. Mask detection data from Experiment 3. Bars show the percentage of 'mask' responses to Open, Prompted, and 2AFC questions about the experimental confederate. Responses are broken down by realism (Low-realism mask, left panels AB; High-realism mask, centre panels CD; Real face, right panels EF) and by viewing distance (Near, upper panels ACE; Far, lower panels BDF). For the High-realism mask and Real face conditions, responses are shown separately for Own-race (light grey) and Other-race (mid grey). Sample sizes for each panel: A, 44; B, 41; C, 82; D, 78; E, 81; F, 81.

Analysis of 2AFC responses

Effects of realism. Figure 2.5 shows a clear separation between the *Low-realism mask* and *Real face* conditions, with intermediate performance in the *High-realism mask* condition. Chi-square analysis confirmed a significant difference between conditions [$X^2(1) = 179.28$, $p < .001$, $v = .66$]. Post-hoc tests revealed that 'mask' responses in the *High-realism* condition (42.5%) were significantly less frequent than in the *Low-realism* condition (96.5%) [$X^2(1) = 141.61$, $p < .001$, $v = .53$], and significantly more frequent than in the *Real face* condition (8.6%) [$X^2(1) = 112.61$, $p < .001$, $v = .39$]. Interestingly, the rate of 'mask' responses in the *High-realism* condition was not significantly different from 50% [$X^2(1) = 3.60$, $p = .058$], indicating low consensus or low confidence in these responses.

Effects of race. The rate of 'mask' responses was higher overall in the *Other-race* condition than in the *Own-race* condition [$X^2(1) = 16.23$, $p < .001$, $v = .22$]. Importantly, this effect was present not only in the *High-realism* condition [$X^2(1) = 12.38$, $p < .001$, $v = .28$], but also in the *Real face* condition [$X^2(1) = 7.55$ $p = .005$, $v = .22$], suggesting that it may reflect a decision bias rather than a difference in perceptual discrimination.

Effects of viewing distance. Overall, 'mask' responses were more frequent in the *Near* condition than in the *Far* condition [$X^2(1) = 16.66$, $p < .001$, $v = .21$]. Post-hoc comparisons revealed that this effect was due to increased mask responses in the *High-realism* condition only [$X^2(1) = 26.70$, $p < .001$, $v = .41$]. There was no effect of viewing distance for the *Real face* condition [$X^2(1) = 2.66$, $p = .103$, $v = .13$] or the *Low-realism* condition [$X^2(1) = .42$, $p = .520$, $v = .07$].

Discussion

Hyper-realistic masks were very rarely detected in this experiment. At the longer viewing distance (20 m), no one in the *High-realism* condition reported a

mask. Even at close range (5 m), only 2 out of 82 viewers reported a mask spontaneously, rising to 5 out of 82 for the prompted question. For the 2AFC question, the proportion of participants who guessed that they were in the *Mask* condition ranged from 7.5% (*Own-race*, *Far* condition) to 71% (*Other-race*, *Near* condition), depending on race and viewing distance. Importantly, these factors similarly affected responses in the *Real face* condition.

One possible explanation for the elevated 'mask' responses in the *Other-race* condition is that participants' judgements incorporated demographic base-rates. In Japan, Western faces are less frequent than Asian faces. In the UK, Asian faces are less frequent than Western faces. This uneven distribution gives rise to different prior probabilities. At the same time, the finding that 'mask' responses were more frequent in the *High-realism mask* condition than in the *Real face* condition, and more frequent in the *Near* condition than in the *Far* condition, implies that subtle visual cues also played a role. Taken together, these observations suggest separable contributions from prior probability and visual evidence to participants' decisions.

2.6 General Discussion

Part of our interest in hyper-realistic masks stems from their use in security settings. At first sight, it is difficult to credit that a person wearing a full mask could board a plane unchallenged. How are we to make sense of such incidents? Do they reflect inattention on the part of the observer, or perhaps an unwillingness to confront the mask wearer? Or could it be that, in these situations, hyper-realistic masks are indistinguishable from real faces? In our experiments, almost no one reported noticing the mask, despite attending to the mask and answering several questions about its appearance. This was true for photographic images presented onscreen. It was also true for live confederates presented outdoors. The numbers are sobering. Of the 280 participants who viewed hyper-realistic masks in these studies (60 in Experiment 1; 60 in Experiment 2; 160 in Experiment 3), only 2 spontaneously reported the mask, and only 3 more reported the mask following

further prompting. Interestingly, all 5 of these participants viewed the mask live (Experiment 3), and at the closer viewing distance of 5 m. These are low detection rates. Evidently, the information available even in near-distance, live viewing (visual detail, 3D form, motion) did not allow viewers to distinguish hyper-realistic masks from real faces with any generality. Nevertheless, the clustering of these few participants by viewing condition suggests that the available information may have some diagnostic value, above and beyond that which is available at longer viewing distances or in photographic presentations.

Other aspects of our results bear out this interpretation. In Question 3 of each experiment, we asked participants to guess whether they were in the *Mask* condition or the *No mask* condition (2AFC). The intention here was to draw out more covert detection of hyper-realistic masks—perhaps arising from an uncanny valley phenomenon. We anticipated that the wording of Question 3, combined with the sensitivity of 2AFC as a measure, might lead to a ceiling effect in responses, with all participants guessing that they were in the *Mask* condition. As it turned out, 2AFC performance did not approach ceiling in any of the experiments (with the planned exception of the low-realism masks in Experiment 3). Instead, 'mask' responses were the minority in Experiment 1, Experiment 2, and the *Far* condition of Experiment 3. Even in the *Near* condition of Experiment 3, ‘mask’ responses were not reliably above 50%.

Presumably, there must be some critical distance at which viewers spontaneously and accurately distinguish hyper-realistic masks from real faces. After all, painted silicone and human skin are different materials with different surface properties (Motoyoshi et al., 2007).

We do not know what this critical distance might be, but we can now be confident that the Near distance in Experiment 3 (5 m) exceeds it. That finding may have implications for mask detection in the real world. Classic work on proxemics (Hall, 1966) divides interpersonal space into four radial zones. In this scheme, intimate distance (0–1.5 feet; 0–0.5 m) is associated with physical contact and whispering, personal distance (1.5–4 feet; 0.5–1.2 m) is reserved for interactions among close friends or family, social distance (4–12 feet; 1.2–3.7 m)

accommodates interactions among acquaintances, and public distance (>12 feet; >3.7 m) is occupied by strangers. Our upper bound of 5 m suggests that any critical distance for mask detection falls within social space (4 to 12 feet; 1.2 m to 3.7 m) or closer in this scheme. But most people do not enter this space. Strangers in particular tend to be seen at longer range, where we now know mask detection is unreliable. One important exception is photo-ID checks (e.g. passport control), which are typically carried out at a distance of one or two metres (Noyes & Jenkins, 2017; Verhoff, Witzel, Kreutz, & Ramsthaler, 2008). Future studies should assess mask detection performance at this closer range. However, anecdotal reports of mask use on airlines (Zamost, 2010), and the prevalence of identification errors in live-to-photo comparisons (Davis & Valentine, 2009; Kemp, Towell, & Pike, 1997; White et al., 2014), do not inspire confidence.

These proxemic considerations raise some interesting questions about the appearances of hyper-realistic masks and their social effects. To date, mask manufacturers have followed a single strategy for evading detection: the pursuit of ever greater realism. An interesting direction for future research would be to assess the viability of a complementary strategy: evading detection by manipulating the behaviour of onlookers. It is almost tautological that the less approachable a mask looks, the less inclined viewers will be to approach it, and the less likely they will be to reach the critical distance for detection. A similar argument could be made for attractiveness. To the extent that facial attractiveness summons attention (Shimojo et al., 2003; Sui & Liu, 2009) and increases dwell time (Leder et al., 2010), a less attractive mask should receive less scrutiny. Based on such principles, it may be possible to devise a hyper-realistic mask that deflects observers' minds by (i) maximising viewing distance and (ii) minimising visual attention. A brutish-looking pickpocket might arrive at a different set of priorities, favouring a highly approachable mask that allows them to move closer to a target.

In future studies, it would be interesting to isolate the information that leads viewers to guess they are in the *Mask* condition. The fact that ‘mask’ responses were more prevalent in the *Near* condition than the *Far* condition suggests that high spatial frequency information plays an important role. However, it is not clear

whether decisions are driven by local visual features (e.g. surface discontinuities around the eyes or mouth), by more holistic visual features (e.g. wrinkle patterns over the whole face), or by higher-level inferences that are abstracted from such information (e.g. social attributions based on facial appearance). If reliable cues can be established, they could potentially form the basis of a training program aimed at enhancing mask detection. For passive viewing situations, such as reviewing recorded footage, this could be as simple as encouraging observers to monitor for particular visual features.

For interactive situations, such as live identity checks, more active approaches may be feasible. Our informal observation is that wearing a hyper-realistic mask attenuates some forms of facial movement. Even with good contact between the face and the mask, manipulating the mask places additional demands on facial muscles, relative to normal facial movement. Moreover, movements that may be clear and distinct at the internal surface of the mask (where they are initiated) will be partly absorbed by the silicone on their way to the external surface (where they are seen). These attenuation effects may be negligible for coarse movements such as rotation of the head on the neck, and opening and closing of the jaw. But emotional expressions such as smiles and frowns generally appear muted, and subtle expressions are often lost altogether. This, in turn, may be affecting social inference judgements (see appendix 2.3 for summary of social characteristic ratings in Experiment 1 and 2).

The overall facial impression, at least in extended interactions, is one of blunted animacy. It is possible that, under appropriate testing conditions, this impression might be enough to cue detection of a hyper-realistic mask, perhaps by tipping the interaction into the uncanny valley. However, it may also encourage false positives for low-animacy real faces. Thus blunted animacy in the face may be more diagnostic when it is paired with incongruous animacy cues from the body or voice. Various aspects of facial appearance—including apparent age, gender, and emotion—can shape viewers' expectations about how a person is likely to move and speak (e.g. Johnson, McKay, & Pollick, 2011; Lander et al., 2007; Montepare & Zebrowitz-McArthur, 1988; Van den Stock, Righart, & De Gelder, 2007). Violations of those expectations, such as sprinting centenarians,

may allow viewers to infer the presence of a mask, even if the mask itself is entirely convincing.

Speech could be revealing for other reasons too. Normal speech comprehension is strongly supported by visual lipreading (Campbell, 2008; McGurk & MacDonald, 1976). However, the lips of a hyper-realistic mask fully cover the lips of the wearer (see Figure 2.1). This arrangement has a number of implications for speech and lipreading. First, it introduces a physical barrier between the wearer's lips, presumably impeding production of phonemes that require contact between the lips (e.g. /b/, /p/, /m/), or between the teeth and the lower lip (e.g. /f/, /v/). Second, it reduces the pliability of the whole mouth area, presumably impeding articulation more generally. Reduced lip movement implies reduced visual support for speech understanding (Campbell, 2008). It also suggests that hyper-realistic masks may affect the auditory stream in distinctive ways. Ironically, auditory information may provide the best hope of solving this difficult visual task.

Perception of emotional expression, uncanny valley effects, cue integration, and speech comprehension are all matters that can be unpicked experimentally. Our observation (Experiment 1) of elevated detection rates for participants with prior knowledge of hyper-realistic masks suggests that training to enhance performance is possible at least in principle. The optimal form of training remains to be determined.

We also tested for other-race effects in mask detection. Other-race effects were originally observed in the context of face *identification*—a task that requires fine perceptual discriminations. Given that distinguishing hyper-realistic masks from real faces also requires fine perceptual discriminations, we wondered whether performance would be poorer for other-race faces than for own-race faces. The evidence on this particular point was not very clear. Floor effects in the Open question and Prompted question make it difficult to draw any conclusions about race effects in overt detection, beyond noting that the task defeated own-race and other-race viewers alike. The same manipulation did have some impact on responses to the 2AFC item, but even here the different experiments present a

mixed picture. Experiment 1 (UK participants) and Experiment 2 (Japanese participants) were both based entirely on Western face images. Comparing across experiments, Japanese viewers were somewhat more likely than UK participants to guess that they were in the *Mask* condition (rather than the *No mask* condition), but this difference was not statistically significant. Experiment 3, using a fully crossed design and a larger sample, found a significant difference in the same direction: other-race viewers were more likely than own-race viewers to guess that they were in the *Mask* condition. On its own, this effect might suggest an other-race *advantage* in distinguishing real faces from hyper-realistic masks. That would contrast with the other-race *disadvantage* that is standard in identification tasks. But the *Real face* condition undermines this interpretation. For real faces too, other-race viewers were disproportionately likely to guess that they were in the *Mask* condition. That finding is not consistent with an other-race advantage in distinguishing real faces from hyper-realistic masks. Instead, it suggests an overall bias towards guessing 'mask'.

This interpretation of the 2AFC data accords with the array challenge findings (Experiments 1 and 2). In the array challenge, Japanese participants picked out the mask significantly less often than the UK participants. Given that the stimuli were Western face images, this pattern resembles the expected disadvantage for other-race faces. It is not obvious how one might square an other-race disadvantage in the array challenge with an other-race advantage in the 2AFC. But no such tension arises between an other-race disadvantage in the array challenge and a decision bias in the 2AFC.

Why might other-race viewers be especially inclined to guess that they are in the *Mask* condition? One possibility is that, at least in the campus locations we tested, other-race faces are simply less prevalent than own-race faces. That being the case, if the confederate presents an other-race face, the participant has to explain the balance of probabilities. Either they just happen to be witnessing a (relatively) rare event, or they are subject to an experimental manipulation. Presumably, some proportion of participants finds the latter explanation more compelling than the former. If this argument is sound, we expect that equating the frequencies of own-race and other-race stimuli in a laboratory experiment should

give rise to an other-race disadvantage.

Hyper-realistic masks fool most of the people most of the time. This finding should be unsettling, not least because it indicates a new frontier in deception. Covering the face may be grounds for suspicion when the intent is to conceal identity. But historically, such deception has been easy to detect. In hyper-realistic masks, we confront the prospect of face coverings that shroud the wearer, yet are themselves accepted as real faces. It is difficult to estimate how many of these masks are already in circulation. But as documented cases attest (see Appendix 1.1), their proliferation poses a challenge for face recognition in applied settings, including crime prevention and border control. We expect that increasingly sophisticated manufacturing techniques will continue to improve the quality of these masks and to drive prices down. Keeping pace with these improvements will require increasingly sophisticated countermeasures, perhaps including consciousness raising, personnel development, and supplementary imaging methods. Machine vision researchers have made some interesting progress on this front (e.g. Erdogmus & Marcel, 2014; Kose & Dugelay, 2013). The conditions are conducive to a new arms race in face identification between deception and detection.

Chapter 3.

A Turing test for synthetic faces

3.1 Summary

In Chapter 2 we showed that hyper-realistic face masks can pass for real faces during live viewing. However, live viewing embeds the perceptual task (mask detection) in a powerful social context that may influence respondents' behaviour. To isolate the perceptual component of the task, we assessed viewers' ability to distinguish photos of hyper-realistic masks from photos of real faces in a computerised 2AFC procedure. In Experiment 4 (N = 120), we observed an error rate of 33% when viewing time was restricted to 500 msec. In Experiment 5 (N = 120), we observed an error rate of 20% when viewing time was unlimited. In both experiments we saw a significant performance cost for other-race comparisons relative to own-race comparisons. We conclude that viewers could not reliably distinguish hyper-realistic face masks from real faces in photographic presentations. As well as its theoretical interest, failure to detect synthetic faces has important implications for security and crime prevention, which often rely on facial appearance and personal identity being related.

3.2 Introduction

Technologies often imitate natural objects, giving rise to artificial diamonds, artificial flowers, artificial fur, and countless other artifacts. How are we to judge the success of such imitations? In 1950, Alan Turing proposed an influential answer for the specific case of artificial intelligence: an imitation is successful when we cannot distinguish it from the real thing (Turing, 1950). In his original argument, Turing imagined a human evaluator engaged in natural language conversations with a real human and a computer designed to generate human-like responses. The evaluator would be informed that one of the two partners is a

computer, and asked to determine which one. To focus the evaluation on quality of thought rather than quality of speech, the dialogue would be mediated by text only (e.g. keyboard and screen). If the evaluator cannot reliably distinguish the computer from the human, the computer is said to pass the test (see Figure 3.1).

As a target of imitation, intelligent conversation is enormously complex. No current machine is close to passing the Turing Test. However, the logic of the test itself is straightforward, and provides a means for assessing the maturity of imitation technologies generally: given the imitation alongside the real thing, can an observer tell which is which?

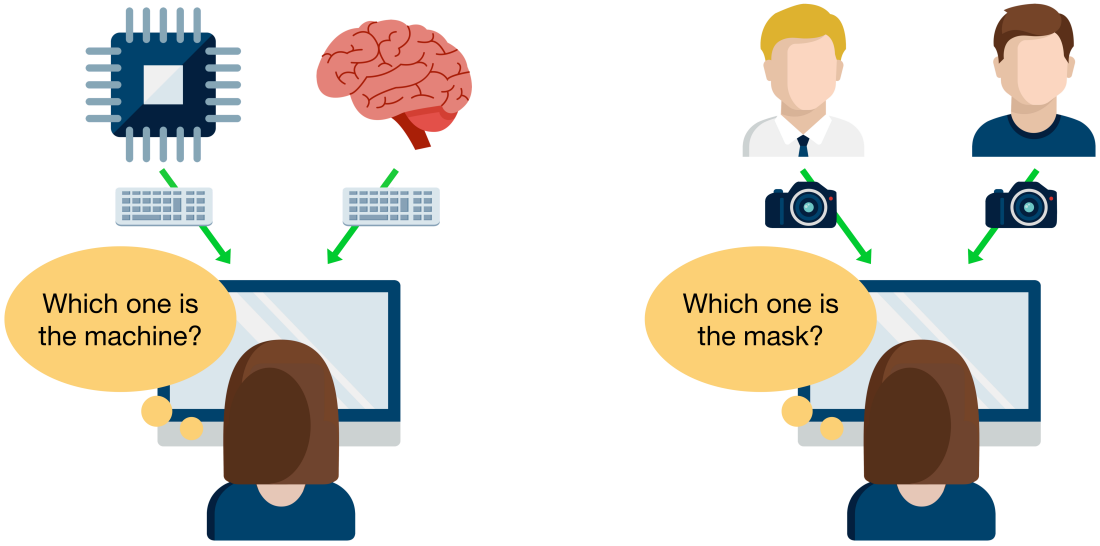

Figure 3.1. Schematic illustrating parallels between the standard Turing Test (left) and a similar test for synthetic faces (right). In both cases, an evaluator is given the task of trying to determine which presentation is the genuine article and which is the imitation. The evaluator is limited to using a computer interface to make the determination.

Here we bring this logic to bear on a much more tightly circumscribed imitation technology—artificial faces. The past decade has seen increasing interest in the realism of computer generated faces (Holmes, Banks, & Farid, 2016; Nightingale, Wade, & Watson, 2017). Our concern is artificial face images of a very different kind, specifically, unretouched photos of artificial faces in the real world. Images in this category differ from digital images in at least two important ways. First, digitally generated or manipulated images are not snapshots of reality. They only

exist in print and on screen, and that limits the ways in which they can be encountered. Our focus is physical artifacts that exist in the real world and are caught on camera. Second, digital image manipulation has been a part of mainstream media for a generation. As such, the level of public understanding that images may be 'photoshopped' is high. One consequence of this development is that photorealistic images carry less evidential weight than they once did—all images are suspect (see Kasra, Shen, & O'Brien, 2018). Since the real world cannot be photoshopped in the same way, physical artifacts are more protected from this slide in credibility.

Artificial faces in the real world may not be intended to pass for genuine faces, even when they strive for realism in some sense. A marble bust might capture the proportions of a real face, but none of the movement; a robotic head might capture some facial movements, but remain disembodied. Hyper-realistic silicone masks differ from these examples in that they are worn by a real person, and so are seen in the context of a real body. Moreover, they are constructed from a flexible material, so they relay the wearer's rigid and non-rigid head movements. These characteristics set hyper-realistic masks apart from other artificial faces, as they allow them to be fully embedded in natural social situations.

These natural social situations place unusual demands on imitation technologies, as humans tend to be especially attuned to social stimuli. Face perception offers abundant evidence of such tuning. For example, humans are predisposed to detect face-like patterns (Robertson, Jenkins, & Burton, 2017), and this tendency is present from early infancy (Morton & Johnson, 1991). Faces capture our attention (Langton et al., 2008; Theeuwes & Van der Stigchel, 2006), and having captured attention, tend to retain it (Bindemann et al., 2005). While viewing a face, we make inferences about the mind behind it, including emotional state from facial expression (Ekman & Friesen, 1971; Young et al., 1997), and direction of attention from eye gaze (Baron-Cohen et al., 2001; Friesen & Kingstone, 1998). For people we know well, we also identify the individual (Burton, Bruce, & Hancock, 1999; Burton, Jenkins, & Schweinberger, 2011), which can trigger retrieval of personal information from memory (Bruce & Young, 1986). All of these processes require high sensitivity to subtleties of facial

appearance. There is even some evidence that these processes can become tuned to specific populations through social exposure. For example, children tend to be better at recognising young faces than old faces (and vice versa; Anastasi & Rhodes, 2005; Neil et al., 2016); Japanese viewers tend to be better at recognising East Asian faces than Western faces (and vice versa; O'Toole et al., 1994). Perhaps most relevant for the current study, discrimination between faces and non-face objects can be accomplished rapidly and accurately. Using saccadic reaction times, Crouzet, Kirchner, and Thorpe (2010) found that viewers could differentiate images of faces versus vehicles at 90% accuracy in under 150 milliseconds—significantly faster than discriminations that did not involve faces. Although Crouzet, Kirchner, and Thorpe's (2010) findings were based on images from different categories, they nonetheless provide an interesting baseline against which to compare the more nuanced discriminations investigated here.

Taken together, these findings suggest that faces may be particularly difficult objects to imitate. Faces attract the glare of attention, and details of their appearance convey socially significant information. Even so, there is some evidence that hyper-realistic silicone masks can pass for real faces, at least in certain situations. In Chapter 1 (Experiment 3), passers-by consistently failed to notice that a live confederate was wearing a hyper-realistic mask, and showed little evidence of having detected the mask covertly. Out of 160 participants in the critical condition, only two spontaneously reported the mask, and only three more reported the mask following further prompting. These low detection rates are consistent with the idea that hyper-realistic masks successfully imitate real faces. However, several aspects of the experimental procedure complicate this interpretation. For example, masks were not mentioned during the main phase of data collection, and participants had no reason to expect to see a mask. It is possible that participants might have detected the masks more often had they been expecting them. Moreover, responses were collected in a live social setting. It is possible that respondents were reluctant to inspect or to discuss the appearance of a person who was physically present (albeit out of earshot)—and especially reluctant to declare that person's face to be artificial.

These matters of interpretation arise in part from our approach to testing,

which prioritised ecological validity over experimental control. Here we adopt the complementary approach of two-alternative forced choice testing (2AFC), which strikes the opposite balance (see Bogacz et al., 2006 for a review). The 2AFC method originated in psychophysical research (Fechner, 1860/1866), where it was developed to measure quantities such as perceptual acuity. Our application is closer in spirit to the Turing Test, in that our main interest concerns the realism of artificial stimuli.

In 2AFC testing, the participant is presented with two stimuli, one of which is the target, and is forced to choose which is the correct stimulus. This contrasts with the tasks that we used in Chapter 1 (Experiment 1 and 2), in which participants viewed individual stimuli, and later made categorical judgements. There are several reasons why the proposed 2AFC testing should sharpen observers' ability to distinguish hyper-realistic masks from real faces. First, the task instructions ensure that participants are aware in advance of the masks. Second, social influence is minimised, as the task is computer based. Third, the task always involves two stimuli at a time: one is always a mask and the other is always a real face. Thus even when participants are uncertain whether one of the images is the target, they can still solve the task indirectly if they are certain about the other image.

To test for other-race effects in this task, we collected data in both the UK and Japan. Although other-race effects are most strongly associated with identity-based tasks, such as face recognition (Meissner & Brigham, 2001) and face matching (Megreya, White, & Burton, 2011), our question here is whether they can also arise when distinguishing hyper-realistic masks from real faces (Robertson, Jenkins, & Burton, 2017). The live viewing study in Chapter 1 could not address this point fully, as in naturalistic settings, the background probabilities of encountering own-race and other-race faces are not well matched. The 2AFC task gets around this limitation by allowing us to present own-race and other-race items equally often. We expect that equating background probabilities in this way will allow us to reach a more definitive answer.

3.3 Experiment 4:

Discriminating masks from faces with limited exposure time

To assess participants' ability to distinguish hyper-realistic masks from real faces, we constructed a computer-based 2AFC task in which participants viewed pairs of on-screen images (one face and one mask), and indicated via key press which of the two images showed the mask. To isolate effects of mask realism, we compared performance for high-realism masks and low-realism masks that were easy to detect. We expected that reaction times would be markedly slower in the high-realism condition than in the low-realism condition.

To test for other-race effects, we also presented equal numbers of own-race and other-race trials. The standard perceptual explanation of the other-race effect is that viewers become attuned to the variability that surrounds them, and remain relatively insensitive to variability outside of this range (e.g. O'Toole, Deffenbacher, Valentin, & Abdi, 1994). These differences in perceptual experience lead to more efficient perceptual discrimination for own-race faces than for other-race faces. Although these effects are usually demonstrated using identification tasks, the same argument also applies to distinguishing hyper-realistic masks from real faces. We thus predicted shorter response latencies for own-race faces than for other-race faces in this task.

Method

Ethics statement. Ethical approval for the experiments in this study was obtained from the departmental ethics committee at the University of York and Kyoto University.

Participants. 120 volunteers took part in exchange for a small payment or course credit. These were 60 members of the volunteer panel at the University of

York (39 female, 21 male; mean age = 23, age range 18–39 years) and 60 members of the volunteer panel at Kyoto University (27 female, 33 male; mean age = 22, age range 18–50 years). Testing took place on site at Kyoto University, Japan, and the University of York, UK.

Design and Stimuli.

Three types of photographic image were used to construct the stimulus pairs—High-realism masks, Low-realism masks, and Real faces. To allow a fully-crossed design, we collected an equal number of Asian and Caucasian images for each category. To ensure that we sampled real world image variability, we used ambient images throughout (Jenkins et al., 2011). In the High-realism condition, a real face was paired with a hyper-realistic silicone mask. In the Low-realism condition, a real face was paired with a non-realistic party mask.

High-realism mask images. To collect images of high-realism masks, we entered the search terms ‘realistic masks’, ‘hyper-realistic masks’ and ‘realistic silicone masks’ into Google Images. We selected images that (i) exceeded 150 pixels in height, (ii) showed the mask in roughly frontal aspect, (iii) showed the eye region without occlusions, and (iv) included eyebrows made with real human hair. We used the same criteria to search the websites of mask manufacturers (e.g. RealFlesh Masks, SPFX, CFX) and topical forums on social media (e.g. Silicone Mask Sickos, Silicone mask addicts). For each of the Asian and Caucasian image sets, we gathered 37 hyper-realistic mask images that met the inclusion criteria (74 High-realism mask images in total).

Low-realism mask images. For comparison, we collected 74 images of low-realism masks by combining the search terms ‘Caucasian’ and ‘Asian’ with terms such as ‘Halloween’, ‘party’, ‘mask’, ‘masquerade’, ‘face-mask’, and ‘party mask’ in Google Images, and selecting the first images that met the inclusion criteria i-iii above. For low-realism mask images, race referred to the mask wearer, and was apparent from the parts of the face that were not occluded.

Real face images. We also collected 148 real-face images to pair with the 74 high-realism and 74 low-realism mask images (148 mask images in total). To ensure that the demographic distribution among our real face images was similar to that portrayed by the high-realism masks, we combined the search terms 'Caucasian' and 'Asian' with the terms 'young male', 'old male', 'young female', and 'old female' in Google Images. We then accepted images that met criteria i-iv until the distribution of faces across these categories was the same as for the High-realism mask images. All photos were cropped to show the head region only and resized to 540 pixels high x 385 wide for presentation (see Figure 3.2).

To create the stimulus displays, we paired each Real face image with a mask image from either the High realism or the Low realism set. On each trial, the mask was equally likely to appear on the left or right side of the display. Stimuli always paired two images showing the same race (i.e. both Asian or both Caucasian). Within these constraints, image pairings were randomised separately for each participant, such that each participant saw each image exactly once, but judged different image combinations. In both the UK group and the Japan group, participants were randomly assigned to either the own-race or the other-race condition.

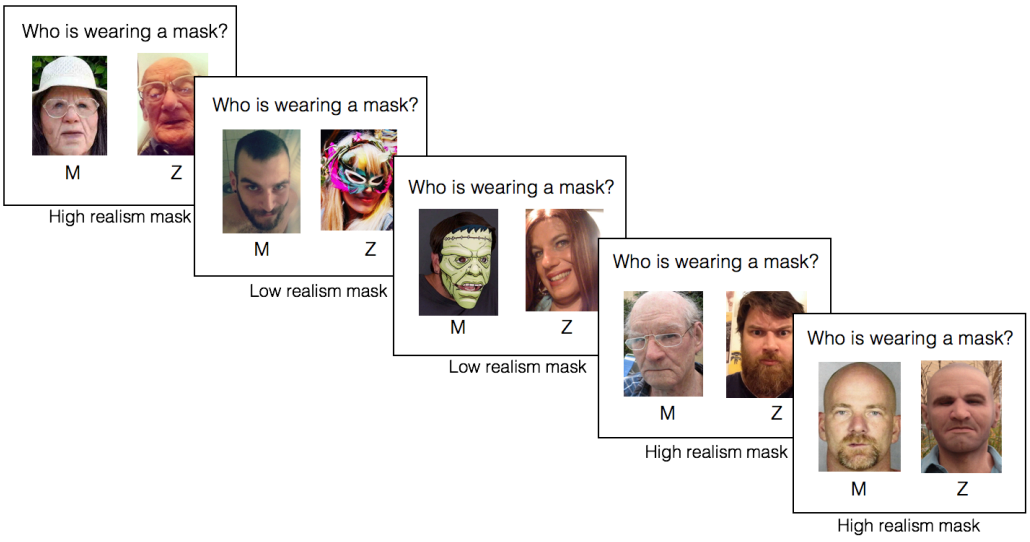

Figure 3.2. Example trials from Caucasian image set. Each mask image was randomly paired with one real face image from the set, independently set for each participant.

Procedure. Participants were instructed that each stimulus pair contained one real face and one mask, and that the task was to indicate via keypress which image showed the mask. Each trial began with an image pair presented at the centre of the screen for 500 msec with the caption '*Who is wearing the mask?'* immediately below, and response options 'Z' and 'M' below the left and right images respectively. After 500 msec, the images were removed, and the question and response options remained onscreen until response. Participants pressed 'Z' for the left image, or 'M' for the left image as quickly and accurately as possible, and the response initiated the next trial. Each participant saw three practice trials followed by 74 recorded trials in a random order. The entire experiment took approximately 10 minutes to complete

Results

Reaction time and error data are summarised in Figure 3.3.

Reaction Times. Participants' mean correct reaction times (RTs) were submitted to a 2 x 2 mixed ANOVA with the within-subjects factor of Mask Type (High Realism, Low Realism), and the between-subjects factor of Race (Own, Other).

As expected, there was a significant main effect of Mask Type, with slower responses for High Realism trials (M = 1258 msec, SE = 40.8, CI = 1178 – 1339) than for Low Realism trials (M = 921 msec, SE = 29.3, CI 857 – 971), [$F(1,118) = 204.6$, $p < .001$, partial $\eta^2 = .63$].

There was also a significant main effect of Race, with slower RTs in the Other-race condition (M = 1197, SE = 103.5, CI = 994 – 1399) than in the Own-race condition (M = 976 msec, SE = 76.6, CI = 826 – 1125), [$F(1,118) = 11.97$, $p < .001$, partial $\eta^2 = .09$]. The interaction between Mask Type and Race did not reach significance [$F(1,118) = 3.60$, $p = .060$].

Simple main effects confirmed that there was a significant effect of Mask

Type for both Own-race [F(1,118 = 76.96, p < .001, partial η^2 =.40] and Other-race faces [F(1,118 = 131.26, p < .001, partial η^2 =.53]. The effect of Race was also present in both the High Realism condition [F(1,118) = 11.62, p = .001; partial η^2 = .09] and the Low Realism condition [F(1,118) = 9.61, p = .002; partial η^2 = .08].

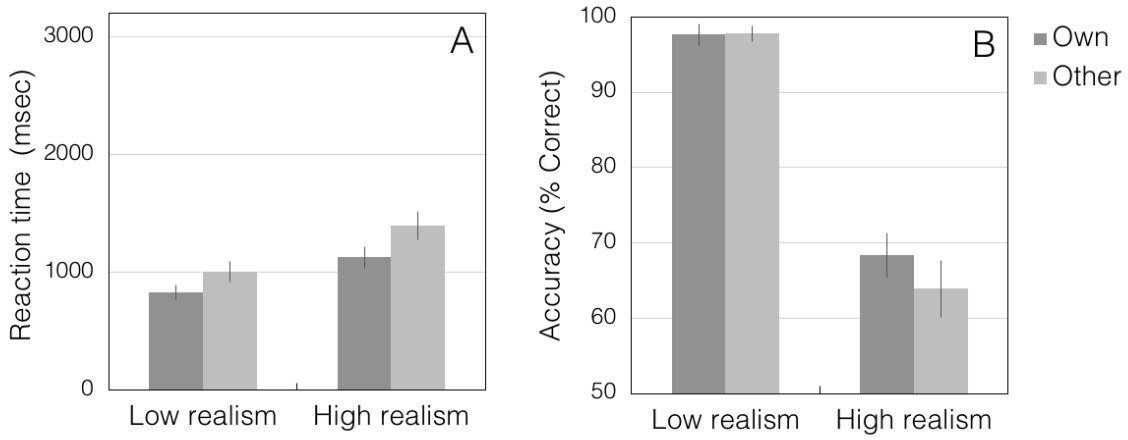

Figure 3.3. Reaction times (A) and percentage correct performance (B) in Experiment 4. Error bars show 95% confidence intervals.

Errors. Mean percentage correct scores were likewise submitted to a 2 x 2 mixed ANOVA with the within-subjects factor of Mask Type (High Realism, Low Realism), and the between-subjects factor of Race (Own, Other).

The analysis revealed a significant main effect of Mask Type, with lower accuracy for High Realism trials (M = 66.2%, SE = 1.2, CI = 63.8 – 68.8) than for Low Realism trials (M = 97.7%, SE = 0.4, CI = 96.9 – 98.6), [F(1,118) = 635.8, p <.001, partial η^2 =.84].

There was no main effect of Race (Own-race: M = 83.0%, SE = .8, CI = 81.5 – 84.6; Other-race: M = 80.9%, SE = .9, CI = 79.2 – 82.5), [F(1,118) = 2.69, p = .104], and no significant interaction between Mask Type and Race [F(1,118) = 3.44, p = .066].

Simple main effects confirmed that there was a significant effect of Mask Type in both the Own-race condition [F(1,118 = 272.85, p < .001, partial η^2 =.70]

and the Other-race condition [$F(1,118 = 366.33$, $p < .001$, partial $\eta^2 = .76$]. Despite the numerical trend, there was no significant effect of Race in the High Realism condition [$F(1,118) = 3.45$, $p = .066$], nor in the Low Realism condition [$F(1,118) = .02$, $p = .880$].

Discussion

Analysis of reaction times showed that 2AFC discrimination of masks from real faces was indeed slower for High realism masks than for Low realism masks (~300 msec RT cost). As it turned out, the more interesting effect was in the error data. Participants performed almost perfectly in the Low realism condition (98% accuracy). That is perhaps not surprising, given the simplicity of the task. However, accuracy in the High realism condition was just 66%, in the context of chance performance being 50%. An error in this 2AFC task is striking, as it requires the observer to choose the real face over the alternative, *when the alternative is a mask*. The implication is not merely that the hyper-realistic masks looked human. In some cases, they appeared more human than human in this task. That was the judgement in one-third of the High realism trials.

We also observed an effect of Race in reaction times (~200 msec cost), though not in the accuracy data. If reliable, this is an intriguing finding, as it potentially extends the classic other-race effect from identification tasks to the very different task of differentiating real faces from synthetic faces (masks).

One aspect of our experiment that complicates interpretation is the limited exposure duration for the stimuli (500 msec). Limiting stimulus duration is standard practice when the task would otherwise be too easy (Bogacz et al., 2006). As it turned out, the High realism condition was far from easy. In the next experiment, we removed this time limit.

3.4 Experiment 5:

Discriminating masks from faces with unlimited exposure time

In Experiment 4, mask realism affected not only the speed of mask/face discriminations, but also their accuracy. One plausible interpretation of this result is that the hyper-realistic face masks were difficult to distinguish from real faces. However, another possibility is that the stimulus presentations were too brief (500 msec) to allow proper comparison of the two images. To distinguish these alternatives, we repeated the preceding experiment with one important change: stimuli now remained on screen until the participant responded. If errors in Experiment 1 were due to insufficient viewing time, then unlimited viewing time should eliminate them. On the other hand, if the errors were due to the similarity of the masks to the faces, the error rate in the High realism condition should remain high.

Method

Participants. 120 new volunteers, none of whom participated in Experiment 4, took part in exchange for a small payment or course credit. These were 60 members of the volunteer panel at the University of York (51 females, 9 males; mean age = 20, age range 18–29 years) and sixty members of the volunteer panel at Kyoto University (23 females, 37 males; mean age = 21, age range 18–38 years). Once again, testing took place on site at Kyoto University, Japan, and the University of York, UK.

Stimuli and design. The stimuli and design were the same as in Experiment 1, except that the stimulus pairs now remained on screen until the participant responded.

Procedure. The procedure was also the same as in Experiment 4, except for the unlimited viewing time. Task instructions were modified to emphasise that the task was self-paced and that there was no time limit.

Results

Reaction time and error data are summarized in Figure 3.4.

Reaction Times. As in Experiment 4, participants' mean correct reaction times (RTs) were submitted to a 2 x 2 mixed ANOVA with the within-subjects factor of Mask Type (High Realism, Low Realism), and the between-subjects factor of Race (Own, Other).

Once again, there was a large main effect of Mask Type, with slower responses for High Realism trials (M = 2146 msec, SE = 109.6, CI = 1931 – 2360) than for Low Realism trials (M = 977 msec, SE = 33.9; CI = 911 – 1044), [$F(1,118) = 213.2$, $p < .001$, partial $\eta^2 = .64$].

There was also a significant main effect of Race, with slower RTs overall for Other-race trials (M = 1787 msec, SE = 219.8, CI = 1356 – 2217) compared with Own-race trials (M = 1337 msec, SE = 142.9, CI = 1057 – 1617), [$F(1,118) = 11.7$, $p < .001$, partial $\eta^2 = .09$]. On this occasion, there was a significant interaction between Mask Type and Race [$F(1,118) = 21.3$, $p < .001$, partial $\eta^2 = .15$].

Simple main effects confirmed that there was a significant effect of Mask Type in both the Own-race condition [$F(1,118 = 49.86$, $p < .001$, partial $\eta^2 = .30$] and the Other-race condition [$F(1,118 = 184.66$, $p < .001$, partial $\eta^2 = .61$]. The effect of Race was driven specifically by the High Realism condition [$F(1,118) = 15.70$, $p < .001$, partial $\eta^2 = .18$], not the Low Realism condition [$F(1,118) = 1.40$, $p = .238$].

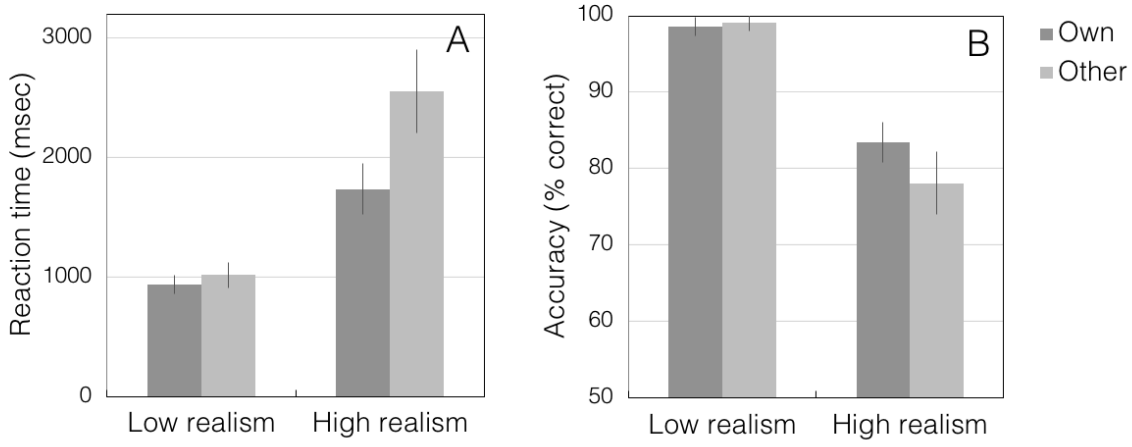

Figure 3.4. Reaction times (A) and percentage correct performance (B) in Experiment 5. Error bars show 95% confidence intervals.

Errors. Mean percentage correct scores were also submitted to a 2 x 2 mixed ANOVA with the within-subjects factor of Mask Type (High Realism, Low Realism), and the between-subjects factor of Race (Own, Other).

Accuracy was again lower for High Realism trials (M = 80.8% correct; SE = 1.3; CI = 78.3 – 83.2) than for Low Realism trials (M = 98.6%, SE = 0.42; CI = 98.0 – 99.7), [$F(1,118) = 228.4$, $p < .001$, partial $\eta^2 = .66$].

There was no overall main effect of Race on accuracy (Own-race: M = 91.0%, SE = .69, CI = 89.6 – 92.3; Other-race: M = 88.6%, SE = .95, CI = 86.8 – 90.5), [$F(1,118) = 2.73$, $p = .101$]. However, there was a significant interaction effect between Mask Type and Race [$F(1,118) = 6.08$, $p = .015$; partial $\eta^2 = .49$].

Simple main effects confirmed that there was a significant effect of Mask Type in both the Own-race condition [$F(1,118 = 79.97$, $p < .001$, partial $\eta^2 = .40$] and the Other-race condition [$F(1,118 = 154.47$, $p < .001$, partial $\eta^2 = .57$]. There was a significant effect of Race in the High Realism condition [$F(1,118) = 4.54$, $p = .035$; partial $\eta^2 = .40$], but not in the Low Realism condition [$F(1,118) = .47$, $p = .495$].

Discussion

Performance in the Low realism condition was virtually identical to Experiment 4. Accuracy was again almost perfect (99%) in this easy task. Response times were also similar, despite the unlimited presentation time, suggesting that additional time could not be used to further optimise performance. In the High realism condition, responses were much slower compared with the Low realism condition (~1100 msec cost), and compared with the High realism condition in Experiment 4 (~800 msec cost). Participants spent much longer on these difficult decisions, given the chance. However, even this unlimited viewing time did not come close to eliminating errors. For one out of every five High realism trials, participants judged the real face to be the mask.

As in Experiment 4, there was also an effect of Race in reaction times (~400 msec cost). This effect was carried mainly by the High realism condition. This time however, the other-race cost in accuracy was also statistically robust—again, in the High realism condition specifically (5% cost). Together, these measures indicate that distinguishing hyper-realistic masks from real faces was harder for other-race faces than for own-race faces.

3.5 General Discussion

To assess the realism of synthetic faces, specifically, hyper-realistic silicone masks, we tested how well viewers could distinguish photos of masks from photos of real faces in a 2AFC task. For low-realism masks, decisions were both fast and accurate. For high-realism masks, decisions were not only slower, but also surprisingly error prone. That was the finding in Experiment 4, when viewing time was restricted (33% errors). It was also the case in Experiment 5, when viewing time was unlimited (20% errors). Whether making snap decisions (Gladwell, 2005) or more deliberative judgements (Kahneman, 2011), participants could not reliably distinguish hyper-realistic face masks from real faces.

It was already evident from previous experimental work (Experiment 1-3)

and from real-world criminal cases (e.g. Malm, 2016; Weisman, 2018; Stanton, 2015) that hyper-realistic face masks can pass for real faces during live viewing. In principle however, other factors besides mask realism could account for those observations. For example, live viewing can place complex demands on attention, and challenging another person's appearance may be socially awkward. The current studies reach similar conclusions based on comparison of photographs under laboratory conditions.

Although the error rates seen here are high, they almost certainly underestimate error rates that would arise in more realistic settings. We chose the 2AFC task precisely because it provides a highly sensitive measure. Participants knew from the outset that their task was mask detection, whereas in everyday life that is not the default mindset. Participants also knew that every display contained a mask, whereas outside of the lab, the prevalence of hyper-realistic face masks is low (base rate is potentially important, as rare items are often missed; Wolfe, Horowitz, & Kenner, 2005). Finally, the mask in our displays was always one of two alternatives. The real world seldom presents the problem in such a convenient form. The more common task is to decide whether a single item is a mask or not (e.g. Stanton, 2015; Weisman, 2018). Experimentally, viewers make many more errors in that task, even when they are briefed in advance about hyper-realistic face masks (Experiment 1 and 2, Array challenge); and many more again when they are not (Experiment 1-3, Open and Guided questions).

None of this means that hyper-realistic mask detection is perceptually impossible. Accuracy in the current experiments was well above the chance level of 50%. However, in securing that level of performance, we have retreated quite far from the applied problem, and it is important not to lose sight of that retreat.

Both experiments showed a clear cost for other-race comparisons relative to own-race comparisons. This cost emerged in reaction time measures (Experiments 4 & 5) and also in error rates (Experiment 5). Other-race effects have been shown repeatedly in the context of identification tasks. The present study demonstrates a similar effect in the very different context of discriminating real faces from synthetic faces. This aspect of our findings is reminiscent of recent

work on social groups. Hackel, Looser, and Van Bavel (2014) presented stimuli that were generated by morphing real faces with doll faces to create intermediate blends. Viewers perceived less humanness in a morphed face when it was assigned to an out-group than when it was assigned to an in-group, indicating out-group dehumanisation. The same phenomenon could account for the other-race effect seen here, if out-group dehumanisation blunts the distinction between real faces and hyper-realistic face masks. One way to test this possibility would be to assess mask/face discrimination for identical stimuli using a 'minimal' group manipulation (Dunham, Baron, & Carey, 2011).

We began by comparing the challenge of distinguishing synthetic faces from real faces to the Turing Test. Our findings suggest that synthetic faces nearing the point where they fool viewers consistently. We see no reason to expect this imitation technology to stop improving now. People are rightly wary of photorealistic images because they know they can be manipulated. We may be entering a time where the same concerns apply to facial appearance in the real world.

Chapter 4.

Individual differences in mask detection

4.1 Summary

Hyper-realistic masks present a challenge to security and crime prevention. In Chapter 1 and 2 we have shown that people's ability to differentiate these masks from real faces is extremely limited. In this chapter, we consider individual differences as a means to improve mask detection. Participants categorised single images as masks or real faces in a computer-based task. Experiment 6 revealed poor accuracy (40%) and large individual differences (5–100%) for High-realism masks among Low-realism masks and Real faces. Individual differences in mask categorisation accuracy remained large when the Low-realism condition was eliminated (Experiment 7). Accuracy for mask images was not correlated with accuracy for real face images, or with prior knowledge of hyper-realistic face masks. Image analysis revealed that mask and face stimuli were most strongly differentiated in the region below the eyes. Moreover, High performing participants tracked the differential information in this area, but Low performing participants did not. Like other face tasks (e.g. identification), hyper-realistic mask detection gives rise to large individual differences in performance. Unlike many other face tasks, performance may be localised to a specific image cue.

4.2 Introduction

In a number of high-profile criminal cases, offenders have used hyper-realistic face masks (Figure 4.1) to transform their facial appearance, leading police to pursue suspects who looked nothing like the actual offenders (e.g., different race, age or gender; Sabawi, 2018; Weiner, 2017). In a separate incident, an airline passenger wearing a hyper-realistic mask boarded an international flight without the deception being noticed (Zamost, 2010). These cases suggest that, in

practical settings, hyper-realistic face masks can be difficult to distinguish from real faces. Experimental evidence bears out this conclusion. In a series of studies in Chapter 1, we examined incidental detection of unexpected but attended hyper-realistic masks in both photographic and live presentations. In all of these studies, viewers accepted hyper-realistic masks as real faces. These findings extend a tradition of research into realism of artificial stimuli. The Uncanny Valley phenomenon originally considered a range of human-like stimuli from puppets to robots (Mori, 1970; Mori, MacDorman, & Kageki, 2012). In recent years, the focus has shifted somewhat to computer-generated images (e.g. Nightingale, Wade, & Watson, 2017), but the very success of computer graphics has raised awareness that on-screen images may be digitally generated or enhanced. In Chapter 3, our data suggests that hyper-realistic masks too are nearing the point where they fool viewers consistently from photographic images. One of the interesting aspects of hyper-realistic masks is that they also fool the eye in the physical world (Chapter 2, Experiment 3), where digital image manipulation has not yet encroached.

Figure 4.1. Hyper-realistic face mask (left) worn by Dr Rob Jenkins (right).

The finding that spontaneous mask detection is unreliable suggests that

specific measures may be required if detection rates are to be improved. Here we pursue an individual differences approach to the problem. Over the last decade, individual differences have become an important topic in face perception research, not least because they suggest a route to improving performance in applied settings. For face identification, the range of ability is bracketed by two extremes. At the high end, super-recognisers who rarely make errors (Bobak, Hancock, & Bate, 2016; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016; Russell, Duchaine, & Nakayama, 2009), and at the low end, people with developmental prosopagnosia who rarely exceed chance performance (Behrmann & Avidan, 2005; Duchaine & Nakayama, 2005). Between these extremes, there is a spectrum of ability on standardised face identification tests (e.g. Burton, White, & McNeill, 2010; Duchaine & Nakayama, 2006).

These findings have led some researchers to suggest that personnel selection could play a useful role in optimising occupational face recognition (White et al., 2014). For example, Metropolitan Police super-recognisers have been found to score unusually high on a range of face identification tests (Robertson et al., 2016).

For mask detection, the cognitive situation is somewhat different. Here the challenge is not individuation at the subordinate level (Rosch et al., 1976), but rather categorisation at the basic level, albeit for the unusual case where one basic category (masks) deliberately mimics the other (faces). As the current task involves face/non-face categorisation, it arguably has more in common with face detection than with face identification (see Bindemann & Lewis, 2013, for a careful dissection of these issues).

The analogy with face detection may have some broad predictive value for the present case. Large individual differences in face detection ability have recently been reported (Robertson, Jenkins & Burton, 2017), and they appear to dissociate from face identification ability. However, one important difference is that face detection hinges on the presence or absence of a face-like pattern (e.g. two eyes above a nose above a mouth). That criterion will not help the viewer in the current task, as hyper-realistic face masks and real faces both present face-like

patterns. Thus, the intuition is that hyper-realistic mask detection will require finer discrimination than face detection tasks demand.

As yet, very little is known about individual differences in this finer perceptual task. For example, we do not know the expected range of ability. Nor do we know any factors that might differentiate high performers from low performers. The present studies address these issues by asking whether some people are better than others at categorising masks and faces, and what they may be doing that allows them to perform well. The overarching aim is to establish whether an individual differences approach might be as useful in hyper-realistic mask detection as it has been in face identification.

We begin in Experiment 6 by comparing detection of low-realism and high-realism masks in the context of real faces. In Experiment 7, we eliminated low-realism masks to focus participants on the harder comparison (high-realism masks vs real faces). Finally, we undertook an image analysis to compare use of information for high- and low- accuracy subgroups.

4.3 Experiment 6:

Discriminating high and low realism masks from real faces

Previous studies of hyper-realistic mask perception have assessed spontaneous detection of masks during an orthogonal task (social inference ratings; Experiment 1 and 2, see Appendix 2.3). Detection rates approached floor levels in that situation, precluding individual differences analysis. In this study, we sought to increase detection rates by (i) explicitly instructing participants that the task was to distinguish masks from real faces, (ii) presenting masks and faces equally often (50% prevalence), and (iii) explaining this prevalence rate to participants. These measures were intended to license ‘mask’ responses, even when participants were not certain. We expected that Low realism masks and real faces would be categorised accurately. Our main interest was in the range of performance for High realism masks

Method

Ethics statement. Ethics approval for all experiments was obtained from the departmental ethics committee at the University of York.

Participants. Thirty members of the volunteer panel at the University of York (21 female, 9 male; mean age = 22, age range 18–41 years) took part in exchange for a small payment or course credit.

Stimuli and Design. To collect images of high-realism masks, we entered the search terms ‘realistic masks’, ‘hyper-realistic masks’ and ‘realistic silicone masks’ into Google Images. We selected images that (i) exceeded 150 pixels in height, (ii) showed the mask in roughly frontal aspect, (iii) showed the eye region without occlusions, and (iv) included real hair eyebrows. We used the same criteria to search the websites of mask manufacturers (e.g. RealFlesh Masks, SPFX, CFX) and topical forums on social media (e.g. Silicone Mask Sickos, Silicone mask addicts). Our aim here was to sample ‘ambient’ photos of hyper-realistic masks that represent the range of the mask images in the visual world (Jenkins et al., 2011). For this reason, we avoided promotional studio photographs of the masks, and instead used photos of the masks *in situ*. This search resulted in 37 hyper-realistic mask images that met the inclusion criteria.

For comparison, we collected 37 images of low-realism masks by entering search terms such as ‘Halloween’, ‘party’, ‘mask’, ‘masquerade’, ‘face-mask’, and ‘party mask’ in Google Images, and selecting the first images that met inclusion criteria i-iii above.

We also collected 74 real-face images for use as fillers in the mask/face categorisation task. To ensure that the demographic distribution among our real face images was similar to that portrayed by the high-realism masks, we entered the search terms ‘young male’, ‘old male’, ‘young female’, and ‘old female’ into Google Images. We then accepted images that met criteria i-iii until the distribution of faces across these categories was the same as for the High-realism mask images. All photos were cropped to show the head region only and resized to 540 pixels high x 385 wide for presentation (see Figure 4.2).

The final image set consisted of 148 photographs (37 High Realism masks; 37 Low Realism masks; 74 Real Faces). Each participant viewed the 148 images intermixed in a different random order (within-subjects design).

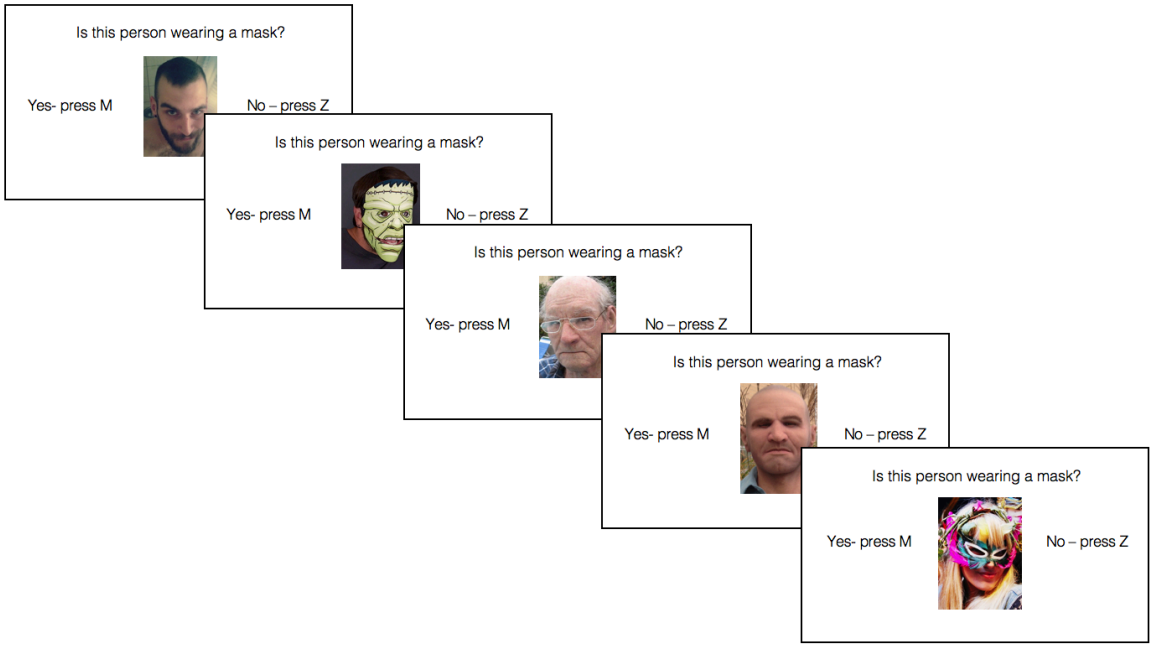

Figure 4.2. Example trials from Experiment 6. Correct responses: Z, M, M, M, M. See main text for details.

Procedure. Participants were instructed that half of the images showed real faces and half of the images showed masks. They were also informed that Mask trials would contain both Low-realism masks and High-realism masks. Each trial consisted of a centrally-presented image (a mask or a face) together with the prompt '*Is this person wearing a mask?'* and response options '*Yes - Press M*' and '*No – Press Z*'. The display remained on screen until response, upon which the following trial began automatically. No time limit was imposed. Participants completed 3 practice trials, followed by 148 experimental trials in a unique random order. The entire experiment took approximately 10 minutes to complete.

Results and Discussion

Group performance. Real face images were correctly classified on 96.3% of trials and were not analyzed further. Performance on mask trials is summarised in Figure 4.3. As expected, Low-realism masks were categorised reliably (M = 98.2%, SE = .4, CI = 97.6 – 99.0). High-realism masks were categorised much less reliably (M = 40.4%, SE = 5.6, CI = 29.2 – 51.5), meaning that the clear majority of these masks (59.6%) were misclassified as real faces. A within-subjects t-test confirmed that this difference in accuracy was statistically significant [t(29) = 10.29, p <.001].

Reaction time (RT) data followed a similar pattern. Correct responses to Low-realism mask trials were relatively fast (M = 895 msec, SE = 35, CI = 831 – 959). Indeed, RTs to High-realism masks were twice as long 1629 msec (SE = 142, CI = 1352 – 1901). Again, the difference between mask conditions was statistically robust [t(29) = 5.86, p <.001].

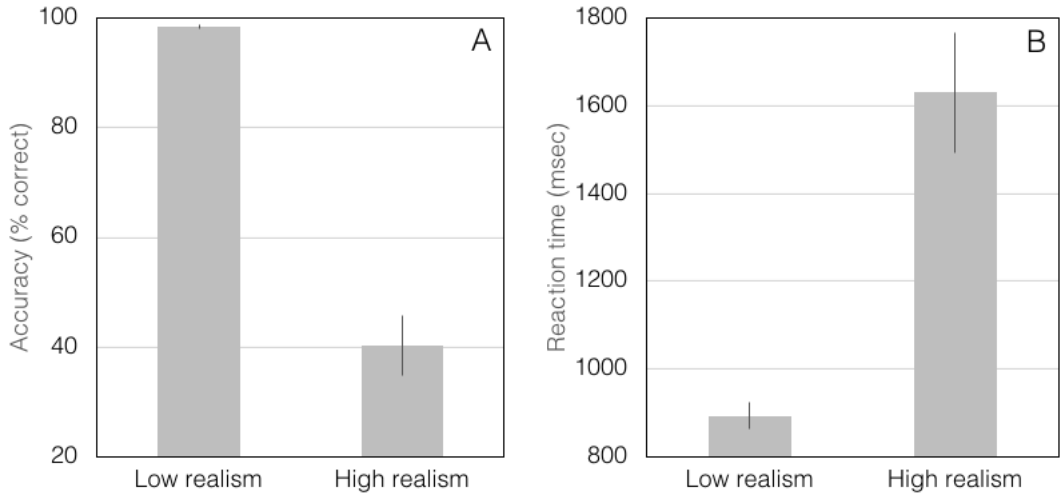

Figure 4.3. Mean accuracy rates (A) and correct reaction times (B) across participants as a function of mask condition in Experiment 6.

Individual differences. As can be seen in Figure 4.4, there was little

variability in accuracy in the Low-realism mask condition (range 95–100%), with performance compressed against ceiling for this easy task. In contrast, accuracy in the High-realism condition spanned the entire range (5–100%). Unsurprisingly, there was no correlation between High and Low realism mask trial performance ($r = .182$, $p = .335$).

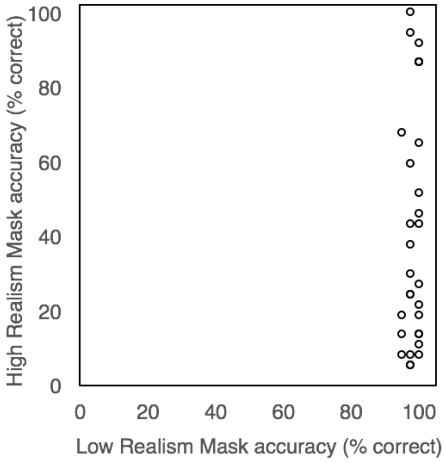

Figure 4.4. Scatterplot showing participants' mean categorisation accuracy rates in the High-realism and Low-realism mask conditions in Experiment 6.

Overall, classification judgements were much harder for High-realism masks than for Low-realism masks. More importantly for the current study, the data reveal striking individual differences in performance for the High-realism condition. A few observers detected hardly any hyper-realistic face masks in this experiment, but a few detected nearly all of them.

One possible interpretation of this pattern is that low-realism masks make high-realism masks hard to detect, by encouraging viewers to draw the category boundary in the wrong place ([*real faces* + *high-realism masks*] versus [*low-realism masks*], as opposed to [*real faces*] versus [*high-realism* + *low-realism masks*]). Prior knowledge of hyper-realistic face masks could protect against this error, leading to high overall accuracy. To address this possibility, we next repeated the experiment without the low-realism mask condition. We also asked participants whether they had encountered hyper-realistic face masks before the

experiment.

4.4 Experiment 7:

Discriminating high realism masks from real faces

This experiment was the same as Experiment 6, except for the following changes. First, we replaced the low-realism mask stimuli with high-realism mask stimuli, in order to focus participants on the difficult judgements (real faces versus hyper-realistic face masks). As before, we informed participants that half of the trials would contain real faces, and half of them would contain masks. We expected the new composition of trials to elicit errors in both directions (i.e. masks mistaken for faces and faces mistaken for masks). Our main interest was the distribution of performance in this situation. To test for effects of prior mask knowledge on performance, we also collected self-report ratings at the end of the experiment.

Method

Participants. Thirty members of the volunteer panel at the University of York (24 female, 6 male; mean age = 20, age range 18–24 years) took part in exchange for a small payment or course credit.

Stimuli and Design. Additional stimuli were collected via internet search, using the method described in Experiment 6. Once again, the proportions of young male, old male, young female, and old female items were matched across Real face and High-realism mask images. The final image set consisted of 148 photographs (74 High-realism masks and 74 Real faces). Each participant viewed the 148 images intermixed in a different random order (within-subjects design).

Procedure. The procedure was the same as for Experiment 6, except that the low-realism trials were replaced with high-realism trials (see Figure 4.5). To test whether individual differences in performance could be explained by prior knowledge of hyper-realistic face masks, we asked participants to rate their prior

knowledge on a 7-point Likert scale at the end of the experiment.

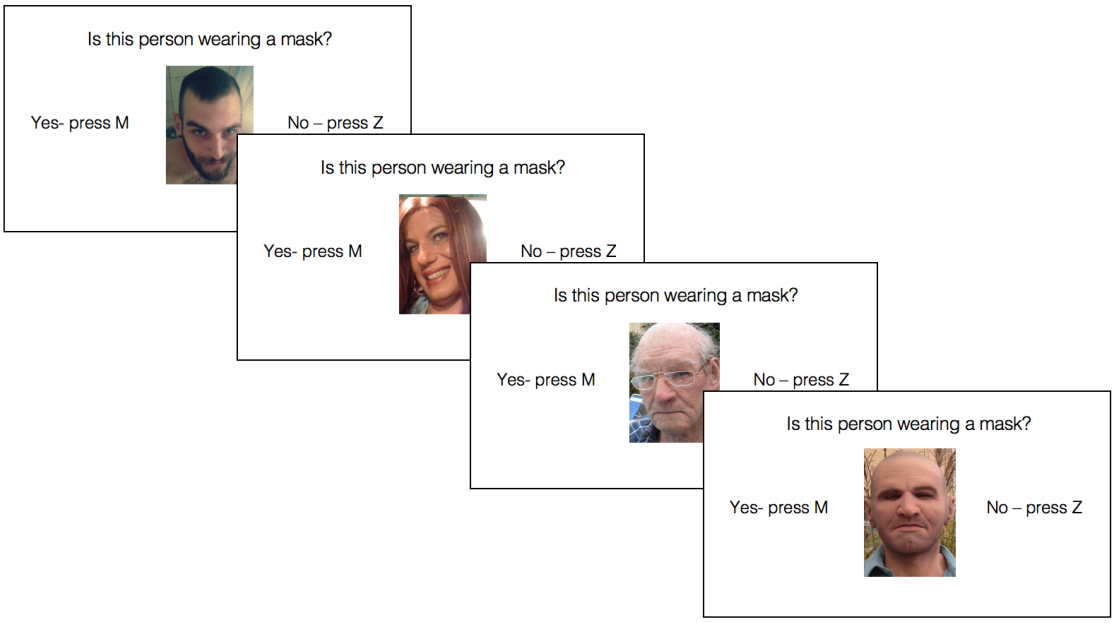

Figure 4.5. Example trials from Experiment 7. Correct responses: Z, Z, M, M.

Results and Discussion

Group performance. Overall categorisation performance is summarised in Figure 4.6. As can be seen from the figure, classification of Real face images was accurate, but not at ceiling (M = 91.2%, SE = 2.0, CI = 87.3 – 95.1). Accuracy for High-realism masks was relatively low (M = 73.7%, SE = 2.7, CI = 68.3– 79.0), indicating that hyper-realistic masks were frequently misclassified as real faces (26.3%). A within-subjects t-test confirmed that this difference in classification accuracy was statistically significant [t(29) = 6.78, p <.001].

There was no significant difference in reaction times between Real face (M = 1301 msec, SE = 93, CI = 1121 – 1480) and High-realism mask trials (M = 1283 msec, SE = 71, CI = 1145 – 1421); [t (29) = .34, p =.730].

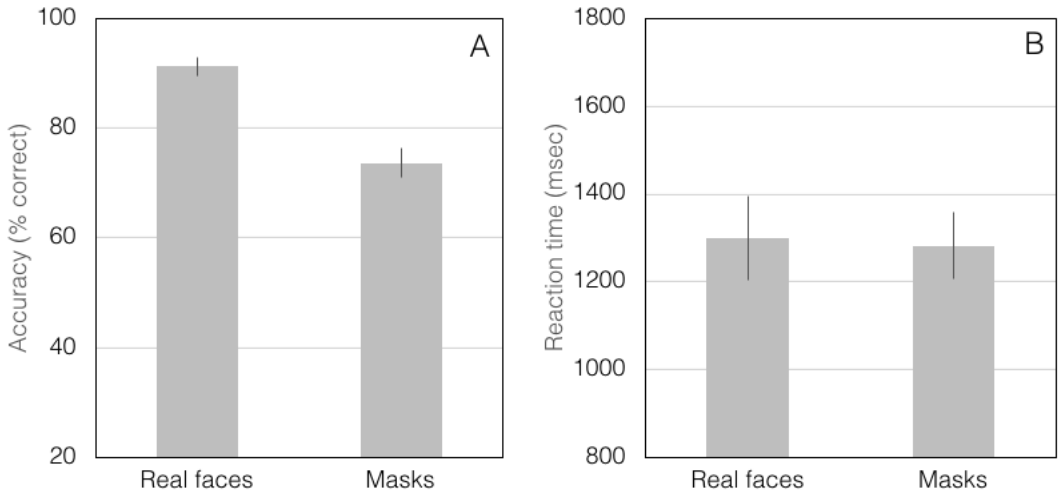

Figure 4.6. Mean accuracy rates (A) and correct reaction times (B) across participants as a function of experimental condition in Experiment 7.

Individual differences. As can be seen in Figure 4.7, almost everyone performed above chance in both conditions. Classification accuracy ranged from 65–100% in the Real face condition, and 43–91% in the High-realism mask condition. Interestingly, there was no correlation in performance between the two conditions [r = -.04, p = .830].

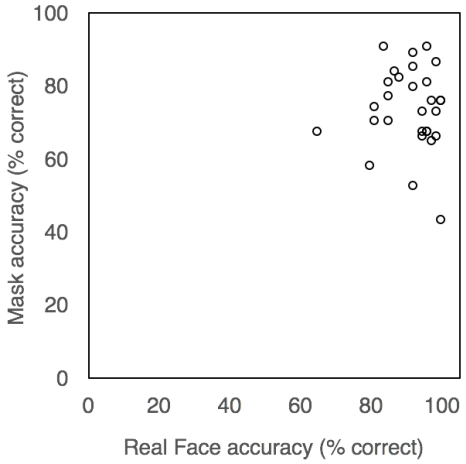

Figure 4.7. Scatterplot showing participants' mean categorisation accuracy rates in the Real face and High-realism mask conditions in Experiment 7.

Prior mask knowledge. Self-report ratings of prior mask knowledge were generally low (M = 2.67, SD = 1.03), suggesting little or no exposure to hyper-realistic face masks before the experiment. More importantly, there was no significant correlation between prior mask knowledge and performance in either the High-realism mask condition [r = .025, p = .898] or the Real face condition [r = .319, p = .092] (see Figure 4.8).

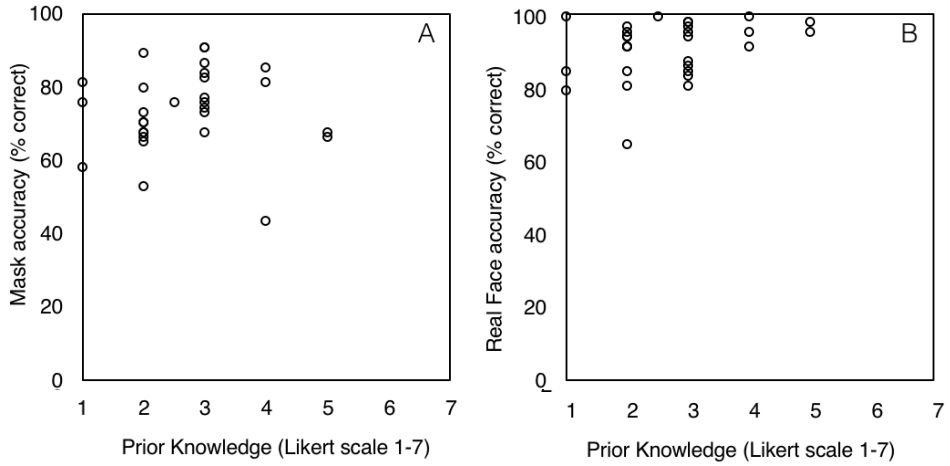

Figure 4.8. Scatterplots showing (A) accuracy for High-realism masks, (B) accuracy for Real faces by prior mask knowledge in Experiment 7.

Overall error rates were high (20%) despite the simplicity of the task, and despite the fact that participants were informed about the prevalence of mask and real face trials. We note that error rates were somewhat higher in the mask condition (30%) than in the real face condition (10%), meaning that overall, masks were mistaken for faces more often than faces were mistaken for masks. Interestingly, some participants were highly accurate in correctly categorising the masks. However, accuracy in the mask condition was not explained by accuracy in the real face condition, nor by prior exposure to hyper-realistic face masks. In the final study, we ask whether high-performing individuals are using specific visual cues to support their accurate judgements.

4.5 Image Analysis

The purpose of the image analysis was to compare the use of visual information by high classification accuracy and low classification accuracy participants in Experiment 7. Our specific interests were (i) the availability of visual cues—that is, whether mask and face images differed reliably, (ii) the nature of any reliable visual cues—specifically, their spatial location, and (iii) whether high-performing and low-performing participants made different use of these cues. We addressed these issues by using categorisation data from Experiment 7.

The logic of this image analysis is as follows. The appearance of the mask stimuli and the face stimuli can be summarised by generating an average image for each stimulus category (an average mask and an average face). Systematic differences between these two categories can then be visualised by subtracting the average face from the average mask to create a difference image. This difference image indicates which regions of the stimulus are most informative for mask/face classification. Our hypothesis is that high-performing participants tracked this information more closely than low-performing participants. To test this hypothesis, we used categorisation responses from Experiment 7 to generate difference images for the high-performing and low-performing subgroups. This allowed us to compare the *perceptual* difference images (based on participants' categorisation of the stimuli) against the *physical* difference image (based on the actual stimulus categories). By undertaking this comparison for different slices of the image, we were able to quantify participants' tracking of category-level regularities across different face regions.

Method

Participant subgroups. To establish a strong manipulation of the independent variable (categorisation accuracy for masks), we divided participants into performance quintiles (N = 6 per subgroup) and contrasted the highest and lowest quintiles. A 2 x 2 mixed ANOVA with the within-subject factor Image Type

(Mask, Real face), and the between-subjects factor of Subgroup (High, Low) confirmed that these subgroups were statistically distinct with respect to their classification scores. Consistent with the whole-group analysis, we found a significant main effect of Image Type, with higher accuracy for Real face trials (M = 90.0%, SE = 1.4, CI = 83.6 – 95.7) than for Mask trials (M = 72.5%, SE = 1.5, CI 65.9 – 79.1), [$F(1,10) = 13.76$, $p = .004$, partial $\eta^2 = .58$]. More importantly, there was also a significant main effect of Subgroup, with the High accuracy group (M = 90.2%, SE = .8, CI = 86.7 – 93.8) reliably outperforming the Low accuracy group (M = 72.1%, SE = 2.1, CI = 62.9 – 81.2), [$F(1,10) = 85.44$, $p < .001$, partial $\eta^2 = .89$]. There was no significant interaction between these factors [$F(1,10) = 1.78$, $p = .212$].

Face averages. We next constructed six average images (Burton et al., 2005) from the following six image sets: (i) actual masks (N = 74), (ii) actual faces (N = 74), (iii) perceived masks for High performers, (iv) perceived faces for High performers, (v) perceived masks for Low performers, and (vi) perceived faces for Low performers (weighted averages of images as classified; N > 50 for all). Seven images (5 masks, 2 real faces) were excluded from this analysis because the camera angle did not allow accurate landmarking of the photographs (see Kramer, Jenkins, & Burton, 2017 for implementation details). The six weighted texture averages for the remaining images are shown in Figure 4.9.

Difference images. To ask what distinguishes masks from real faces, we next computed a difference image (average mask minus average face) separately for the veridical categories, the High performance group, and the Low performance group. These three difference images are shown in Figure 4.9 (lighter regions indicate greater difference). The veridical difference image (Figure 4.9, center) indicates that the surrounding of the eye is especially informative, presumably because the eye holes in the mask can produce local anomalies in appearance (e.g. surface discontinuities if the mask is not flush with the wearer's face; complexion discontinuities if the skin around the wearer's eyes is exposed). The question is whether observers pick up on these subtle cues. Visual comparison confirms that the difference image for the High performer group (Figure 4.9, left) closely resembles the veridical difference image (Figure 4.9,

center). The difference image for the Low performer group (Figure 4.9, right) resembles the veridical difference image less closely. This global pattern is perhaps to be expected, given the formation of the subgroups: if high performers didn't track the veridical categories, they would not be high performers. However, local variations in this pattern may reveal specific cues that high performers exploit, and that low performers overlook. We investigated this possibility by comparing correlations between different image slices.

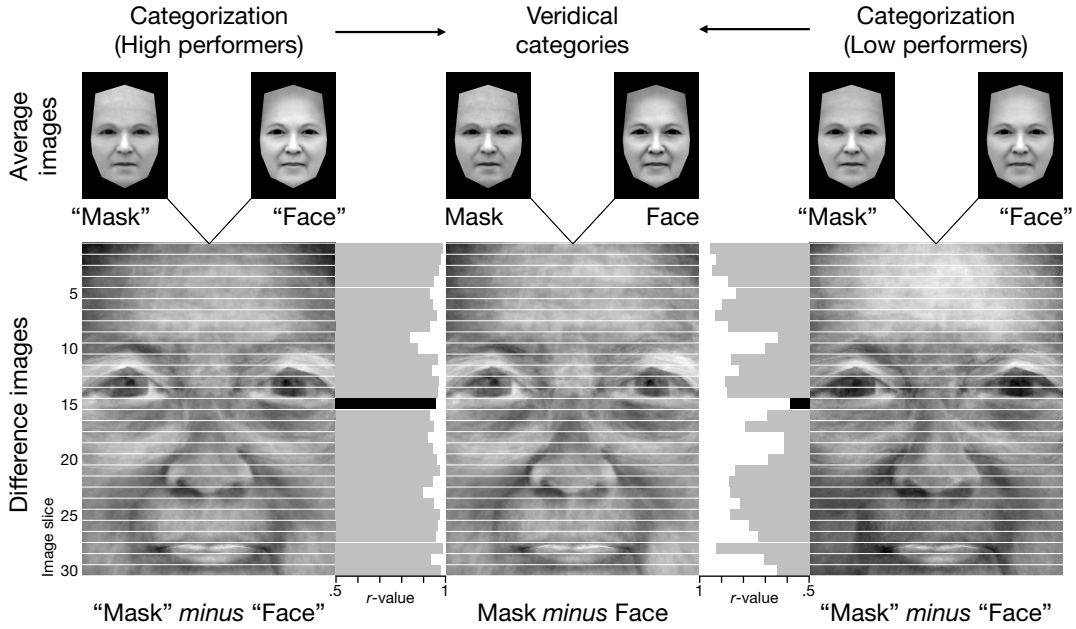

Figure 4.9. Summary of image analysis. Average images show mean pixel intensities across images in each category, separately for High performers (Left), Low performers (Right), and veridical categories (Centre). Difference images are subtractions of pixel intensity (Mask minus Face; rescaled for visualisation). Lighter colours indicate larger differences. Note the light region around the eye in the veridical difference image. The y-axis shows 30 horizontal image slices. Correlations between difference images (grey bars) are shown for each image slice. The largest discrepancy between High and Low performers is shown at Slice 15 (black bars). High performers closely tracked categorical differences in this region. Low performers did not.

Image correlations. To avoid spurious inflation of correlation values by black background pixels, we first cropped the background from each difference

image to create rectangular face image (300 pixels high x 228 pixels wide) that retained all of the internal features. To allow direct comparison across equally sized regions, we then divided each rectangular image into 30 horizontal slices (10 pixels high x 228 pixels wide; see Figure 4.9). Successive rows of pixels can be concatenated to form a single vector of pixels for each slice (1 pixel high x 2280 pixels wide), in which the greyscale intensity of each pixel is specified by an integer value between 0 (black) and 255 (white). These intensity values formed the input to the correlation analysis.

Figure 4.9 shows the results of these image correlations (r values), separately for each slice. As can be seen from the figure, correlations between the veridical difference image and the High-performer image are consistently high across image slices (range .87–.99). The correlations between the veridical difference image and the Low-performer image are lower overall and much more variable (range .59–.95). Most strikingly, there is a distinct notch in correlation values between the Low-performer and veridical difference images, directly under the eyes (image slice 15; r = .59). In fact, this was the lowest correlation in entire analysis. Importantly, that notch does not appear in the correlations between the High performer and veridical difference images (image slice 15; r = .95).

To summarise, our comparison of mask and face images suggests that the eye surround is the most informative region for separating these two categories. High performers appear to use information below the eye in a way that low performers do not. What information could be in this region? We suggest two possibilities. First, in a real face, the region below the eyes normally includes the lower eyelashes—an area of high local contrast. The masks in our stimulus set do not include eyelashes. If the mask covers the wearer's eyelashes, it will typically reduce local contrast. Reduced local contrast under the eye may be a cue to mask detection. Second, in a real face, skin complexion below the eyes normally changes gradually on a local scale. The masks in our stimulus set do not necessarily match the complexion of the wearer. If the mask exposes any skin below the wearer's eyes, it may cause an apparent discontinuity in skin colouration. Discontinuity in complexion under the eye may be a cue to mask detection. Each of these possibilities suggests that the precise fit of the mask

around the wearer's eyes is critical. Shade from the brow will tend to conceal cues in the upper eye region, at least under normal illumination conditions (light source above). However, the same illumination conditions will tend to highlight cues in the lower eye region, making them more salient.

4.6 General Discussion

Across three studies, we investigated individual differences in hyper-realistic mask detection—specifically, the ability to categorise images as masks or real faces. In Experiment 6, we found large individual differences in a mask/face categorisation task for High-realism masks, Low-realism masks, and Real faces. Although Low-realism masks (and real faces) were categorised accurately overall (>98% correct), High-realism masks were not (40% correct). More importantly from an individual differences perspective, accuracy in the High-realism condition ranged from floor (5%) to ceiling (100%), despite the consistently high accuracy for other stimulus types.

In Experiment 7, we discarded the Low-realism mask condition to focus exclusively on the difficult categorisation—hyper-realistic masks versus real faces. Perhaps surprisingly, removing the easy condition improved performance in the difficult condition considerably (74% correct). This seemingly paradoxical result underscores the importance of the context in which a categorisation decision is taken. The absence of an obvious category distinction (cf. Experiment 6), combined with information about the distribution of stimuli, presumably led participants in Experiment 7 to approach the task differently. Nevertheless, we still observed a wide range of performance, even in this very different cognitive situation. Accuracy ranged from near chance (43%) to near ceiling (91%). Interestingly, accuracy in the Real face condition was also varied (65–100%). However, performance in these two conditions was uncorrelated, and was not explained by previous exposure to hyper-realistic face masks.

Both of these experiments revealed large individual differences in hyper-realistic mask detection, in the sense that some people were much more accurate

than others at categorising masks and real faces. These findings suggest that stable differences in ability may be worth pursuing. It is too early to say whether some individuals exhibit a special talent for this task. Conclusive evidence would require estimates of test-retest reliability and consistently high performance across a range of tasks (Robertson et al., 2016; Russell, Duchaine, & Nakayama, 2009;). Until then, we suggest another possible route to improved detection rates—one that does not depend on screening for high-aptitude individuals. In our image analysis, we asked what high-performers are doing that low-performers are not. This analysis revealed a candidate visual cue that these subgroups used differently—the area under the eyes. Hyper-realistic mask images and real face images diverged more strongly in this area than in other areas. Moreover, high-performers and low-performers diverged strongly in the extent to which the area under the eyes predicted their responses. This intriguing finding raises the question of whether mask detection could be improved by drawing attention to this region. If so, it could pave the way for a simple training intervention. This is a tantalising prospect, especially as benefits of training in face identification tasks have proven difficult to pin down (Towler, White, & Kemp, 2014, 2017; White, Kemp, Jenkins, & Burton, 2014). Eye-tracking data in combination with accuracy rates, before and after training, should elucidate the potential of this approach.

Finally, it is worth returning to the somewhat artificial nature of this task. The experiment was specifically contrived to encourage detection of hyper-realistic masks. For example, we focused on masks in the task instructions, and spelled out the distribution of mask and face stimuli. In view of this strong framing, the detection rate for these masks seems rather low. Nevertheless, it almost certainly overestimates the rate of spontaneous detection when a mask framing is absent. In Chapter 1 we reported extremely low rates of spontaneous detection, both for photographic presentations in the lab (Experiment 1 and 2) and live viewing of mask wearers outdoors (Experiment 3). On the other hand, none of these studies has measured detection during active social interaction with the mask wearer (e.g. conversation). We expect that, in a more interactive context, additional cues from speech and movement could increase detection rate, but that is a matter for future studies.

We do provide a more accurate estimate of a single item mask inspection than the Turing test in Chapter 3 allowed (19% error rate). With Low realism distractors, single item inspection increases errors to 60%. In such real-world decisions real-world decision (Stanton, 2015; Weisman, 2018), the only hope is that it is more akin to a faces/masks discrimination with little minimal distraction such as tested in Experiment 7, where error remains at approximately 25%, even in this more complicated task.

For now, we show that distinguishing hyper-realistic masks from real faces is a difficult task. Some people are much better than others at picking out hyper-realistic masks, and these large individual differences are not readily explained by correct categorisation of real faces, or by prior exposure to hyper-realistic masks. We suggest that they may be explained by differential use of specific visual cues, and identify the region under the eyes as a promising candidate.

Chapter 5.

Demographic profiling through a hyper-realistic face mask

5.1 Summary

In a number of criminal cases, perpetrators have used hyper-realistic silicone masks to transform their appearance. When it becomes apparent that such a mask has been used, investigators may try to use demographic profiling of the person underneath the mask as a first step towards face identification. For example, the discovery that the FBI listed bank robber known as the Geezer Bandit was wearing an 'old man' mask led investigators to estimate the perpetrator's age and gender through the mask. To assess the accuracy of such estimates, we asked participants to guess the age, gender, and racial group of confederates wearing hyper-realistic face masks under live viewing conditions. Error rates were high for each of these demographic traits. In addition, the data seem to show a systematic bias in which demographic characteristics of the mask were attributed to the wearer. These findings are discussed in terms of the fundamental attribution error, and suggest that inferences about the wearer through a mask should be treated with great caution.

5.2 Introduction

Hyper realistic masks (see Figure 5.1) are an extreme form of deliberate disguise, available to buy online without restriction (e.g. www.realfleshmasks.com, www.spfxmasks.com, www.immortalmasks.com). In recent years, these masks have been used in criminal settings, where they successfully passed for real faces (Zamost, 2010; Winer, 2017). Our research shows that this is not an anomaly. In a variety of live viewing conditions these masks were missed 99% of the time if unprompted (Chapter 1) and only barely distinguished from real faces under

photographic conditions (Chapter 2-3). Such masks allow the culprit to avoid being identified through face recognition (just like regular masks), but they also divert attention to a completely different target demographic (unlike regular masks, see figure 5.1).

Figure 5.1. All images display Dr Rob Jenkins with the same facial expression. All photographs were taken on the same day.

An example of such a case is that of the 'Geezer Bandit'. The 'Geezer Bandit' is thought to have robbed over 16 banks between 2009-2011 and is still on the FBI's most wanted list in 2018. He owes his nickname to his appearance, with the FBI describing him as a '60-70 year old white male' (see Figure 5.2). It took over one year, and the help of a producer of realistic masks for the FBI to realise that this was not at all the culprit's appearance, but that he/she was rather wearing a hyper-realistic face mask of those demographics. With this knowledge, the FBI adjusted their demographic estimates of the wearer to that of a 30-50 year old male (Weisman, 2018). This leads to the following question: once it is established that a realistic mask is being used, can we tell anything about the wearer beneath the mask?

WANTED
BY THE FBI

"Geezer Bandit" San Diego, California

UNKNOWN BANK ROBBER

Alias:
"Geezer Bandit"

DESCRIPTION

Age:	60 to 70 years old	**Hair:**	Unknown
Height:	Approximately 6'0"	**Eyes:**	Unknown
Weight:	Approximately 190 pounds	**Sex:**	Male
Build:	Average	**Race:**	White

CAUTION

An unknown male robber is suspected of committing at least 11 bank robberies around the San Diego, California, area since August of 2009. Typically, the man enters the bank, approaches the teller, and presents a demand note for cash. He carries a small caliber pistol that he threatens to use if the teller does not comply with his demands. After receiving the money, he walks out of the bank.

SHOULD BE CONSIDERED ARMED AND DANGEROUS

If you have any information concerning this person, please contact your local FBI office or the nearest American Embassy or Consulate.

Figure 5.2. A wanted poster for the 'Geezer Bandit', issued by the FBI in 2010. Image retrieved from: https://bit.ly/2Lzlws8

Testing performance accuracy for demographic estimates of a disguised face is useful for indicating the reliability of such attempts in the real world. Demographic estimation is a common process, as it is the first step toward identifying the culprit. To take US bank robberies as an example, a 2007 guide on US bank robberies for police (Wiesel, 2007) notes that more than half of bank crimes use facial disguise. 2017 statistics on bank robberies also state that racial group and sex of the culprit was undetermined in only .4% of cases (Department of Justice, 2017). These statistics suggest that approximately 50% of cases are being profiled in spite of their facial disguise.

This prevalence is problematic. Shapiro and Penrod's (1968) meta-analysis identified facial disguise as one of the key factors to reduce correct identification and increase incorrect identification in situations which rely on eyewitness testimonies (see Olsen & Wells, 2003 for a more recent review). The few small-scale studies that assess identification accuracy of people wearing simple deliberate disguise experimentally confirm this issue. Specifically, Dhemacha et al. (2014) examined face matching performance and Terry (1994) and Righi, Peissig & Tarr (2012) examined face memory performance for faces covered with (sun)glasses and/or wigs. They consistently found that more errors were made for disguised than non-disguised faces. Terry (1994) attributed this reduction in performance 1) to facial disguise obstructing facial information useful for identification and 2) to facial disguise being encoded as part of the wearer's identity even when it is recognisably *not* part of the face. In the case of hyper-realistic face masks, the latter point is particularly interesting.

In some ways, this overextension of disguise properties to person properties is reminiscent of the *Fundamental Attribution Error* (FAE; Jones & Harris, 1967). In this research participants were asked to guess a typical fellow student's true attitude towards Fidel Castro from reading an essay they had supposedly written. Participants were made firmly aware that the writer had been randomly assigned to write a Pro-Castro essay, but results nevertheless showed that estimates of the writer's attitude swayed toward their randomly assigned category**.** The authors termed this tendency to explain perceived behaviour through dispositional factors rather than situational factors as the FAE (Jones & Harris, 1967; Ross, 1977). The FAE is the most famous effect amongst a family of biases brought forth by studying Attribution Theory (Kelley, 1967), including the correspondence bias, self-serving bias, actor-observer bias, culture bias (Kruglanski & Ajzen, 1983). Attribution Theory broadly studies "how the social perceiver uses information to arrive at causal explanations for events" and "how it is combined to form a causal judgment" (Fiske & Taylor, 1991).

There has been much debate on where the FAE fits amidst the other biases, variously described as either a 'mother' bias or not (Harvey, Town and Yarkin, 1981; Langdridge & Butt, 2004; Sabini, Siepmann & Stein, 2001a and

commentaries: Sabini, Siepmann & Stein, 2001b). Complexity arose when research showed that both situational and dispositional factors to behaviour can be overestimated, depending on the manipulation (Miller, Smith & Uleman, 1981) and how it was characterised (Harvey, Town and Yarkin, 1981; Sabini, Siepmann & Stein, 2001a, b; Solomon, 1978). Some have dissociated the effects as being Dispositional and Situational Attribution Errors (e.g. Solomon, 1978), others have simply used the term Attribution bias (Harvey, Town and Yarkin, 1981). Despite the confusing terminology, the effects are highly replicable (Malle, 2006) and are considered to be an integral component to social integration, with consequences for many principles of social behaviour including achievement and self-efficacy (Chassin et al., 1990; Kok, Den Boer, De Vries, 2014; Mabe & West, 1982), altruism (e.g. Learning, 2003), and blame (e.g. Roesh & Wiener, 2001; Shaver, 2012). From here on out we focus on the subset of the FAE which concerns the overemphasis of external factors, which are known to be non-descriptive to the person.

Terry's (1994) study suggested that there may be such an FAE in face perception, such that covering the face does more than obstruct key memory encoding information; it also alters the perception of that person's identity. This seems somewhat harmless in case of glasses or wigs, as the extent of the effect of non-human artifacts on person perception do no more than increase error rates, as they do not carry inherent demographic or characteristic information which could bias the viewer's perception of the wearer in any particular direction.

Hyper-realistic masks do carry such information. Unlike a balaclava or a pair of sunglasses to cover the face, these masks depict clear demographic and identity signals, which if attributed to the wearer, may influence profiling. Hence if we observe an FAE in face perception, the appearance of the hyper-realistic mask could sway the viewer's perception of the wearer beneath the mask, even if the mask is recognisably not a part of the face. For example, people might be more likely to think that the Geezer Bandit's *Grandpa* mask, which resembles an old white male, is being worn by an older white male, despite knowing that there is no inherent relationship between the mask and the wearer.

Generally, studies investigating the effect of disguise test for identification performance, but as hyper-realistic masks contain more complex signals than regular disguises (such as sunglasses and/or wigs) and obstruct most of the face, we start in with a much simpler task.

5.3. Experiment 8:

Live demographic estimates of wearer beneath a mask

In this experiment, we sought to establish whether viewers would be able to guess basic demographic details (age, gender and racial group) of the wearer underneath the mask. As a baseline measure all wearers were rated without the mask (No mask condition). These same wearers were then rated in a mask (Mask condition), with and without their body being visible (Mask + Body condition), to simulate a variety of real-world identification situations, particularly as a visible body could provide demographic cues that conflict with those from the mask.

We expected near perfect performance in the No mask condition, with performance degrading as more visual information is obstructed. As body form and posture contain cues to gender (Hall, 1978; Mather & Murdoch, 1994) and age (Farage et al., 2013) we expect largest deterioration in performance when the body and face are not visible for these two estimates. For racial group, bodily cues are minimal between the two tested confederate groups (Asian and Western), hence we expect facial cues to play a much more important role.

Method

Ethics statement. Ethics approval this experiment was obtained from the departmental ethics committee at the University of York and Kyoto University

Participants. Four hundred volunteers participated in the study. Two-hundred-twenty-two participants (164 males; mean age = 27, age range 17–80 years) were attendees of the Bradford Media Museum Open Day (N=119) and

Open Days for prospective students (N=102) at the University of York, UK and one-hundred-seventy-seven (106 males; mean age = 27, age range 18–80 years) were students at Kyoto University, Japan.

Stimuli and Design. Four confederates in the UK and four in Japan (counterbalanced for gender and racial group in Japan and the UK; see Figure 5.3) were briefed on the aims of the study.

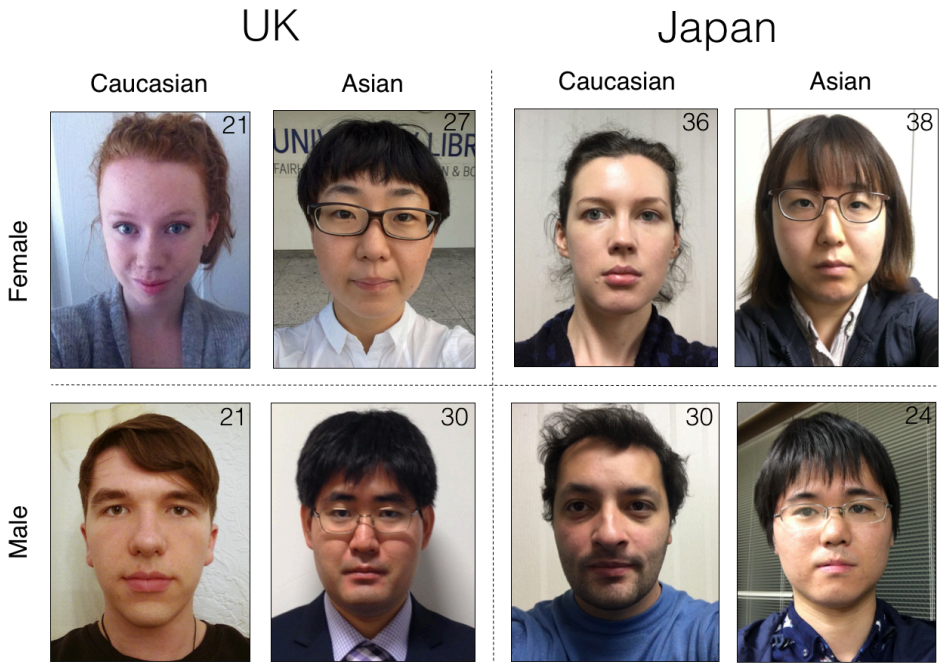

Figure 5.3. Counterbalancing of mask wearers by gender and racial group in two locations. Numbers in top right corner of each photograph denote ages of wearers at time of experiment. Dotted lines separate confederate pairs.

All confederates wore jeans and a loose-fitting blue t-shirt to ensure that clothing would not bias participant guesses. For the *Mask* and *Mask + Body* conditions, we used two different masks produced by Realflesh Masks, Quebec, Canada: *The Pensioner* (Western Old Male Mask) and *The Asian* (Asian Old Male Mask). We ordered punched human hair eyebrows on both masks, and a goatee beard and horseshoe hair on *The Asian* (Figure 5.4).

Figure 5.4. Hyper realistic face mask ‘The Asian’ (left) and ‘The Pensioner’ (right).

All wearers wore a mask with a loose fitted blue t-shirt to cover the shoulders. For the *Mask* condition the confederate was stood behind a poster board, which occluded the body from the shoulders down. For the *Mask* + *Body* condition the poster board was removed so that the participant could see the masked confederate entirely. All participants viewed two confederates: one confederate in the *No mask* condition, and one confederate in the *Mask* condition followed by the *Mask* + *Body* condition (see Figure 5.5). Each of the four confederate pairs rotated through four conditions: 1) confederate A wore the Asian mask whilst confederate B was in the No mask condition; 2) confederate A was in the No mask condition + confederate B wore the Asian mask; 3) confederate A wore the Western mask + confederate B was in the No mask condition; 4) confederate A was in the No mask condition + confederate B wore the Western mask.

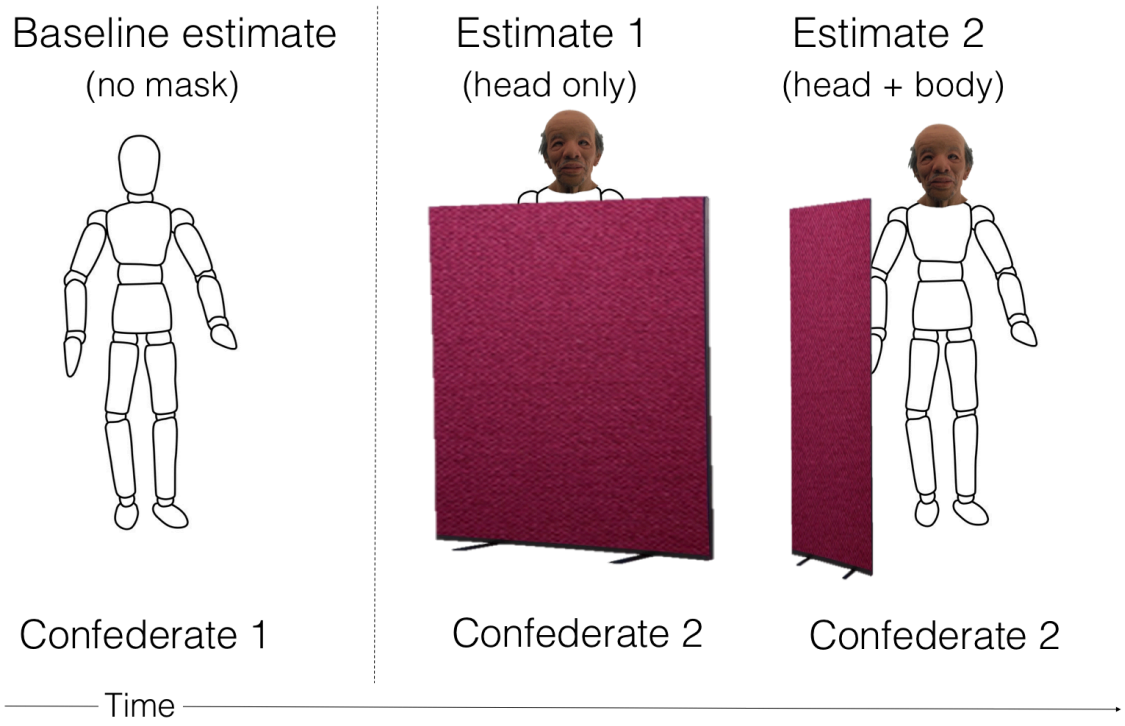

Figure 5.5. Schematic of participant experience over time.

To manipulate Age estimation accuracy of the person beneath the mask we created three different conditions of Model visibility (IV1). To manipulate gender estimation accuracy (DV2), we compared across the same conditions of Model visibility (IV1) and compared male to female wearers (IV2). Finally, to manipulate racial group estimation accuracy we compared across three conditions of Model visibility (IV1), compared wearers (IV2) and masks (IV3) of Asian and Western appearance, and ran the experiment in the UK and in Japan (IV4). Each participant saw all three condition of Model visibility; one confederate without the mask and another in one of the *Mask* and *Mask + Body condition* (within subjects), whereas all other Model visibility conditions were between subject. This resulted in a total of 3 conditions for the Age estimate, 6 conditions for the Gender estimate (3 Model visibility conditions x 2 confederate genders) and 20 conditions for the Racial group estimate (3 Model visibility conditions x 2 confederates x 2 mask types x 2 testing locations) in total.

Procedure. Testing took place at a Bradford Media Museum Late night event, prospective student open days at the University of York, and by the entrance of a university shop at Kyoto University. For the duration of the testing session, two confederates - one standing behind a poster board wearing a realistic face mask, one wearing a lab coat in front of the poster board – remained in a location with reliable foot traffic. An experimenter recruited viewers at approximately 1-2 m distance from the masked confederate to individual passers-by. The participant was told that the confederate behind the poster board was wearing a hyper-realistic face mask, and that their task was to guess who was underneath the mask. To illustrate to the participant how easy this task is without the mask, the participant was first asked to guess the age (open response), gender (male or female response options), and racial group (Asian or Western response options) of the assistant without a mask (control condition). Next the participant was asked to guess the same demographics for the person underneath the mask, whom the viewer could only see from the shoulders up (Mask condition). Next participants were asked to view the masked confederate through a gap between the poster boards – allowing them full view of the confederate's body along with their masked face (Mask + body condition) and were asked to make a final guess of the same three demographics. After responding, the participant was also asked to provide some demographic information about themselves (age, gender, racial group) and rate how realistic they thought the mask was on a Likert scale of 1-7. Responses were written down by the participants, using prepared response sheets. The entire procedure lasted approximately two minutes for each participant.

Results

We will address the results separately by dependent variables Age, Gender and Racial group. In each section we will break down the results by the independent variables.

Descriptives

Table 5.1 summarises the distribution of participants across conditions.

Test location	Wearer gender	Wearer race	Control condition	Masked Conditions	
				Asian mask	Western mask
Japan	Male	Asian	46	22	24
		Western	46	23	23
	Female	Asian	44	21	21
		Western	42	22	22
UK	Male	Asian	49	23	30
		Western	53	25	24
	Female	Asian	42	39	39
		Western	78	22	20

Table 5.1. Number of participants tested in each of the Model visibility conditions shown separately for testing in UK and Japan. The grey/white colour coding in the Control condition corresponds to the within subject data collection in the Masked conditions in the same colour.

Age

Effects of Mask and Model visibility. Average Age estimates (mean deviation in years from confederate real age) were submitted to a One-Way Repeated Measures ANOVA by Model visibility. Results of this analysis are summarised in Figure 5.6.

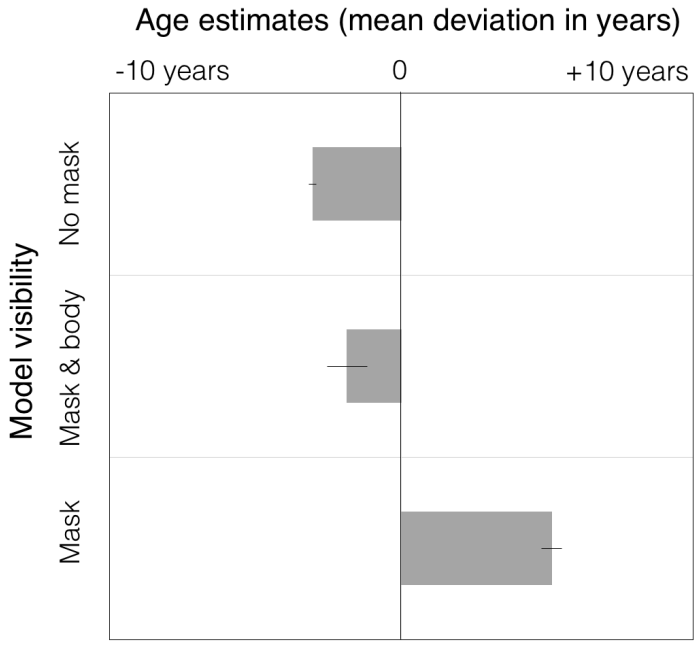

Figure 5.6. Age estimates – expressed as deviation from the confederate's real age – by Model visibility condition. Error bars display standard error.

As expected, there was a significant main effect of Model visibility [(F (2,782) = 81.90, p <.001, partial η^2 = .63]. Turkey's HSD post-hoc tests confirm that the Age estimate in the Mask condition (M = 5.2 years, SE = .79, CI = 3.6 – 6.7) is significantly higher than the Mask + Body condition (M = -2.2 years, SE = .54, CI = -3.1 – -1.3) [Mean difference = M = -7.3, SE = .70, CI = -8.70 – -5.93, p <.001] and the No mask condition (M = -2.6 years, SE = .45, CI = -3.0 – -2.2, [Mean Difference = -7.75, SE = .81, CI = -9.3 – -6.2, p < .001]), but that there is no difference between the No mask and Mask + Body conditions [Mean Difference = .44, SE = .49, CI = -.53 – 1.4, p = .377).

Gender

Gender guesses were submitted to two separate analyses to compare performance accuracy across Model visibility (within subject) and across wearer

gender (between subject), with a Bonferroni correction for multiple comparisons (p = .025). Results of this analysis are summarised in Figure 5.7.

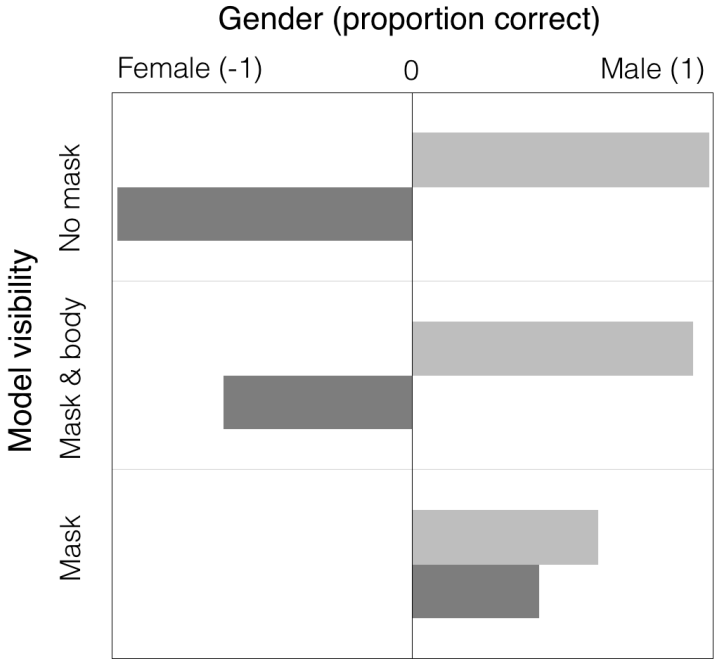

Figure 5.7. Proportion of correct Gender guesses of the wearer beneath the mask by the gender's wearer and Model visibility condition.

Effects of Model visibility. To determine whether there was a difference in performance accuracy under different Model visibility conditions (No mask: 98.5% correct; Mask + Body: 95.5% correct; Mask: 73.3% correct), we submitted Correct and Incorrect Guesses to a Cochran Q Test for comparing more than 2 related categorical samples - in this case for Model visibility, and corrected for multiple comparisons (p = .025). As expected, there was a significant difference between Conditions [Cochran's Q (2) = 217.2, p <.001]. To determine where the differences lay, we did pairwise comparisons using Exact McNemar tests with further Bonferroni correction (p = .008) which indicated significantly higher accuracy in the No mask vs. Mask + Body condition [X^2 (1, 395) = 29.3, p <.001], Mask + Body vs. Mask condition [X^2 (1, 396) = 101.1, p <.001] and No mask vs. Mask condition [X^2 (1, 395) = 132.1 p <.001].

Effects of Mask. To determine whether there was an attribution error of the mask (male) to the wearer (Male and Female), we established a percentage correct performance (averaged across the three Model visibility conditions) and submitted this to an independent sample t-test by Wearer Gender with a Bonferroni correction (p = .025). We compared the percentage of correct responses when the mask wearer was Female versus when the wearer was Male. As the mask was always male, if the gender of the mask is attributed to the gender of the wearer, it should result in a bias towards Male responses.

As expected, participants have a significantly higher overall score across the conditions when the wearer is Male (M = 93.3%, SE = 1.5) than when the wearer is Female (M = 70%, SE = 1.3) [Mean Difference = 21.3%, SE = 2.0, CI = 17.6-25.3; t(356.1) =11.2, p < .001]. Pairwise comparisons using Chi-squared analysis of Correct and Incorrect Guesses separately for Male (No mask: 99.5% correct, Mask + Body: 96.9% correct; Mask: 80.9% correct) and Female Wearers (No mask: 99.0% correct, Mask + Body: 82.9% correct; Mask: 31.9% correct) for each condition with further Bonferroni corrections (p =.008) indicate that the bias towards male guesses only occurs in masked conditions [Mask: X^2 (1) = 97.1, p <.001, φ = .49; Mask + Body: X^2 (1) = 21.0, p <.001, φ = .23], and not in the No mask Condition [X^2 (1) = .60, p < .99], with the largest effect size in the Mask condition.

Racial group

Racial group guesses were also submitted to three separate analyses to compare 1) performance accuracy across Model visibility conditions (within subject), 2) across test locations (between subject) and 3) across mask race and wearer race (between subject) separately, with a Bonferroni correction for multiple comparisons (p = .016). Results of this analysis are summarised in Figure 5.8.

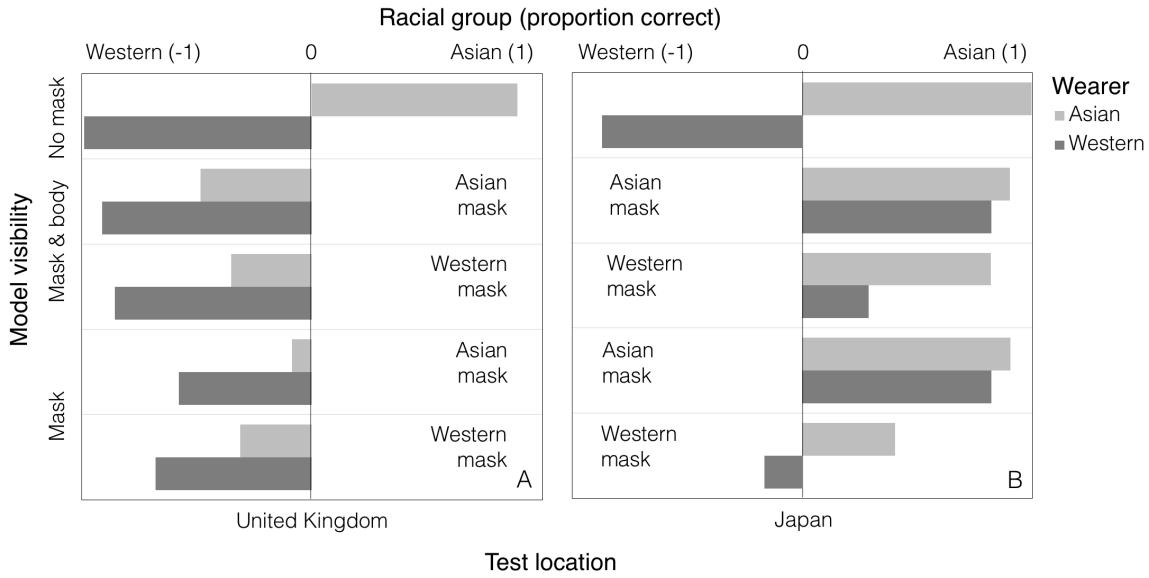

Figure 5.8. Proportion of correct Racial group guesses of the wearer beneath the mask by the Racial group of the Wearer, Racial group of the Mask, Test location and Model visibility.

Effects of Model visibility. To determine whether there was a difference in performance accuracy under different Model visibility conditions (No mask: 95.5% correct; Mask + Body: 55.8% correct; Mask: 56.3% correct), we submitted Correct and Incorrect Guesses to a Cochran Q Test by Model visibility (Bonferroni correction: p = .016). As expected, there was a significant difference between Conditions [Cochran's Q (2) = 229.1 p <.001]. Pairwise comparisons using Exact McNemar tests with further Bonferroni correction (p = .008) indicate significantly higher accuracy in the No mask vs. Mask + Body condition [X^2 (1, 396) = 137.9, p <.001] and No mask vs. Mask condition [X^2 (1, 396) = 140.7 p <.001], but no difference between the Mask + Body vs. Mask condition [X^2 (1, 396) = .01, p =.912].

Effects of Test Location. To determine whether the pattern of performance accuracy differed across test locations we submitted a percentage correct performance (based on the number of correct Racial group guesses per participant) to an independent sample t-test by Test Location (Bonferroni correction: p = .016). This indicated that performance accuracy was nearly

identical across the UK (M = 70.7%, SE = 2.0) and Japan (M = 70.0%, SE = 2.3) [Mean Difference = -.7%, SE = .3, CI = -.72 – -.53; t(394) = -.22, p = .821]. Pairwise comparisons using a Chi-squared analysis of Own and Other-race Guesses by the UK (No mask: 60.9%` Own-race, Mask + Body: 85.4% Own-race; Mask: 69.2% Own-race) and Japan (No mask: 53.7% Own-race, Mask + Body: 80.5 % Own-race; Mask: 74.2% Own-race) for each Model visibility condition independently with further Bonferroni corrections (p =.005) indicate a clear tendency towards base rate guesses in both Masked conditions [Mask + Body: X^2 (1) = 170.7, p < .001, φ = .66; Mask: X^2 (1) = 74.2, p <.001, φ = .43], but not in the No mask condition [No mask: X^2 (1) = 9.67, p =.005].

Interactions of Mask x Wearer. To identify whether the appearance of the mask is being attributed to the appearance of the wearer we conflate across Test Locations and submitted the percentage correct Racial group guesses per participant to a 2 x 2 between subject ANOVA of Mask Race (Own, Other) and Wearer Race (Own, Other; Bonferroni correction: p = .016). There was a significant main effect of wearer, with higher performance for participants viewing Own (M = 89.7%, SE = 1.7, CI = 86.0 – 93.0) vs. Other-race Wearers (M = 53.9%, SE = 1.6, CI = 50.8 – 57.1) [F (1, 392) = 224.97, p <.001, partial η^2 = .37], no main effect of Mask, with performance accuracy for Own (M = 69.7%, SE = 1.7, CI = 66.3– 73.0) and Other-race Masks (M = 73.8%, SE = 1.7, CI = 70.5 – 77.1) [F (1, 392) = 3.02, p =.083], but we do see an interaction effect between Mask and Wearer [F (1, 392) = 10.35, p =.001, partial η^2 = .03]. These results clearly show that signals of wearer's race prevail through the mask, but that viewer's guesses are swayed depending on the interaction between mask and wearer they observe.

To unpack this interaction, we followed up with a simple main effects analysis. This confirmed that there was a significant effect of Wearer in the Own-race Mask condition [F(1, 392) = 166.3, p <.001; partial η^2 = .30] and in the Other-race Mask condition [F(1, 392) = 69.3 p <.001; partial η^2 = .15], with lower accuracy for viewers of Other-race Wearers than Own-race Wearers in both groups. There was also a significant effect of Mask for Other-race Wearer [F(1, 392) = 13.7, p <.001; partial η^2 = .03], with higher accuracy for viewers of Other-race Masks than for those of Own-race Masks, but no difference between masks

for Own-race Wearer [F(1, 392) = .99, p =.320]. This suggests that viewers of the Other-race Mask worn by Other-race Wearers attributed this mask to the wearer. We do not observe this same effect in the Own-race Mask + Own-race Wearer condition, presumably because the default option (being a Western Guess in the UK or an Asian Guess in Japan) makes the Own-race guess nearly at ceiling across conditions.

Finally, to identify which conditions of Model visibility drove the Mask x Wearer effect, we ran pairwise comparisons using a Chi-squared analysis of Correct and Incorrect Guesses by Own-race and Other-race Masks and by Own and Other-race Wearer to compare only the two Masked conditions. This analysis revealed significant differences across both conditions [Mask: X^2 (1) = 80.70, p <.001, φ = .45; Mask + Body: X^2 (1) = 184.90, p <.001, φ = .68], with the largest effect size in the Mask + Body condition.

5.4 Discussion

We set out to determine what a viewer could tell about a wearer beneath a hyper-realistic face mask. To test for demographic estimation accuracy, we analysed participant performance for age, gender and racial group guesses of a person beneath a mask in a live viewing task, comparing a Mask to a Mask and Body view. We also set out to establish whether the appearance of the mask affected the perception of the person beneath the mask. Across demographic estimates we varied mask and wearer appearance to test for the attribution of the mask to the wearer.

What can we tell about a wearer beneath the mask? As it turns out, not that much. We expected that Age, Gender and Racial group estimates would be poor in the Masked conditions (Mask and Mask + Body), but that performance would improve with visible bodily cues (Mask + Body), especially for gender. First, we show that Age estimates of the person beneath the mask on average deviated 5 years from the wearer's real age when just a masked head is visible. This may seem like precise performance, considering that the FBI's age typical profiling

range is 20 years (Rae, 2012). However, the distribution of guesses presents a different picture. First, Guesses ranged from 16 to 85, and only 64% of participants guessed the wearer's age within 20 years of the wearer's actual age. Second, in estimating gender, the Mask condition leave 1/4 of participants choosing the wrong gender in. Considering that 50% is chance level, any individual profiler is likely to make a mistake. In fact, participants are three times as likely to think that the wearer was male compared to female. This indicates a perceptual bias towards guessing the male gender. Third, performance for Racial group estimates was only just above chance level (56% accuracy) in both masked conditions. We did not expect such low performance, as the wearer's eyes are visible through the mask and are known to hold racial group cues in distinguishing Asian from Caucasian individuals (Zhuang et al., 2010). Further research should consider how these effects may vary with different racial groups, but for now we assume that estimates of a wearer beneath a hyper-realistic face mask are unreliable across demographics.

In real world settings, performance is likely to be much worse. Here we designed the experiment so that participants could view the wearers for as long as they wanted and from any angle. They were also immediately made aware that the person was wearing a mask. In reality, we expect profilers will more likely be working with fleeting footage, blurry pictures, or confused eyewitnesses (Burton et al., 1999; Bruce et al., 1999; Wells & Olsen, 2003). There were also experimental constraints that allowed participants to make inferences that might not work in a real-world scenario. For the racial group estimate, we see that participants rely on the local base rate: participants think the wearer is Asian in Japan and Western in the UK. This is quite logical, but it is not always true. We also see reliance on an availability heuristic, where people use an example that comes easily to mind to explain what they observe (Schwarz et al., 1991). For example, Age estimates were correlated with their own age ($r = .27$, $p < .001$) which is clearly irrelevant to the task, but still affects their performance. In Racial group guesses we also see evidence that a congruency heuristic is being used. This heuristic helps to explain a given piece of information in the context within which we observe it (Wason,1960; 1968). In the context of face perception, we often talk about holistic face processing, which assumes that facial features are integrated to be

perceived as one (Tanaka & Gordon, 2011). If a face is perceived holistically, a congruent exhibit of social information (e.g. Western appearing features, Western appearing eyes) is more convincingly perceived as a real person than an incongruent exhibit social information (e.g. Western appearing features, Asian appearing eyes). Our data shows that in case of incongruency, the sum of information from facial features portrayed by the mask (knowingly irrelevant to the task) outdo the benefit of seeing unmasked feature such as the eyes (knowingly relevant to the task). Finally, our data suggests that the perception of the wearer is systematically biased by the appearance of the mask, which we will discuss in more detail in the paragraphs below.

Does the appearance of the mask affect the appearance of the wearer? This short answer is yes. For the Age guess, we predicted an overestimate, owing to the old appearance of the masks, and that is exactly what we see. The overestimate cannot be explained by a base rate inference (University-appropriate guesses would have resulted in an underestimate) and results are not fully explained the other heuristic principles explained above. Nonetheless the age data alone are not necessarily convincing, because the Age estimate does not allow for a dissociation of mask or wearer.

The Gender estimate does. We expected that if viewers were merely guessing (due to a lack of visual information or gender cues visible through the mask) we should see no bias towards the wearer being male or female. If viewers would rely on visual information provided by the mask (always male), we should see a bias towards male guesses. This is indeed what our results show. In fact, 75% of observers think the wearer is male, which is not representative of any other visible cue we measured, whether relevant (gender of the wearer) or irrelevant (base rate, experimenter gender etc.) to the task at hand. Hence, it is highly likely that the appearance of the mask biased the perception of the wearer.

The most convincing evidence comes from the racial group estimate, where we dissociate the race of the mask and the race of the wearer from one another. Here too we expected that if viewers were merely guessing the wearer's race or relying on visual information of the wearer through the mask, we should

see no bias towards the wearer being of Own or Other-race. In fact, the data clearly reveal two biases. Most dominant, we see the representativeness bias causing a ceiling effect towards the Own-race guess. It is likely that the base rate tendency pushing performance in Own-race Guesses towards ceiling obstructs the mask effect from appearing in the Own-race mask conditions. The interaction between the mask and wearer does show that wearing an Other-race mask increases the frequency of Other-race guesses. We recommend that further research explores the extent to which these affects could be removed from ceiling, for example by replicating this work in a multicultural environment. For now, we conclude that although the appearance of the mask is clearly irrelevant to the task of the observer, it affects their performance across demographic estimates. This shows us that the appearance of the mask is misattributed to that of the wearer.

This FAE of facial disguise could in fact improve our understanding of the FAE's assumptions. Not only is the attribution of the mask to the wearer an error (participants are aware that the mask is highly irrelevant to their task of judging the wearer), perception is at the core of social interaction. Showing that an FAE is apparent in this perceptual task, highlights how fundamental the FAE must be to cognitive processing and social interaction. In summary, we show that telling who is beneath a hyper-realistic mask is not just affected by a number of well-established heuristics, it is also affected by a perceptual attribution error. If profiling relies on heuristics and is biased by cues irrelevant to the facial features which prevail through the mask (e.g. eyes and face shape), it is a highly unreliable practice.

Knowing that these biases affect performance does not mean estimating mask wearer's demographics would be pointless. In combination with other information it could be rather useful.

First, when the confederate's body is visible too, the Age overestimate is completely abolished and nearly all viewers get the participant's Gender right (96% accuracy). Indeed, there are known cues to age (the wearer's hands: (Farage et al., 2013; posture: Van den Stock, Righart & De Gelder, 2007) and

gender (Body: Mather & Murdoch, 1994; Skin colour: Fink, Grammer & Matts, 2006, Frost, 1988) which would explain these accuracy levels. This too may appear better than it is as 1) bodily cues are likely to combine with the heuristic information (which we just established is more useful in this experiment than in reality) and 2) bodily cues would be less revealing if the wearer would have incentives to hide their identity. In the set-up of this experiment, we put no effort into masking body shape, gate, clothing style or skin colour, whereas this would be relatively easy (e.g chest-binding, padded clothing, gloves or other props). Racial group estimates do not benefit from viewing the wearer's body, but in testing only Asian and Western wearers we do not know whether other racial groups with more distinctive skin colour of features (Maclin & Malpass, 2001; Parra, Kittles & Shriver, 2004) could be discriminated from bodily cues despite a whole-head disguise or hyper-realistic face mask. We recommend that both deliberate disguise of the body and discrimination between other racial groups is empirically studied. At present we cautiously conclude that, provided the observer has a good view of the wearer's body, age and gender profiling of the mask wearer may sometimes be a viable option. Racial profiling does not seem viable in this context.

Second, demographic profiling might be more effective if a crowd analysis were to be used. Crowd analysis is based on the idea that a group tends to outperform any one individual (Galton, 1907; Krause et al., 2011; White, Burton, Kemp & Jenkins, 2013) and can be particularly useful when no one person is an expert (e.g. categorisation of an unfamiliar individual). Using a crowd analysis relies on the principle that some facial cues are accessible to some degree. The 75% accuracy rate in the Gender guess is a sign that some such cues must prevail through the mask. In Racial group estimates too, we observe a strong main effect of wearer, indicating that some information must be visible through the mask. These levels of accuracy make it highly likely that any individual observer would make a mistake, but collectively results could become very accurate.

To sum up, demographic estimates of the wearer beneath a hyper-realistic face mask are poor if only the wearer's head is visible. Gender and Age estimates reliably improve when the body is visible, but Racial group guesses do not.

More broadly, our findings are first to address the effects of whole-head disguise on person categorisation. Up to recently, using partial disguise (e.g. hat or moustache) rather than a regular whole-head disguise (e.g. balaclava) to avoid capture may have been the most appealing choice, as regular whole-head disguises draw more attention and immediately flags a person as dangerous. Hyper-realistic masks change this situation. As hyper-realistic masks become more realistic, less detectable, and through the distracting demographics of face mask could even protect a wearer's identity, it seems likely that whole-head disguise will be used increasingly.

Chapter 6.

Identifying the person beneath the mask

6.1 Summary

Capturing a culprit who used a hyper-realistic mask relies on 1) detecting the mask and 2) identifying the person beneath the mask. In some situations, identification may be achieved by simply removing the mask. However, in other situations, this may not be possible, and the wearer must be identified through a mask. In the current study, we simulate the latter situation using face photographs. Two confederates (1 female, 1 male) were photographed wearing each of three different hyper-realistic face masks (1 female, 2 male). Experimental participants were asked to categorise these photographs according to the identity of the wearer. Those who were unfamiliar with the mask wearers exceeded chance performance, achieving 75% accuracy. Surprisingly, those who were familiar with the mask wearers did no better. In fact, the range of performance was the same for both subgroups (Unfamiliar 41–86%; Familiar 43–86%). Viewers tended to choose whichever identity was consistent with the gender of the mask, suggesting that judgements of the wearer's identity were influenced by the mask's appearance.

6.2 Introduction

Hyper-realistic masks have gone undetected in a number of high profile criminal cases (see Appendix 1.2). In previous chapters we have shown experimentally that they are both difficult to detect (Chapter 1-3), and argue that they are better at protecting the wearer's identity compared with regular whole-head disguises (see Chapter 5). With this in mind, hyper-realistic masks will likely become an increasingly popular choice of facial disguise, as we have seen from media

reports over the last 9 years (see Appendix 1.1). This presents a potential problem for the legal system. Prosecution today relies heavily on facial appearance for person identification (Robertson, 2018). If we do not improve our means to detect these masks, we can no longer take for granted that a face contains valid cues to identity. The first half of my thesis focused on detection of hyper-realistic masks. Once a mask is detected, we still need to identify the wearer. In the previous chapter we showed that profiling of the wearer, as a first step towards identification, is highly error prone. As a logical next step, we focus on actual identification of the wearer. Based on performance in Chapter 5, we expect that recognition of a mask wearer is poor, at least for individuals who are unfamiliar with the face of the wearer. However, familiarity confers a huge advantage in face identification tasks (Bruce, 2012). If this advantage extends to situations in which the face is masked (perhaps exploiting shape cues or exposed regions of the face), this may offer a path to identification of masked individuals in applied settings.

One of the most consistent factors to mediate recognition or face matching accuracy is viewer familiarity with the target individual. For example, Jenkins et al., (2011) explored the differences in performance for familiar and unfamiliar viewers in a card-sorting task: containing two local celebrities from the Netherlands, unfamiliar to a British audience. Dutch and British participants sorted both sets of celebrities by identity. Results showed that unfamiliar viewers generally perceived 7.5 identities in the set, whereas familiar viewers nearly always selected just two. In sum, face matching is much easier task when you are familiar with a person compared to when you are unfamiliar.

There are a number of reasons to expect that familiar viewers would also perform better under disguised conditions. People generally identify familiar faces with little effort, despite possibly large variations of lighting, viewpoint and expressions, use or spectacles and hats. Moreover, familiarity with a face permits identification even from very low quality images or under conditions of disguise.

For example, Burton et al. (1999) compared performance of participants on familiar and unfamiliar face recognition in a face memory test. Participants were

shown a series of video clips from poor-quality CCTV and were later shown good-quality photographs. Their task was to decide whom they had and whom they had not been shown on video. For items participants were familiar with they performed almost perfectly, whereas for their performance for items they were unfamiliar was only just above chance.

Dhamecha, Singh, Vatsa and Kumar (2014) looked at the familiarity advantage for models who were asked to disguise themselves using a variety of props. Participants were shown these images in pairs of match or mismatch trials and simply asked to decide whether 2 images were shown of the same or different individuals. They found familiar viewers outperformed unfamiliar viewers.

Noyes and Jenkins (under review) also considered the effect of deliberate disguise on face matching accuracy for familiar and unfamiliar viewers, with a much-improved paradigm. Models were explicitly incentivised to induce identification errors through altering their appearance. They found that unfamiliar observers were less accurate for disguise items than for undisguised items, even when they were informed of the disguise manipulations. Familiar observers, on the other hand, saw through most (but not all) disguises, even under these viewing challenging conditions.

Ideally, we would be able to train professionals how to become familiar with faces quickly, but there is little evidence that face recognition ability can be trained (Towler, White, & Kemp, 2014, 2017; White, Kemp, Jenkins, & Burton, 2014). We do, however see large individual differences in performance. For example, ability on a live-to-photo face-matching task by White et al. (2014) found performance to vary from 70% to 100% accuracy in only forty-nine individuals. In fact, face recognition ability is thought to be on a spectrum, from individuals that completely lack the ability to recognise faces (*congenital prosopagnosics;* Behrmann & Avidan, 2005) to highly skilled face recognisers on a variety of tasks (*super-recognisers*; Russell, Duchaine & Nakayama, 2009). A recent study on face-matching classed super-recognisers (Robertson et al., 2016; n = 4) working for the London Metropolitan police force showed that they consistently performed above normal levels as measured in police trainees on the GFMT (n = 194), and a

student sample on the Model Face Matching Test (similar, but more difficult version of GFMT, n = 64) and Pixelated Lookalike Test (face matching for pixelated images of famous individuals and their lookalikes, n = 30). Although training individuals is unlikely to be a solution to the face recognition problem, it is possible to recruit individuals that are naturally good at face recognition.

Hence, even if unfamiliar viewers struggle to identify the face beneath the mask, we still see two possible routes towards improved performance. First, if familiar viewers outperform unfamiliar viewers on this task, identification might be more reliable when attempted by individuals who are familiar with suspected wearer. Second, if we see large individual difference in performance, it may be possible to recruit individuals who are particularly good at identifying faces even under these difficult conditions.

Given that Chapter 5 indicated that demographic estimation was poor, we started with a very easy face identification task. We adopt a two-alternative forced choice (2AFC) design (Fechner, 1860/1866). developed to measure quantifiable perceptual acuity (Bogacz et al., 2006). Here we use this approach to ask participants to decide who is beneath the mask in each of a series of photos, with a choice of just two possible wearers.

6.3 Experiment 9:

Computerised 2AFC recognition of wearer beneath a mask

As a primary interest, we compare familiar and unfamiliar viewers and individual differences in performance. Given that viewers who are familiar with the wearer are expected to have a clearer mental image of the wearer's face (Bruce, 2012), we predicted that performance would be higher for familiar viewers than for unfamiliar viewers. By measuring the range in recognition performance, we aimed to explore the prospects for recruiting high-performing participants for this task.

As a secondary interest, this experiment allows us further to explore the

attribution error observed in Chapter 5. For all demographic estimates of the wearer, we saw that the appearance of the mask influenced the demographic judgement of the wearer, despite participants being explicitly informed to disregard the information portrayed by the mask. In this task, we consider a more abstracted version of this FAE. In both tasks viewers are also explicitly informed to ignore the demographic and identity cues portrayed by the mask, but in Chapter 5 the FAE remained in a task specific domain: the mask showed clear age, gender and racial group cues and the participant's task was to judge exactly that. In this task, the masks show clear age, gender and racial group cues, but the task is to categorise identity. Hence the interference of the mask would have to cross into a different task domain. An attribution bias in this context would provide evidence of the fundamental nature of the error (Harvey, Town, & Yarkin, 1981; Sabini, Siepmann, & Stein, 2001).

To this end, we recruited two confederates (1 female, 1 male) who served as mask wearers in this study. If a cross-domain FAE occurs in this context, we expect that the gender of the worn mask (1 female, 2 male) should influence identification accuracy for the two wearers. Specifically, we predict that male masks will promote male identity judgements for the wearer, and that female masks will promote female identity judgements. Given that familiar viewers should be more firmly tied to their prior knowledge of the wearer, we expected the attribution error to be smaller for familiar viewers than for unfamiliar viewers.

Method

Ethics statement. Ethics approval for this experiment was obtained from the departmental ethics committee at the University of York.

Participants. Fifty members of the volunteer panel at the University of York (20 familiar, 30 unfamiliar; 14 males; mean age = 24, age range 18–56 years) took part in exchange for a small payment or course credit.

Design and Stimuli. We asked two groups of participants (IV1; Familiar and

Unfamiliar; between subjects) to make an identity judgement (DV) for two confederates (IV2; Mladen and Florence; within subject). They judged pictures of each confederate without a mask and whilst wearing three different hyper-realistic face masks (IV3; No mask, OFM, OMM, YMM; within subject).

As in Chapter 1, we used three different models of masks from Realflesh Masks, Quebec, Canada: *The Pensioner* (Old Male Mask; OMM), *The Fighter* (Young Male Mask; YMM), and *The Grandma* (Old Female Mask; OFM).

To generate images, we took 10 photographs of the 2 confederates wearing each of the three masks and without a mask (60 Mask images in total, 20 No mask images). Photographs depicted the confederate's head in frontal view with no occlusions and were taken indoors and outdoors under different viewing conditions to approximate the range of variability seen in natural face images (Jenkins et al., 2011). All photos were cropped to show the head region only and resized to 540 pixels high x 385 pixels wide for presentation (see Figure 6.1).

Figure 6.1. Variable face photographs of wearer ‘Mladen’ and wearer ‘Florence’ without a mask (top row) and in three different masks (bottom rows).

Both familiar (n = 20) and unfamiliar (n = 30) viewers judged each mask image in a 2AFC design, in two blocks. To establish baseline categorisation accuracy, participants decided whether No mask images showed ‘Florence’ or ‘Mladen’ in the first block (20 images). To evaluate the effect of the mask in the same task they decided whether Mask images showed ‘Florence’ or ‘Mladen’ in the next block (60 images). Images in both blocks were presented in a random order. For each photo, participants made a Florence/Mladen judgement via keypress. Each image stayed on screen until a response was made.

Procedure. Participants were informed that there were two mask wearers, and that the same masks could appear more than once, so that judgments should be assessed on an image-by-image basis. They were informed that no time limit was imposed for this task, but were asked to make their judgement as quickly and accurately as possible.

In the instructions, participants were introduced to their task of identifying whether the image showed wearer Mladen or wearer Florence. They were informed that 'Mladen [was] male with dark hair and dark eyes', and that 'Florence [was] female with dark hair and dark eyes, and were shown a single image of each individual to give an indication of their appearance (Figure 6.2).

Figure 6.2. Image of Florence (left) and Mladen (right) shown to participants with the task instructions.

Participants then completed the first block, in which they judged the 20 No mask images, followed by the second block in which they judged the 60 Masked images. In each trial, a single image was presented in the centre of the screen. In the first block, this was headed with the caption '*Who is shown in this picture?'*. In the second block, this was headed with '*Who is underneath the mask?'.* Both blocks showed response options 'Florence – Press Z' to centre left and 'Mladen - Press M' to centre right of the image. Participants pressed 'Z' if they thought they saw a picture of Florence, or 'M' if they thought they saw a picture of Mladen, which initiated the next trial. Each participant completed 3 practice trials of a

different individual, followed by 1 block of the 20 No mask images in a random order and 1 block of 60 recorded Mask images in a random order (80 experimental trials in total).

At the end of the experiment, participants were debriefed and provided a familiarity rating (1-7 Likert scale) for each wearer to be certain they had been assigned to the right group. The entire experiment took approximately 5 minutes to complete.

Results

Mean percentage correct scores were submitted to a 2 x 2 x 2 mixed ANOVA with the between-subjects factor of Familiarity (Familiar, Unfamiliar), and the within-subjects factors of Mask (Masked, No mask) and Wearer (Florence, Mladen). The results are summarised in Figure 6.3.

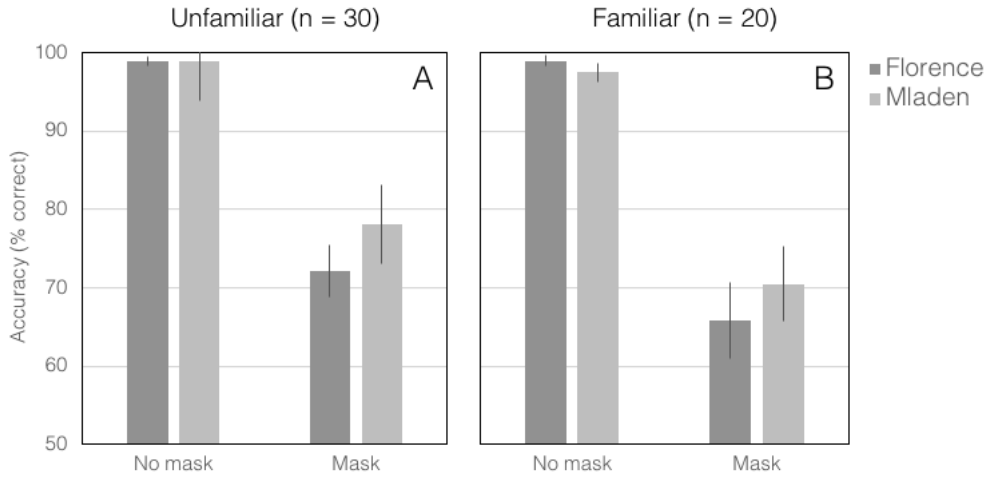

Figure 6.3. Identification accuracy (percentage correct) for (A) Unfamiliar viewers and (B) Familiar viewers, separated by Wearer and by Mask condition.

Identification accuracy. Firstly, we found a main effect of Familiarity. Contrary to expectations, familiar viewers (M = 83.2%, SE = 1.4, CI = 80.4 – 86.0) performed significantly *worse* than unfamiliar viewers (M = 87.1%, SE = 1.2, CI = 84.7 – 89.4) [$F(1,47) = 4.47$, $p = .04$, partial $\eta^2 = .09$], although this effect was numerically small.

We also found a significant main effect of Mask, with poorer performance for Mask trials (M = 71.7%, SE = 1.7, CI = 68.2 – 75.1) than for No mask trials (M = 98.6%, SE = 0.4, CI = 97.7 – 99.5), [$F(1,47) = 243.78$, $p < .001$, partial $\eta^2 = .84$]. Despite the significant reduction in accuracy, we note that participants performed significantly above chance in all conditions [Mean Difference = 22.3%, CI = 18.8 – 25.8; $t(48) = 12.78$, $p < .001$].

There was also a main effect of Wearer, with poorer performance for Florence trials (M = 84.0%, SE = 1.1, CI = 81.8 – 86.2) than for Mladen trials (M = 86.3%, SE = .9, CI = 94.4 – 88.1), [$F(1,47) = 6.84$, $p = .012$, partial $\eta^2 = .13$].

These main effects were qualified by a significant Mask x Wearer interaction [$F(1,47) = 11.03$, $p = .002$, partial $\eta^2 = .190$]. Simple main effects confirmed that there was a significant difference in the Mask and No mask conditions for Mladen trials [$F(1,47) = 198.26$, $p < .001$, partial $\eta^2 = .81$] and for Florence trials [$F(1,47) = 189.42$, $p < .001$, partial $\eta^2 = .80$]. There was a significant difference between Mladen and Florence in the Mask condition [$F(1,47) = 10.065$, $p = .003$, partial $\eta^2 = .18$], but not in the No mask condition [$F(1,47) = 1.464$, $p = .23$].

One possible explanation of this interaction is attribution of the mask to the wearer. We return to this issue later in the chapter. For now, the important finding is that identification accuracy was poor in the Mask condition (26% error rate), and was not enhanced by familiarity with the wearer.

Individual differences. It is possible that performance could be improved by seeking out high-performing individuals. As can be seen in Figure 6.4, there was little variability in accuracy in the No mask condition (range 90–100%). In contrast, accuracy in the Masked condition spanned the entire range (42–100%).

Unsurprisingly, there was no correlation between Mask and No mask trial performance (r = .15, p = .31).

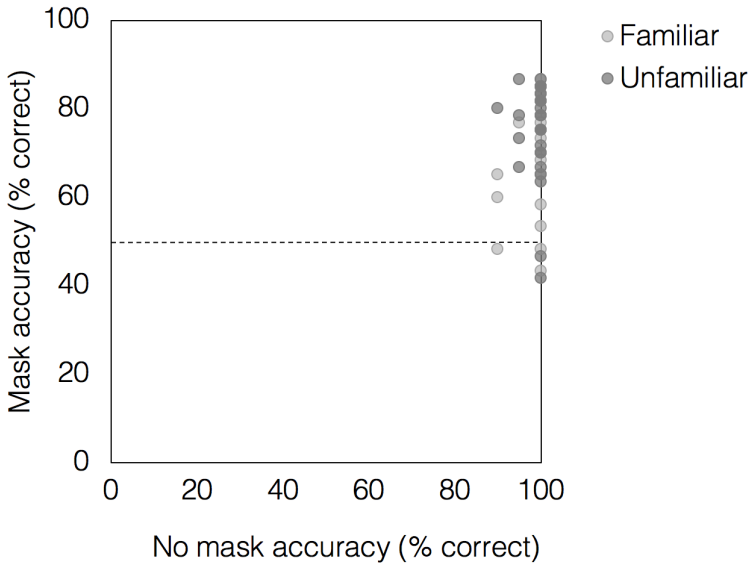

Figure 6.4. Scatterplot showing participants' mean identification accuracy rates in the Mask and No mask conditions. Familiar viewers in light grey, Unfamiliar viewers in dark grey.

Whilst we see near perfect performance for all participants in the No mask condition, the data reveal quite a range in performance for the Mask condition. Some participants were performing at chance, whilst others performed almost perfectly. Interestingly, the performance distributions for Familiar and Unfamiliar viewers coincide completely. Thus, familiarity does not explain the individual differences in performance seen here.

We see this even more clearly when we correlate Familiarity ratings for each wearer (Florence: M = 2.5, SD = 1.8; Mladen: M = 3.0, SD =1.4) with accuracy on Mask trials for each wearer. As can be seen in Figure 6.5, there was no significant correlation between these measures for either Florence [r = -.178, p = .216] or Mladen [r = -.237, p = .097]. However, there was strong positive correlation in performance between the two wearers [r = .656, p < .001] (see Figure 5.5). This consistency across wearers suggests that something in their approach to the task (whether conscious or not) puts some participants at an

advantage over others.

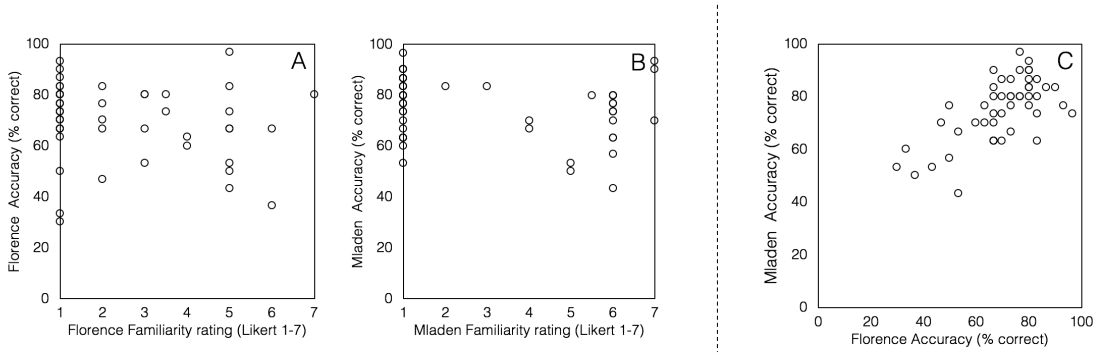

Figure 6.5. Scatterplots showing (A) Familiarity ratings by Accuracy for trials of Florence, (B) Familiarity and Accuracy for trials of Mladen, (C) Accuracy for trials of Mladen and Florence.

Attribution Error. Participants were informed that Mladen trials and Florence trials were equally likely, we note that error rates were somewhat higher for Florence (31%) than for Mladen (26%). One possible explanation for this imbalance is that 2 of the 3 masks were male. That is, the gender of the mask could have been attributed to the gender (hence identity in this task) of the wearer. To test for this possibility, we submitted mean accuracy data to a 3 x 2 x 2 mixed ANOVA with the within-subjects factors of Mask (OFM, OMM YMM) and Wearer (Florence, Mladen), and the between-subjects factor of familiarity (Familiar, Unfamiliar). The results of this analysis are summarised in Figure 6.6.

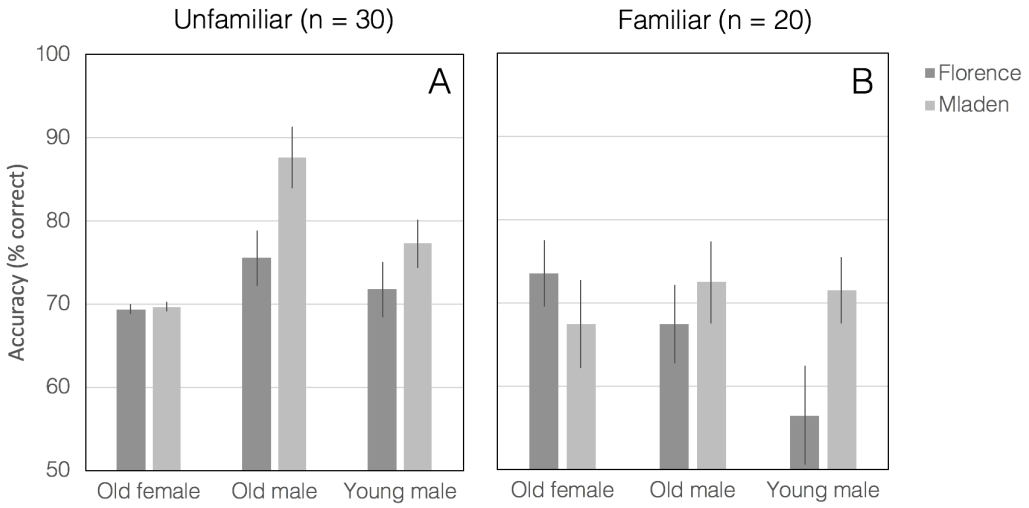

Figure 6.6. Summary of performance accuracy in percentage correct identification for (A) Unfamiliar viewers separated by Wearers and Mask Type (masked images only) and (B) Familiar viewers separated by Wearers and Mask Type (masked images only).

Firstly, we again show main effect of Familiarity, where familiar viewers (M = 68.2%, SE = 2.6, CI = 62.8 – 73.5) performed significantly worse than unfamiliar viewers (M = 75.2%, SE = 2.2, CI = 70.7 – 79.6) [$F(1,47) = 4.14$, $p = .047$, partial $\eta^2 = .08$], and a main effect of Wearer, with lower accuracy for trials depicting Florence (M = 69.0%, SE = 2.1, CI = 64.7 – 73.3) than for trials depicting Mladen (M = 74.3%, SE = 1.7, CI = 71.0 – 77.7), [$F(1,47) = 10.07$, $p = .003$, partial $\eta^2 = .18$].

We also found a significant main effect of Mask [$F(1,47) = 5.2$, $p = .007$, partial $\eta^2 = .100$]. Pairwise comparisons show lowest accuracy for YMM trials (M = 69.2%, SE = 1.8, CI = 65.6 – 72.9 [vs. OMM: p=.009; vs OFM: p =.730), followed by OFM trials (M = 70.0%, SE = 2.2, CI = 65.6 – 74.4; [vs OMM p = .008]) and OMM trials (M = 75.8%, SE = 2.4, CI = 69.2– 80.6).

These main effects were qualified by a significant Wearer x Mask interaction [$F(1,47) = 3.39$, $p = .038$, partial $\eta^2 = .07$]. To follow up this interaction, we conducted a simple main effects analysis. For Florence trials there was no

significant difference between the three masks [F(1,47) = 2.743, p = .075], whilst for performance on Mladen trials there was [F(1,47) = 5.556, p = 007, partial η^2 =.20]. Pairwise comparisons for Mladen trials indicate that accuracy was higher on OMM than on OFM trials (Mean Difference = 11.5%, SE = 3.6, CI = 5 – 18.7, p = .002), but that there was no difference between YMM and OFM trials (p = .182) or OMM and OFM trials (p = .087). In performance on OFM trials there was no difference between the two Wearers [F(1,47) = .563, p = .475], but accuracy was higher on Mladen trials than on Florence trials for both for the OMM [F(1,47) = 10.020, p = .003, partial η^2 =.18] and YMM trials [F(1,47) = 6.259, p = .016, partial η^2 =.12]. This pattern is consistent with attribution of the gender of the mask to that of the wearer, in turn influencing the identity judgement of the person beneath the mask.

6.4 Discussion

In this experiment, we primarily investigated 1) recognition of the person beneath a hyper-realistic mask comparing Familiar and Unfamiliar viewers and 2) individual differences in recognition of the person beneath the mask. We found that unmasked faces were categorised accurately overall (>98% correct), whereas Masked faces were not (70% correct). To our surprise, Unfamiliar viewers (75% correct) somewhat outperformed Familiar viewers (68% correct). We also saw large individual differences in the Masked recognition task, for both Familiar and Unfamiliar viewers. In fact, the range in performance for Familiar viewers (43%-87%) and Unfamiliar viewers (41-87%) was nearly identical. Although target Familiarity does not explain the individual differences observed, a strong correlation in accuracy between the two wearers does.

One unexpected aspect of these findings is the absence (or reversal) of a familiarity advantage for recognising the face beneath the mask. One possibility is that so much of the person's facial appearance is obstructed and/or distorted by the mask, that recognition cannot rely on normal familiar face processing mechanisms, putting unfamiliar viewers at an advantage.

We base this on the following argument. There are few identity cues left to see through the mask (e.g. eye region, face shape). This alone should not abolish recognition completely, as familiar faces can be recognised from the eye region only (Tanaka and Farah, 1993; Peterson, Cock, & Eckstein, 2008; Keil, 2009). For example, Tanaka and Farah (1993) showed that occluding all other facial information still allowed for an 80% accuracy rate in a 2AFC recognition task for familiar viewers.

Having said that, recognition is not all in the eyes. In the same study, Tanaka and Farah (1993) showed that additional facial information improved performance. In the case of realistic face masks, we expect that additional information is drawn from the appearance of the mask, which interferes with the wearer being recognised. This suggestion could be construed as a variation of the face composite effect (Young, Hellawell and Hay, 1987). The standard face composite effect shows that two aligned halves of different familiar faces are processed holistically as one new identity, interfering with recognition of both (Young, Hellawell and Hay, 1987). The suggestion here is that features of the mask (majority) and features of the wearer (minority) similarly fuse into a holistic unit. Holistic processing is thought to be particularly important for familiar face recognition (for reviews see Maurer, Le Grand, & Mondloch, 2002; Tanaka & Gordon, 2011; Degutis et al., 2013), whilst recognition of unfamiliar faces is thought to rely more on processing of featural information (Lobmaier & Mast, 2007; Megreya, 2018). In our study, it is possible that participants who came to the task as Familiar viewers were relying on holistic processing to support recognition. Rather than facilitating performance, this dependence may have increased interference from the mask and disadvantaged Familiar viewers in comparison to Unfamiliar viewers who were relying on feature-based processing instead.

If this argument is accepted, it calls into question when recognising a disguised face transitions from being a face processing task to being a feature-based classification task. Where, Noyes and Jenkins (under review) showed that some of their deliberate disguise manipulations removed familiar viewer from ceiling performance, hyper-realistic mask disguises seem to illustrate a tipping

point of familiarity, reverting even familiar face recognition to levels of performance akin to categorical or feature based classification. Our data supports that performance could in fact be driven by such categorical classification. In Chapter 5, the gender guess based on a live viewing of the confederate comes to 73% accuracy in the Mask Only condition. This is strikingly similar to our recognition performance at 70% accuracy. It suggests that for lack of better cues, recognition may have been reduced to a gender guess. Although replicating these results using a larger variety of confederates would be required to be certain, our findings hint at the possibility that separable principles governing disguised and undisguised face recognition. With more certainty, we can state that increased familiarity does not provide an advantage under all face recognition scenarios.

If this is true, the pattern is potentially important. Familiar viewers are prone to over-estimate their own performance (Bothwell, Deffenbacher, & Bothwell, 1987; Busey et al., 2000; Deffenbacher, 1980). Generally, we assume that we are able to recognise the people whose faces we know well regardless of the circumstances in which we view them. With the high error rate seen here, overconfidence could increase misidentification of a wearer beneath a realistic mask in the real world. This worry is easily dissolved, if confidence is correlated with performance on this task, hence we recommend that future research explores confidence of the viewer for this wearer recognition task.

Our data also suggest another possibility for improving performance. Individual differences in performance were large, and consistent across the two wearers. This suggests that stable cues are used to allow effective recognition by some, and not by others. In Chapter 4 we found that a similar pattern in performance for mask detection. An image analysis pointed towards a cue right beneath the eyes – where the mask breaks to reveal the wearer's eyes – used by high performers. As the eye region also contains the only uncovered facial cues to the wearer's identity, it is plausible that attentiveness to this area of the face may also be advantaging recognition of the wearer. Conclusive evidence would require estimates of test-retest reliability across a range of tasks (Robertson et al., 2016; Russell, Duchaine, & Nakayama, 2009). High performers may fit a 'super-recogniser' pattern across tasks (Robertson et al., 2016), but super-recognition

ability has been hard to define because a range of abilities is observed across face recognition tasks (Davis, Lander, Evans & Jansari, 2016; Noyes, Hill, & O'Toole, 2018). Some have associated strong recognition performance with holistic processing (e.g. DeGutis et al., 2013; see Richler, Floyd, & Gauthier, 2015 for contrasting findings), so as we expect that holistic processing does not advantage performance here, or may rather rely on resisting holistic processing, we may be thinking of an entirely different type of talent. If this is correct, there may actually be better prospects for the recognising the wearer beneath the mask. If viewers are merely focusing on a particular area of the face or cue visible through the mask, recognition rates could be improved by drawing attention to this region. If so, it could pave the way for a simple training intervention, which contrasts with the difficulties experienced in training of face identification in other tasks (Towler, White, & Kemp, 2014; 2017; White, Kemp, Jenkins, & Burton, 2014). Eye-tracking data in combination with accuracy rates could elucidate whether such cues are available. Until then, we suggest screening for high-aptitude individuals as a route for improved recognition of wearers beneath hyper-realistic masks.

Finally, we considered whether there was evidence in this data for attribution of the mask to the wearer. We predicted that accuracy levels for our wearers would be dissociable by gender of the mask. For male masks (OMM and YMM) we indeed see that viewers were significantly more likely to think that the wearer was a male, reflected in a 9% bias towards Mladen as the wearer. This is the case for Familiar and Unfamiliar viewers. For the female mask (OFM), we see a 3% bias towards Florence as the wearer, but this difference was not significant and was driven by Familiar, not Unfamiliar viewers. However, the magnitude of dissociation between Mladen and Florence across the three masks does not come as much of a surprise, as the appearance of the OFM is not as feminine (3% of a bias towards Florence) as the YMM is masculine (9% of a bias towards Mladen; see appendix 2.3 for social character judgements). It would be interesting to repeat this study with wearers and masks with a greater span in femininity-masculinity, to see how these effects unfold. For now we can conclude that, our data supports the FAE but is not conclusive. Consistent with the attribution errors observed in Chapter 5, it does suggest that even though viewers were instructed

to ignore the social cues portrayed by the mask, those cues still affected judgement of the wearer, both directly (manipulation of gender and effect on gender judgement) and indirectly (manipulation of gender and effect on identity judgement). This provides further evidence of the fundamental nature of these attributions.

To summarise, we replicate results from Chapter 5 that with the absence of facial cues to the wearer's identity, judgement of a person wearing a hyper-realistic mask is susceptible to bias. In addition, we show that such bias transfers across domains, with the apparent gender of the mask affecting identification of the wearer.

Chapter 7.

Social inference attribution of the mask to the wearer

7.1 Summary

People make reliable social inferences judgements from facial appearance. These judgements are automatic (Wills & Todorov, 2006) and predictive of how observers treat a target in real-world situations (Todorov et al., 2008). This relies on a heuristic that facial appearance predicts character and behaviour, but there is little evidence that observer's judgements is an accurate predictor of target character or behaviour. Some attribute the lack of evidence to the measurements that are used to study the relationship. Here we test a situation, where we are certain that facial appearance is an *inaccurate* prediction of the actor's character, as a stringent test of social inference error. By asking viewers to judge two confederates in three different masks, we are able to dissociate the attributions of complex character (attractiveness, dominance, trustworthiness; Experiment 10) and personality traits (openness, conscientiousness, extroversion, agreeableness, neuroticism; Experiment 11) of the mask from those of the wearer beneath the mask. We provide evidence of the automaticity with which viewers bind facial appearance to individuals, even when they know that facial appearance has been manipulated. We argue that silicone face masks would allow a realistic manipulation which dissociates the actor's character from their appearance. There is potential to experimentally manipulate appearance along any social trait dimensions and whilst participating in social scenes. This could be used to study social inferences in real-world situations without relying on observational data or artificially manipulated photographs and could to be a useful testbed for studying the accuracy of such effects.

7.2 Introduction

Viewers make instantaneous demographic, emotion and social judgements to evaluate the interpersonal dynamic between themselves and an unfamiliar person. This is an intuitive and automatic process, based largely on physical and mainly facial attributes (Todorov, Said, Engell, & Oosterhof, 2008; Zebrowitz, Voinescu, & Collins, 1996). In fact, a growing body of literature shows that facial appearance is one of the main predictors of an observer's judgement of that person (Willis and Todorov, 2006).

How a person's character is judged has consequences. Amongst the most famous examples are that attractiveness judgements affect dating success (Olivola et al., 2014) and court decisions (Kutys, 2012), that judged competence (Olivola & Todorov, 2010) or dominance (Chen et al., 2014; Chiao et al., 2008; Little et al., 2007) predict political candidate success (Mannetti, Brizi, Belanger & Bufalari, 2016) or CEO salaries (Graham et al., 2014; Rule & Ambady 2008, 2009) irrespective of their performance in respective positions (Graham et al., 2014), and that an untrustworthy face predicts money lending irrespective of their lending history (Chang et al., 2010; Rezlescu et al., 2012; Schlicht et al., 2010; Stirrat & Perrett, 2010; Tingley, 2014; van 't Wout & Sanfey, 2008). More worryingly, defendants who have untrustworthy-looking faces are more likely to receive guilty verdicts (Porter et al., 2010) even when there is less evidence of their guilt (Dumas & Teste, 2006; Porter et al., 2010), receive harsher sentences (Blair et al. 2004), and are more likely to receive the death sentence (Eberhardt et al., 2006).

These consequences are remarkably consistent, considering the accuracy of the inference remains highly debated (Berry & Brownlow, 1989; Bond, Berry, & Omar, 1994; Olivola & Todorov, 2010a,b; Zebrowitz et al., 1996). Some studies have predicted internal traits and behavioural tendencies from facial photographs, such as criminal behaviour (Porter et al., 2008; Valla et al. 2011) and political orientation (Rule & Ambady, 2010; Samochowiec et al., 2010), but all have their non-replication counterparts (see Olivola & Todorov, 2010b; Rule et al., 2013).

One problem in assessing accuracy of social inferences is measurement. Relating social inference judgements to behaviour is nearly always a two-stage process: the behaviour of one group is noted, and another judges their appearance (e.g. Bonnefon, Hopfensitz, & De Neys, 2017; Rule et al., 2013). This approach allows real-world decisions (e.g. prior commitment of a crime) to be related to the facial inference drawn from an actor's facial structure (e.g. trustworthiness), but is also highly unconstrained (e.g. criminal behaviour is not necessarily a proxy for untrustworthiness), often observational (with the goal of predicting real, significant outcomes) and therefore correlational (see Olivola & Todorov 2010b; Rule et al., 2013). In addition, the evaluation of facial appearance is most commonly done with face photographs. This assumes one-to-one mapping from a person's facial appearance to his/her perceived characteristics, which we know not to be the consistent (Jenkins & Burton, 2011). In fact, Jenkins et al. (2011) showed that in using ambient face images, within-person variability in attractiveness (images of the same person) ranged across between-person variability in attractiveness (images of different persons). Using the same approach, Sutherland et al. (2017) and Mileva (2017) replicated this striking pattern for dominance, trustworthiness and attractiveness judgements, and even for emotional valence.

To step away from this two-tiered approach, a small number of studies have systematically manipulated facial appearances to demonstrate that there is a causal relationship between face-based social attributions and various important outcomes (Little et al., 2007; Rezlescu et al., 2012; Schlicht et al., 2010; Tingley, 2014) where images of real people or computer-generated images have been paired experimentally with descriptive evidence to show that facial appearances impacts judgement (Berry & Zebrowitz, McArthur, 1988; Dumas & Teste, 2006; Porter et al., 2010). Although these studies can isolate the effect of the inferred trait, external validity is low and it would be challenging to use this approach to isolate accuracy of the prediction (the effect of changed facial trait on the same person's behaviour).

Moreover, these studies rely on the assumption that a computer-generated face with exaggerated facial features will be equivalent to a real person with

actual social presence (e.g. How likely are you to pick this person as your political candidate?). But we know that non-human faces are not necessarily judged or treated in the same fashion as human faces. If a face image is recognisably not human, but approaches human qualities, they are likely in the *uncanny valley* (MacDorman et al., 2009). The *uncanny valley* (Mori, 1970; see Mori, MacDorman & Kageki, 2012 for an English language translation) refers to a human response which shifts from empathy to revulsion as a humanoid artefact approaches, but fails to attain, lifelike appearance. Given humans' particular sensitivity to face stimuli, computer generated images will tend to fall short of appearing as a true face. Thus, an uncanny computer-generated image attempting to display a certain social inference trait would elicit different responses from an observed as compared to a real face with that actual trait (Dautenhahn, 2007; MacDorman & Ishiguro, 2006).

Here we introduce the use of hyper-realistic face masks as a new means to manipulate social inferences. Hyper-realistic face masks allow experimenters to manipulate rather than measure appearance in experiments, with the focus on real-world behavioural outcomes. Hyper-realistic face masks are particularly interesting for this purpose because unlike props such as glasses or hair changes (which may be used to isolate how they influence judgments, e.g. the nerd defense; Merry, 2012), hyper-realistic face masks are hard to detect (Chapter 1-3) and carry inherent social information (see Appendix 2.3). This allows for an unusual manipulation that could isolate the effect of the face itself for social inference. Most research paradigms in this field assess whether the predictive nature of the face is accurate ('Is the person whose face is trustworthy actually trustworthy?') and whether facial appearance has consequences ('Is the person whose face is trustworthy treated better?'). These paradigms do not allow any form of separation of the social inferences and the effect of character/personality traits. Because realistic masks carry their own social characteristic information, their appearance can be entirely dissociated from the character of the person who wears it. This allows us to invert the usual facial inference task. Rather than assuming that experimental conditions *do not* affect perception of the face and/or perception of a manipulated face (which we know they do, Rosenthal, 1966; Todorov et al., 2015), we can make viewers explicitly aware that the person who

wears the mask is in no way related to the appearance (and social characteristics) portrayed by the mask, and have their task be to *not* let it affect their judgement of the wearer. If we see the same attribution of social traits of the wearer to the appearance of the mask with explicit instruction to ignore the appearance of the mask, we can be certain of the inherent binding facial appearance and the person beneath the face.

If we see inferences of the mask affect judgment of the wearer, masks could be used as a more stringent test of the effect of facial appearance on observer behaviour in real-world manipulation, which can isolate the effect of facial versus character inference.

The findings in the previous two chapters indicated an attribution of the mask to the wearer, even when viewers were instructed to 'see through' the mask when making demographic (Chapter 4) and identity judgements (Chapter 5). Here, we consider in two separate studies, whether the appearance of the mask affects social trait inferences (Experiment 10) and personality judgements (Experiment 11) making the task as explicit as possible.

7.3 Experiment 10:

Social trait judgements of the wearer beneath the mask

In this experiment, we used social trait judgements to assess whether it would be possible for viewers to dissociate their perception of the face – which in case of mask wearing is irrelevant to confederate's social traits – from their judgement of the person beneath it. We used social traits because there is strong evidence for the robustness (Little et al., 2006) and evolutionary roots of attractiveness, dominance and trustworthiness judgements (Zebrowitz, 2004; Zebrowitz & Montepare, 2006). This is indicative of their inherent nature, so we expected that these judgements are particularly persistent and provide a particularly stringent test of the attribution error.

To keep the design simple, we used the image set from Chapter 6, comprising ambient face photographs of two confederates in three different realistic face masks and without a mask. Viewers judged the presence or absence of each social trait.

To assess any impact of the mask, we first needed to establish the appearance of the mask and the appearance of the wearer. Hence, we started with two groups of participants, one judging these images without knowing how the confederates looked, and another made aware of the appearance of the confederate with and without the mask. More importantly, a third group of participants was then asked to judge the wearer through the mask - whilst explicitly requested to ignore the mask. This last group allowed us to isolate whether judgement of the confederate through the mask occurs: is social inference of the wearer through the mask akin to inference made for them without the mask, or akin to inference made of their masked appearance?

We predicted that if there would be no effect of mask, judgements of the wearers would be the same across the 3 mask categories. Alternatively, if the appearance of the wearer was influenced by the appearance of the mask, we would expect to observe consistent differences between masks across both wearers.

Methods

Ethics statement. Ethics approval for all experiments was obtained from the departmental ethics committee at the University of York.

Participants. Ninety members of the volunteer panel at the University of York (23 males; mean age = 27, age range 18–38 years) took part in exchange for a small payment or course credit. All of them were unfamiliar with the confederates, as confirmed by a familiarity check at the end of the procedure.

Design and Stimuli. We used the same image set at in Chapter 6. Participants judged each mask image on social traits (Trustworthiness,

Attractiveness and Dominance) rating each trait in a separate block. The order of blocks randomised, and within each block, the 60 photos were presented in a random order. Given the large number of trials (180 in total), participants were asked to make a simple Yes/No judgement on each trial rather than a Likert scale rating to capture the variability in photos of the same face (Jenkins et al., 2011). Participants were informed that the same masks could appear more than once, and that judgments should be assessed on an image-by-image basis. No time limit was imposed for this task, and each image stayed on screen until a response was made.

We ran this experiment with 3 different paradigms (see Figure 7.1). Participants were randomly assigned to 1) the Unaware paradigm: where they judged the person with the mask, without knowing that there were two confederates; 2) the Aware paradigm: where they judged the person with and without the mask, knowing that there were two wearers and what they looked like, or 3) the Ignore paradigm: where they judged the wearer beneath the mask, and were asked to explicitly to ignore the mask. We first compared the impression of the mask (Unaware paradigm), to the impression of the person without the mask (No mask; Aware paradigm). This sets out the parameters for a possible effect of different masks, and for the wearer through the mask (by comparing the first two conditions to observed in the Ignore paradigm for each character trait).

Task paradigms

In the Unaware paradigm, participants were asked to judge the impression of the mask. They were merely instructed to judge the 60 mask images on how the person looked with the mask, being unaware of the face underneath.

In the Aware paradigm, participants were first asked to judge 20 No mask images for each character trait to capture their impression of the confederates without the mask. They were then asked to judge the traits of the wearers in each mask, in the 60 mask images, without any explicit information on which wearer was underneath the mask to see whether or not these judgements in the Unaware

paradigm would be affected by knowing who was underneath the mask.

In the Ignore paradigm, participants were explicitly instructed to judge the person *beneath* the mask and to ignore the mask itself. To make judging the wearer as easy as possible, participants first categorised 20 images of Mladen and Florence by assigning their confederate names. Participants then judged the 60 Masked images for each trait. For this group, each mask image was captioned with the name of the confederate who was wearing the mask. In addition, participants were given a print out of the 20 images that they categorised in the practice trial, for reference during the experiment.

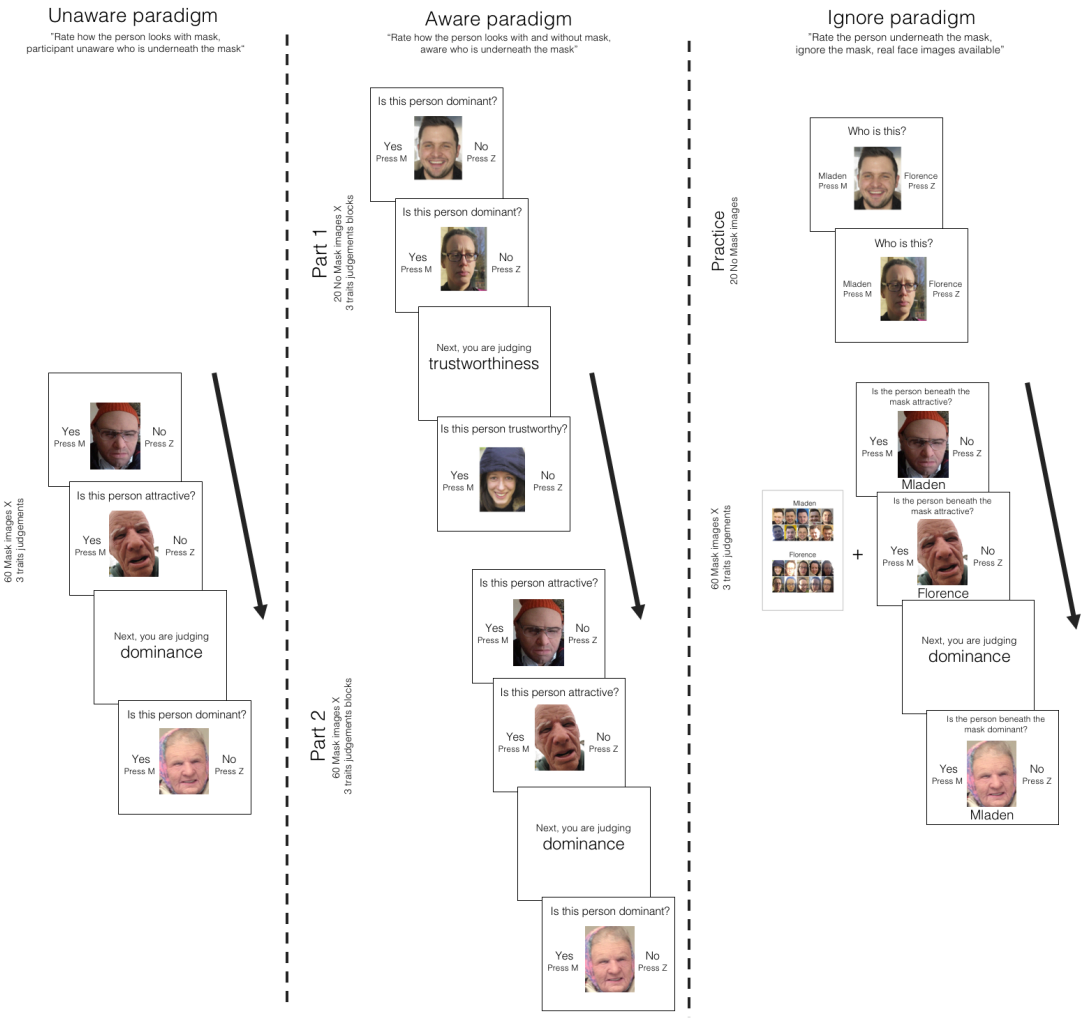

Figure 7.1. Diagram displaying differences between task instructions and trials, by Unaware, Aware and Ignore Paradigms in Experiment 10.

Procedure. Across all conditions, there was a caption instructing which trait was being judged. In the Unaware and Aware paradigms this stated: '*Is this person [e.g. attractive]?'*, and in Ignore paradigm this said: '*Is the person beneath the mask [e.g. attractive]?'*. This caption was immediately above the mask image, with response options 'NO – Press Z' to centre left and 'YES - Press M' to centre right of the image. Participants pressed 'Z' if they disagreed with the statement, or 'M' if they agreed with the statement, which initiated the next trial. Prior to test blocks, participants always completed a practice block for each trait (3 trials),

followed by 3 experimental blocks of 60 Mask image trials per block presented in a random order (180 trials total). In the Aware, this was preceded by an additional 3 randomised blocks of No mask image trials, 20 trials per block (60 trials total). At the end of the experiment, participants were debriefed. The entire experiment took approximately 15-20 minutes to complete.

Results and discussion

For all conditions (Mask x Wearer x Procedure), we calculated the percentage of 'Yes' responses across the 10 mask images per participant by trait judgement. Figure 7.2 summarises these scores for each social trait, separated by Mask, Wearer and Paradigms. For concision, I will only report results relevant for our hypotheses. The following sections set out the same analyses for each of the three social traits in turn.

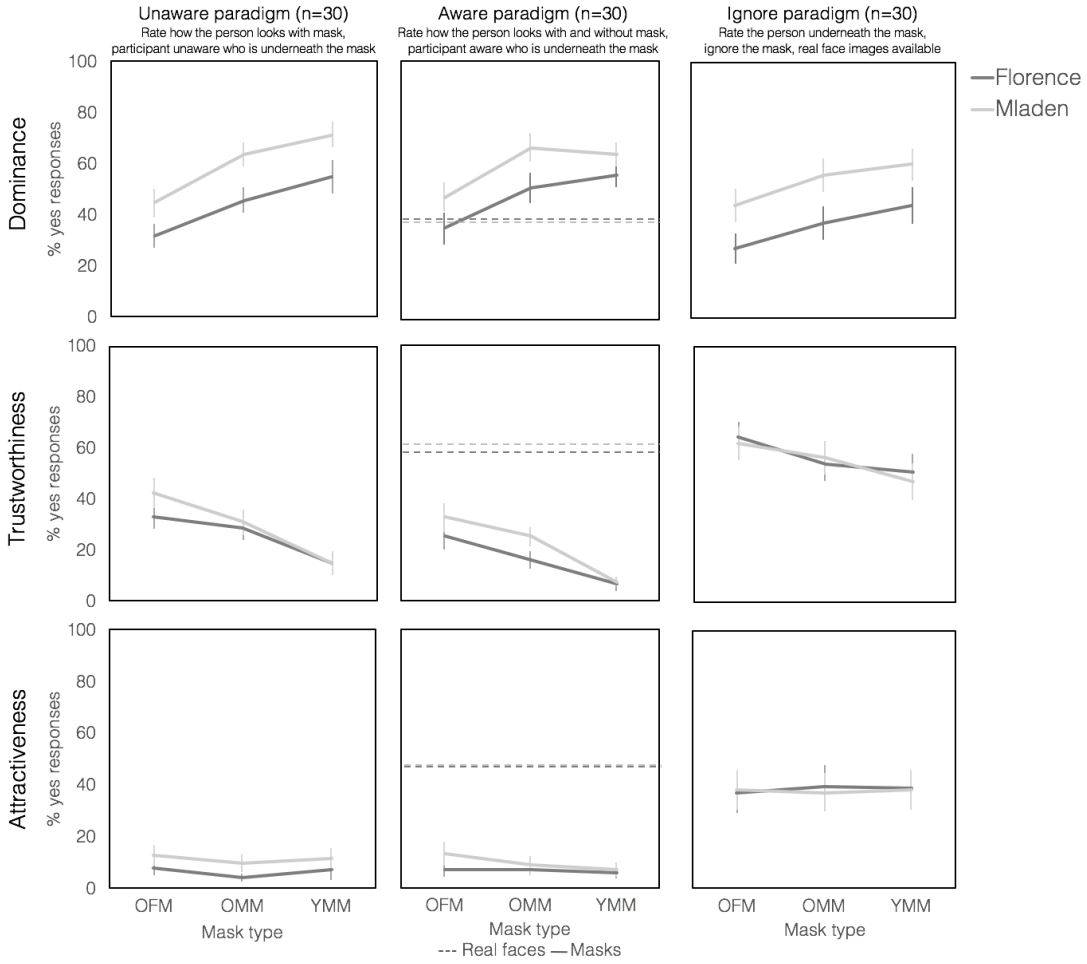

Figure 7.2. Social characteristic judgements (Dominance, Trustworthiness and Attractiveness; rows) of two different Wearers (Florence and Mladen; lines) in three different Masks (Old Female Mask: OFM, Old Male Mask: OMM, Young Male Mask: YMM; x-axis) using three different Paradigms: Unaware paradigm, Aware paradigm and Ignore paradigm (columns) in Experiment 10.

Dominance. Dominance scores were submitted to a 2 × 2 x 3 mixed ANOVA to test for an effect of Mask (OFM, OMM, YMM), Wearer (Mladen, Florence) and Paradigm (Unaware, Aware, Ignore; see figure 7.2). This analysis found a main effect of Mask [$F(2, 85) = 20.28$, $p < .001$, partial $\eta^2 = .22$] and Wearer [$F(1, 86) = 20.5$, $p < .001$, partial $\eta^2 = .19$], but no effect of Paradigm [$F(2, 86) = .439$, $p = .169$] or interaction effects ($p > .05$ for all comparison). For both wearers, participants produced significantly higher dominance scores in the YMM (M =

58.8% yes responses; SE = .27, CI = 53.3 - 64.3, p <.001) and OMM conditions (M = 53.6% yes responses; SE = .27, CI = 48.3 – 58.9, p <.001) compared to the OFM condition (M = 39.0% yes responses; SE = .27, CI = 33.4 - 44.6). Participants also produced significantly higher dominance scores for Mladen (M = 58.1% yes responses; SE = .26, CI = 52.8 – 63.3,) than for Florence (M = 39.0% yes responses; SE = .27, CI = 33.4 - 44.6, p <.001), across all masks. That we do not see an effect of Paradigm shows that the type of instruction given did not impact Dominance judgements.

The Aware paradigm allowed us to compare these scores to the wearer's actual appearance using a 4 x 2 repeated measures ANOVA of Mask (OFM, OMM, YMM, No mask) and Wearer (Mladen, Florence). As expected, the analysis showed a main effect of Mask [$F(3, 84) = 8.21$, $p < .001$, partial $\eta^2 = .23$] and Wearer [$F(1, 28) = 11.26$, $p < .001$, partial $\eta^2 = .29$], but also an interaction effect ([$F(3, 84) = 3.92$, $p = .011$, partial $\eta^2 = .12$]. In line with the attribution error hypothesis, pairwise comparisons revealed that No mask Dominance scores (M = 38% yes, SE = .41, CI = 30.0 – 46.9) were significantly lower from the OMM Dominance scores (M = 58.8% yes, SE = .53 CI = 48.0- 69.6, M difference = -20.3, SE .63, CI = -31.9 - -8.8, p = .001) and YMM Dominance scores (M = 60.2% yes, SE = .46, CI = 50.8 – 69.5, M difference = -21.7, SE = 34.6, CI = -34.6 - -8.8, p = .001), but not those for the OFM condition (M = 41.6% yes, SE = .55, CI = 30.3 – 52.8, M difference = -3.1, SE = 6.3, CI = -16.0 – -9.8, p = .626). Moreover, the interaction effect implies that differences in scores for the wearer underneath the mask must be driven by the masks, as such differences are not apparent in the No mask condition.

A 3 x 2 repeated measures ANOVA of Mask (OFM, OMM, YMM) and Wearer (Mladen, Florence) showed the same pattern in mask Dominance scores when participants were asked to judge the wearer underneath the mask (Ignore paradigm) [main effect of Mask: $F(2, 56) = 7.314$, $p = .002$, partial $\eta^2 = .21$]. To compare these data with judgements of the wearers in the No mask condition we created one score for Mladen and one for Florence per participant, by averaging across the three Masked conditions. We then entered these averaged scores into a 2 x 2 mixed ANOVA with the No mask scores for each wearer in the Aware

paradigm. This analysis revealed no significant difference in scores between the two paradigms [main effect of Paradigm: $F(1, 57) = 2.93$, $p = .092$].

In sum, these data indicate that the average dominance levels of the confederates across Masked conditions are not separable from the No mask conditions. However, each mask very clearly influenced judgements of the wearer's dominance even when viewers were told to disregard the mask's appearance–perhaps akin to an attribution error.

Trustworthiness. Trustworthiness scores were also submitted a 2 x 2 x 3 mixed ANOVA to test for effects of Mask (OFM, OMM, YMM), Wearer (Mladen, Florence), and Paradigm (Unaware, Aware, Ignore). This analysis found significant main effects of Mask [$F(2, 88) = 36.86$, $p < .001$, partial $\eta^2 = .30$] and Paradigm [$F(2, 86) = 19.23$, $p < .001$, partial $\eta^2 = .31$] only, with no main effect of Wearer [$F(2, 88) = 2.43$, $p = .123$], and no significant interaction effects. For both wearers, participants produced the lowest Trustworthiness scores in the YMM condition (M = 23.3% yes responses; SE = .27, CI = 17.9 – 28.7), followed by the OMM condition (M = 34.9% yes responses; SE = .27, CI = 29.5 – 40.4), and the highest scores in the OFM condition (M = 43.3% yes responses; SE = .32, CI = 37.1 – 49.7, $p < .001$ for all comparisons). Participants also produced significantly higher Trustworthiness scores when estimating the person underneath the mask (Ignore paradigm: M = 55.8% yes responses; SE = .44, CI = 47.0 – 64.6) compared with the other two Paradigms (Unaware: M = 26.6% yes responses; SE = .44, CI = 17.9 – 35.2, $p < .001$; Aware: M = 19.3% yes responses; SE = .44, CI = 10.7 – 28.0, $p < .001$).

To compare these scores to the wearer's actual appearance we ran a 4 x 2 repeated measures ANOVA of Mask (OFM, OMM, YMM, No mask) and Wearer (Mladen, Florence) for Aware Paradigm only. As expected, the analysis shows a main effect of Mask [$F(3, 84) = 63.87$, $p < .001$, partial $\eta^2 = .70$] and Wearer [$F(1, 28) = 6.63$, $p = .016$, partial $\eta^2 = .19$], but no interaction effect ([$F(3, 84) = 1.46$, $p = .232$]. More importantly, pairwise comparisons revealed that Trustworthiness scores for the No mask condition (M = 59.8% yes, SE = .29, CI = 54.0 – 65.7) were different from each of the Mask conditions (OFM: M = 29% yes, SE - .52, CI

= 18.4 – 39.6; OMM: M = 20.5% yes, SE = .32, CI = 14.0 – 27.0%; YMM: M = 7% yes, .19, SE = 3.1 – 11.0; p < .001 for all comparisons).

A 3 x 2 repeated measured ANOVA of Mask (OFM, OMM, YMM) and Wearer (Mladen, Florence) showed the same trend persisting when participants judged the wearer underneath the mask (Ignore paradigm) [main effect of Mask: $F(2, 56) = 10.15$, $p < .001$, partial $\eta^2 = .27$]. As with the Dominance analysis, we averaged across the three mask conditions for Mladen and Florence to arrive at one Masked score per participants, which we compared in a 2 x 2 mixed design subject ANOVA with the No mask scores for each Wearer in Aware paradigm. We found no difference in scores between the two Paradigms [main effect of Paradigm: $F(1, 57) = .49$, $p = .486$].

To summarise these trustworthiness findings, judgments of the masked appearance were not dissociable from judgements of the confederates, nor could we dissociate between trustworthiness inferences of the two confederates. However, the inferences made from different masks did significantly influence judgements of the wearer's trustworthiness – again consistent with attribution error.

Attractiveness. Attractiveness scores were similarly analysed via 2 x 2 x 3 mixed ANOVA to test for effects of Mask (OFM, OMM, YMM), Wearer (Mladen, Florence), and Paradigm (Unaware, Aware, Ignore). This analysis only found a main effect of Paradigm [$F(2, 86) = 15.29$, $p < .001$, partial $\eta^2 = .26$], with no effect of Mask $F(2, 86) = .59$, $p = .557$] or Wearer [$F(2, 86) = 1.22$, $p = .272$], and no significant interactions. Participants produced significantly lower attractiveness scores for the Unaware paradigm (M = 8.5% yes responses; SE = .44, CI = -.02 – 17.2) and the Aware paradigm (M = 7.9% yes responses; SE = .44, CI = -.07 – 16.6) compared with the Ignore paradigm (M = 43.3% yes responses; SE = .32, CI = 37.1 – 49.7, p < .001 for both comparisons).

To compare these scores to the wearer's actual appearance we ran a 4 x 2 repeated measures ANOVA of Mask (OFM, OMM, YMM, No mask) and Wearer (Mladen, Florence) for the Aware paradigm only. As expected, the analysis showed a main effect of Mask [$F(3, 84) = 62.58$, $p < .001$, partial $\eta^2 = .69$], but no

effect of Wearer [F(1, 28) = 2.72, p = .110] or interaction effect ([F(3, 84) = .76, p =.520]. Most relevant for our attribution error hypothesis, pairwise comparisons revealed that attractiveness scores for the No mask condition (M = 51.2% yes, SE = .39, CI = 43.2 – 59.2) were significantly higher than for all three Mask conditions (OFM: M = 10.3% yes, SE - .37, CI = 2.7 – 18; OMM: M = 7.9% yes, SE = .29, CI = 1.9 – 14%; YMM: M = 6.4% yes, SE = .26, CI = 1.1 – 11.7; p < .001 for all comparisons).

A 3 x 2 ANOVA of Mask (OFM, OMM, YMM) and Wearer (Florence, Mladen) for Attractiveness scores for the wearer underneath the mask (Ignore paradigm) shows no trend in attractiveness scores either [main effect of Mask: F(2, 56) = .108, p = .898]. Hence, attractiveness ratings do not provide evidence for an attribution error, but this may be due to the lack of diversity in attractiveness among the masks themselves (all were rated low). We see that averaged across the Masked conditions for both Wearers in the Ignore paradigm, compared in a 2 x 2 between-subject ANOVA with the No mask scores for each wearer collected in the Aware paradigm, there is no difference in scores between the two Paradigms [main effect of condition: F(1, 57) = .08, p = .777]. In other words, viewers appeared able to judge the wearer's attractiveness without being affected by the mask's appearance.

Experiment 10 clearly provided evidence for a number of effects. First, Dominance and Trustworthiness judgments of the mask through the wearer indicate that viewers were not able to ignore the social cues that they read from the mask, even when they were irrelevant to the task at hand. In other words, they were binding the physical attributes of the 'face' to the person beneath the face. This suggests for an attribution error in face perception.

Second, judgments for all three traits hovered around the rating of the confederates unmasked. This suggests that viewers were not affected by the appearance of the mask alone, but that they were also incorporating other information. In this case, that information concerned the appearance of the viewer without the mask, but we suggest that the same integration effect could apply to other types of information too (e.g. contextual cues).

We also see that attractiveness judgements remained unaffected by the appearance of the mask. This could indicate that some trait judgements are more prone to face-person binding than others. However, the current data cannot confirm this hypothesis, as we saw very little variation in mask attractiveness in the first place.

In Experiment 11, we see whether these effects replicate with an increased level of abstraction, by having participants judge personality trait of the wearer.

7.4 Experiment 11:

Personality trait judgements of the wearer beneath the mask

When we look beyond the field of face perception, where we mainly think in terms of social inference, the leading model of the structure of personality traits is the Big Five model (see Goldberg, 1993; John & Srivastava, 1999 for reviews). This describes human personality in terms of five dimensions; extroversion, agreeableness, conscientiousness, openness and neuroticism (Goldberg, 1990; McCrae & Costa, 1987). The Big Five model can be used for peer and self-ratings (Goldberg, 1993; John & Srivastava, 1999), and a variety of studies have investigated the judgments of strangers on the Big Five personality dimensions from face photographs, mostly to examine the accuracy of these judgments (Back et al., 2010; Beer and Watson, 2008; Ivcevic and Ambady, 2012; Kramer and Ward, 2010, 2011; Leikas et al., 2013; Little and Perrett, 2007; Penton-Voak et al., 2006; Watson, 1989).

Sutherland et al. (2015) showed how the Big Five related to social inferences by examining how the Big Five personality traits and the major social traits of Approachability/Trustworthiness, Dominance, and Youthful-Attractiveness related to each other in 1000 ambient images. They found that Big Five judgments were separable, but nonetheless related to social trait judgements. Judgements of Openness, Extroversion, Neuroticism, and Agreeableness were all linked to

Trustworthiness, whereas Conscientiousness judgements related to a combination of Approachability and Dominance.

This suggests that personality traits too may be prone to attribution error of the face to the wearer, but may function along separate dimensions. Using the same methodology as in Experiment 10, we aim to isolate the effect of facial appearance on personality inference of the person beneath the face. As in Experiment 10, we predicted that if there is no effect of mask, judgements of the wearers would be the same across the 3 mask categories. However, if the evaluation of the wearer is influenced by the appearance of the mask, we expect to observe consistent differences between masks across both wearers.

Methods

Participants. Ninety members of the volunteer panel at the University of York (31 males; mean age = 27, age range 18–39 years) took part in exchange for a small payment or course credit.

Design and Stimuli. The design was identical to Experiment 10, but participants now judged each mask image on personality traits (Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism), rating each trait in a separate block. In addition, participants were provided with a print out of Costa and McCrae (1992) definitions of each of the traits to refer to throughout the experiment.

Procedure. The procedure was also identical to Experiment 10, but in this case task instructions stated '*Is this person [e.g. extrovert]?'* in the Mask and Aware paradigm, and in Ignore paradigm this said: '*Is the person beneath the mask [e.g. extrovert]?'* Each participant completed a practice block for each trait (5 trials), followed by 5 experimental blocks of 60 Mask image trials per block in a random order (300 trials total). In the Aware, this was preceded by an additional 5 randomised blocks of No mask image trials, 20 trials per block (100 trials total). At the end of the experiment, participants were debriefed. The entire experiment took

approximately 20-30 minutes to complete.

Results and Discussion

A summary of personality trait judgements is shown in Figure 7.3. We used the exact same approach to analysing mask and wearer personality judgements as in the preceding experiment, but in the interest of concision we present only summary findings here. The full results are presented in Appendix 7.1.

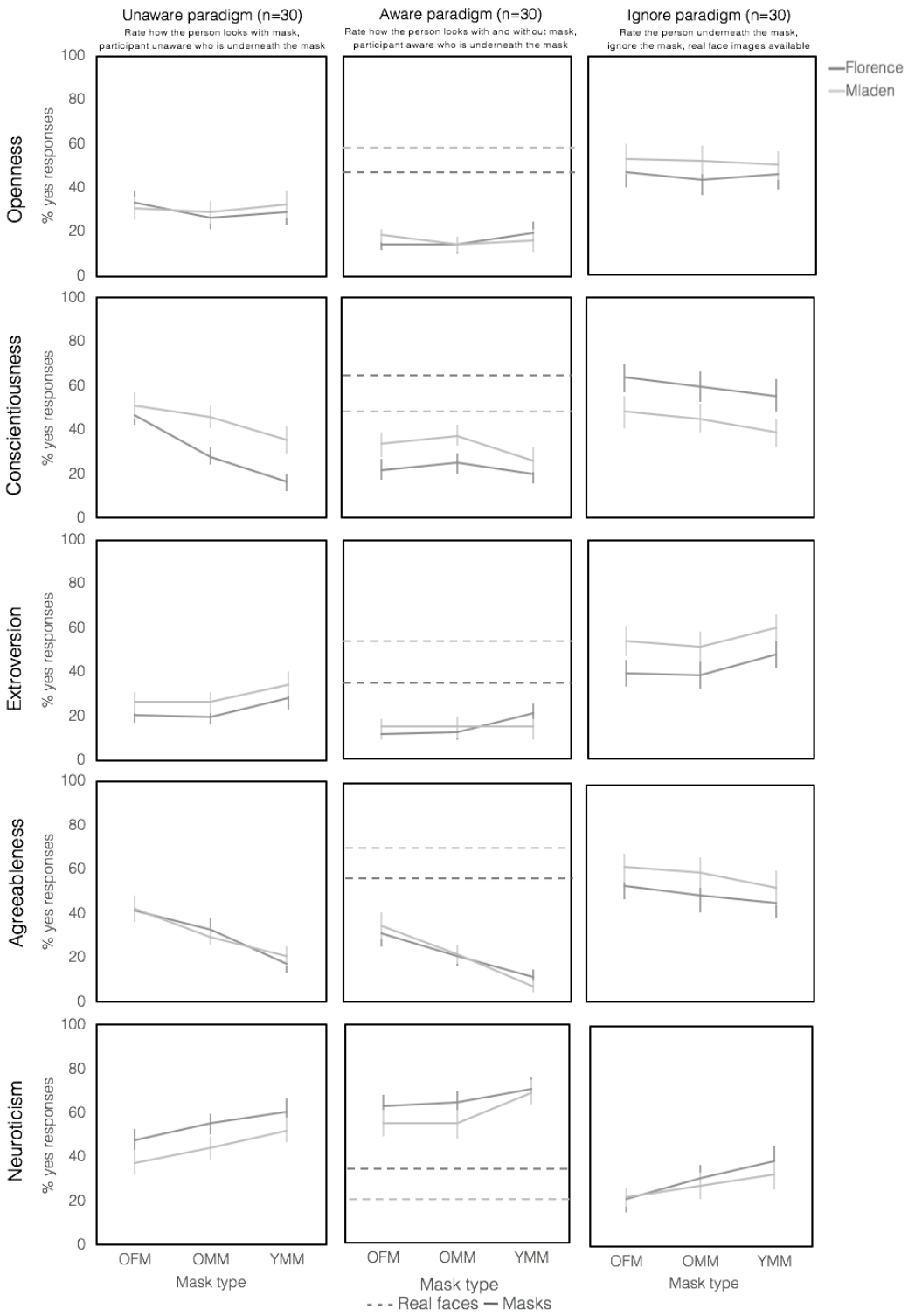

Figure 7.3. Personality judgements (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism; rows) of two different Wearers (Florence and Mladen; lines) in three different Masks (Old Female Mask: OFM, Old Male Mask: OMM, Young Male Mask: YMM; x-axis) using three different Paradigms: Unaware paradigm, Aware paradigm and Ignore paradigm (columns) in Experiment 11.

Openness. The data did not show any attribution of the mask's Openness scores to those of the wearers. However, this may be because Openness ratings did not vary significantly between masks. Covering the face seemed to impair observers' ability to differentiate between the two confederates on Openness (something that clearly comes out when they do not wear a mask).

Conscientiousness. Confederates and masks were clearly separable in terms of perceived Conscientiousness. When judging the confederate's appearance through the mask, viewers could distinguish between the mask and wearer cues. In fact, we observe a reversal between the Unaware and Aware vs. the Ignore paradigm, in that judgements of the person beneath the mask clearly aligned with the confederate, rather than the average of the Masked conditions. In addition – and this is relevant to the attribution of the mask to the wearer – we see a (non-significant) numerical trend in the direction of the appearance of the mask. Even though we clearly see a reversal in judgment in the Ignore paradigm across Masked conditions (where a masked Mladen is judged as more conscientious than a masked Florence, but an unmasked Mladen is judged as less conscientious than an unmasked Florence), viewers were unable to ignore the effect of the mask entirely.

Extroversion. The two confederates were clearly separable in the social inference of extroversion (but not when they are rated in the mask), and so were two of the three masks (OFM and OMM vs YMM). The data show a remarkable merging of these two patterns when viewers focused on judging the wearer beneath the mask. Judgements appear anchored on the No mask condition, and trend in the exact pattern (OFM and OMM vs YMM) as in the condition where the appearance of the mask was rated. In short, when judging the person beneath the mask, viewers distinguished between mask wearers, but were also influenced by the masks.

Agreeableness. Without the mask, the two confederates were clearly separable in terms of how agreeable they look. When masked and judged with the mask, we no longer saw any separation between the two wearers, but we did see a strong effect of mask. As with Extraverson and Conscientiousness, we see that

in judging the person beneath the mask, viewers appeared to anchor on the agreeableness judgement they gave for the unmasked appearance, with a non-significant trend in the direction of the appearance of the masks.

Neuroticism. Interestingly, when judging the confederates with and without the masks the social inference of Neuroticism is dissociable for mask and Ignore paradigms. In judging the person beneath the mask, Neuroticism was similar across wearers, and although viewers were apparently able to approximate the wearer's Neuroticism through the mask, their scores were also significantly influenced by the appearance of the mask.

Overall effects. It is worth noting that across outcome variables we observe a general contrast effect. When viewers rated only the mask images they tended to be more moderate, whilst seeing both the Masked and Unmasked images polarised the two across personality judgements.

In summary, we see that the appearance of the mask is attributed to that of the wearer across different personality traits. Full analyses are presented in Appendix 7.1.

General Discussion

We set out to determine whether viewers attribute characteristics of the mask to the wearer, comparing two types of social inferences: character traits (Experiment 10) and personality traits (Experiment 11). We predicted that if such attribution occurred, we would see scores for the wearer influenced by the appearance of the mask *and* for this pattern to be consistent across wearers.

Across both studies, we see the pattern that we predicted. As the most striking example, Dominance scores of the wearer beneath the mask slope upwards (from OFM, 36% yes responses, to OMM, 41% yes responses, to YMM, 52% yes responses) in similar but separable trends for each of the two confederates. The task instructions seem almost irrelevant: the trend is nearly

identical for the Unaware paradigm and the Ignore paradigm.

The other social inferences measured show very similar patterns, with one key distinction: in addition to an attribution error (visible from the slope across mask conditions), inference judgements also seem to be anchored on the wearer's No mask appearance (as evident in the comparison across paradigms). Extroversion judgements provide a good example. In addition to a trend in the same direction as observed in the Unaware paradigm, Extroversion judgments without the mask (A: 47% yes responses; B: 63% yes responses) were almost identical to averaged Extroversion judgments with the mask (A: 45% yes responses; B: 60% yes responses). This pattern was significant across Trustworthiness, Neuroticism and Extroversion measures. The same trends in Conscientiousness and Agreeableness were not statistically significant, possibly because they are harder to rate from someone's appearance, resulting in larger error bars (Funder, 1995; Petrican, Todorov, & Grady, 2014).

Across all 8 outcome variables, only Attractiveness and Openness judgements showed attribution of the mask to the wearer at all. These exceptions are perhaps not surprising, as the three masks we used did not show much variation in these traits in the first place.

Our findings are highly consistent across mask manipulations in other chapters. In Chapter 5 we saw that the mask influenced demographic judgements, and in Chapter 6 we saw that the gender of the mask influenced judgements of the wearer's identity. Across Chapters 5–7, we see that external characteristics that are known not to describe the person nevertheless affect the perception of that person. This effect applies not only to physical characteristics associated with demographics (age, gender, and race), but also to complex social and personality traits and even across domain. Viewers are apparently unable to ignore this information in the mask, even when they are aware that it is entirely irrelevant to the task. Collectively, this provides strong evidence of the inherent connection between a person's face and their perceived character.

The relationship between facial appearance and character judgement is commonly studied by adding descriptive character dimensions alongside

manipulations of face photographs (Berry & Zebrowitz, McArthur, 1988; Dumas & Teste, 2006; Porter et al., 2010; Reslezcu et al., 2012). Generally, these experiments instruct a participant to focus on character traits that are presented rather than the character portrayed by the face (e.g. money lending history versus a trustworthy face (e.g. Reslezcu et al., 2012) or evidence of a crime versus attractive/trustworthy face (Berry & Zebrowitz, McArthur, 1988; Dumas & Teste, 2006; Porter et al. 2010), but find that the face influences the outcome nonetheless. Such studies are useful in illustrating the consequences of attribution error based on facial appearance, and the magnitude of any effects, but they are not useful for exploring the intrinsic nature of the effect.

The current approach goes beyond previous studies by using an unusual proxy for 'character'. The above studies rely on the assumptions that 1) people want to judge the person for their deeds not the facial appearance (Todorov et al., 2015), but also that 2) people commonly believe that a face displays one's character (Olivia, Funk, & Todorov, 2014). In those traditional studies, there is no direct conflict between the facial appearance and character information provided. In other words, the information is complementary. Participants are not incentivised to resist the influence of the face, which means they are not trying to overcome any bias they naturally experience. In considering the 'inherent' nature of binding between the face and person for the perceiver, we assume that observers could not do anything other than use this mechanism. Hyper-realistic masks overcome these issues by presenting two sets of facial inferences (mask and confederate) in direct conflict with each other. In this situation, viewers must suppress the interference of the mask to be able to accurately complete the task. The current findings show that viewers are not capable of doing so, indicating that the automatic inferences from facial appearance override logic and influence how character is perceived, whether presented as deeds, characteristic descriptions or real faces beneath a mask.

It is worth noting that, considering participants were explicitly asked to resist the information portrayed by the mask, the effects we see are often remarkably large. The biggest difference is for Dominance scores: the range in Dominance scores between wearers (18%) is about the same as the range of

Dominance scores for masked judgements of the wearer (17%). This dissociation is a representative pattern: differences between masks on judgment of the wearer beneath the mask ranges anywhere from 0-17% yes responses, where the effect of the wearer ranges from 0-18% yes responses.

The magnitudes of these effects suggest that hyper-realistic face masks may provide a useful technique for studying a variety of effects in social perception. The masks can be worn as a manipulation of facial appearance that is entirely dissociable from the character of the person who wears it. This manipulation opens the door to studying 1) the effect of facial appearance on observer behaviour in the real-world, 2) the dissociable effects of facial appearance and character on behaviour of the mask wearer.

Chapter 8.

General Discussion

8.1 Overview of Findings

In this thesis, I introduced hyper-realistic face masks to the face perception landscape. In the introductory chapter, I brought to light a number of high profile criminal cases where hyper-realistic face masks were used effectively to disguise the culprits' identity and were seemingly going undetected. I argued that hyper-realistic face masks – if these cases were not anomalies– could raise serious questions for identification practice. I made this case according to the argument that we still rely heavily on identification from face images (e.g. in passport control and everyday social environments), and that this mapping between images and identities assumes that an individual's facial appearance is stable within certain bounds (e.g. age changes over time). Radical changes beyond these bounds (e.g. disguises) are generally easy to spot. Hyper-realistic face masks overturn this assumption by allowing the wearer to look like an entirely different person. If unnoticed, they would break the link between facial appearance and personal identity. This argument gave rise to the research theme in the first half of this thesis: assessing the realism of hyper-realistic face masks under various circumstances.

In Chapter 2, I approached this from a practical angle: Assume that you are in an attentive state, but not expecting to see a mask, would you incidentally detect it? I examined this question in 3 experiments, considering photographic (Experiment 1 and 2) and live viewing situations (Experiment 3). To our surprise, performance was very low (~1% spontaneous hit rate). Viewers generally accepted hyper-realistic masks as real faces. However, this was a difficult task, with the odds stacked against the participant.

I next examined mask realism under more stringent task conditions, by devising a Turing test (Turing, 1950) for discriminating hyper-realistic masks from

real faces. This Turing test allowed for much more favourable to the participant. Viewers simply had to decide which one of two images in an image pair was the mask, with one being a hyper-realistic mask, one a real face. I ran the test with limited (Experiment 4) and unlimited exposure time (Experiment 5), which viewers failed a striking 34% and 18% of the time, respectively. I replicate these results with two adjusted versions of this task as part of Chapter 4, but results only worsened: 60% error rate in Experiment 6 and 26% error rate Experiment 7 with these slightly harder test conditions of inspecting an image one at a time.

As a third test of realism, I checked for an other-race effect in hyper-realistic mask detection, by comparing Japanese and British participants in Chapter 2 and in Chapter 3. The computerised (Experiment 1 and 2) and live detection tasks (Experiment 3) suggested that there was no other-race effect from memory ('What did you think of the person?' and 'Did you notice anything unusual about the person?; all 3 experiments), but that there was when viewers discriminated masks from real faces from images in the array challenge ('Can you pick the mask from the array?'; Experiment 1 and 2 only). I replicated this pattern in Chapter 3, examining performance on the Turing test using a full own-/other-race design with two independent sets of participants. I found a 5% own-race advantage for performance. This outlined the added risk of other-race masks in security context (e.g. such as airport security), it also served as another route to demonstrating the realism of these masks: these synthetic faces are realistic enough not just to be able to trick our face detection system some of the time, but to such an extent that own-race viewers are better at discriminating own-race masks from real faces than other-race viewers.

Using these three methodological routes, my thesis establishes that hyper-realistic masks fool the eye very often, under all tested conditions.

As the obvious next step, I considered means of improving mask detection. In Chapter 4, I assessed untrained expertise or individual differences in an adapted version of our Turing test, revealing large individual differences (5–100%) for High-realism masks among Low-realism masks and Real faces (Experiment 6), which remained when the Low-realism condition was eliminated (Experiment

7). This suggests that we could recruit high-end performers as personnel into settings where hyper-realistic masks could be used.

In addition I used the data in Chapter 2 (Experiment 1 & 2) and Chapter 4 (Experiment 7) to consider whether prior knowledge of hyper-realistic masks may be driving performance, but found no correlation between performance on the task and with prior knowledge of hyper-realistic face masks.

As a final approach toward improving detection performance, I used the data from Experiment 7 in an image analysis of masks and real faces images to capture any physical cues which differentiate masks from real faces, and could be used to train detection. This revealed that mask and face stimuli were most strongly differentiated in the region below the eyes. By comparing high and low performers I could also show that High performing participants tracked the differential information in this area, but Low performing participants did not. Unlike many other face tasks, performance may be localised to a specific image cue.

Resolving the issue of mask detection does not necessarily resolve the problem of that hyper-realistic face masks pose to face recognition. I argued that capturing a culprit who used a hyper-realistic mask relies on 1) detecting the mask and 2) identifying the wearer beneath the mask. The second part of my thesis focused on the latter part of the problem. In the introductory chapter, I characterised the issue of recognition by introducing a high profile criminal case where a bank robber, who became known to be wearing a mask, was profiled through the mask. Considering the high error rates for face recognition accuracy for regular and partial facial disguises (Dhamecha et al., 2014), this called into question what a viewer could accurately estimate beneath a mask.

In Chapter 5, I assessed the accuracy of demographic estimates, of confederates beneath hyper-realistic face masks in a live viewing condition (Experiment 8). Error rates were high across age, gender and racial group estimates, which suggested that profiling of mask wearers should be treated with great caution. As one may expect, recognition too was highly error prone. *How* error prone was a surprise. In Chapter 6 I used the simplest possible recognition task: unfamiliar viewers picked which one of two wearers (one male, one female)

was beneath a mask. Even this highly simplified identification task give rise to a 26% error rate. By comparing these results to those of Experiment 8, we know that this is no better than a ‘gender’ guess.

I also use Chapter 6 (Experiment 9) to test for three means of improving recognition of the person beneath the mask. Assuming that a particular individual is suspected of being beneath a mask, I argued that familiar viewers would outperform unfamiliar viewers in identifying the wearer. To my surprise, the data revealed that Familiar viewers (71% accuracy) did no better than Unfamiliar viewers (73% accuracy). In fact, their performance was significantly worse. I argued that face familiarity is unlikely to be a reliable means of improving performance. More interestingly, face recognition under these conditions of extreme disguise seems to define an upper limit of the Familiarity advantage.

As an alternative method for improving performance, I considered individual differences once again, this time for improving recognition accuracy. I found equally large individual differences for Familiar (43–86%) and Unfamiliar viewers (41–86%). Self-reported familiarity with each of the confederates could explain individual differences in accuracy, but there was no such relationship. Interestingly, there was a strong correlation between accuracy for one confederate and accuracy for the other. I argued that strong performers are good at this task for both confederates, not because of how familiar they are with the confederate, but because they are doing something else consistently across the task. This ‘something else’ could be that they are responsive to certain facial cues which poor performers are not. Chapter 4 provides an early suggestion of which visual cues might be important for this task. I conclude that steering personnel recruitment towards high-end performers could improve performance. Characterising what cues separate strong performers from poor performers is also a potential step towards improving recognition performance across the board (e.g. via targeted training).

In Chapter 5-7, I also argued that for a lack of facial appearance cues through the mask, viewers rely on heuristic information to fill the gaps. I see evidence of a base rate and representative heuristic in Chapter 5 (Experiment 8),

but experimentally manipulate the appearance of the mask and wearer to isolate in particular an attribution error of the mask to the wearer (Chapter 5-7; Experiment 8-11). In Chapter 5, demographic characteristics of the masks influenced estimates of age, gender and racial group of the wearer. This suggests that viewers were unable to ignore the demographic cues provided by the mask, even though they knew these were entirely irrelevant to the prescribed task. In Chapter 6 (Experiment 9), this pattern was replicated, with viewers attributing the gender of the mask to the identity of the wearer. I argued that this cross-domain attribution (gender to identity), which persisted even for familiar viewers, is evidence of the fundamental nature of the process by which inferences from facial appearance are bound to the person in question.

In Chapter 7, I examined this attribution process for social inference judgments. Social inferences are more abstracted from facial appearance than identity and demographic judgements, and there is a lot more debate as to whether they are accurate predictors – unlike for demographic and identity judgements. They are also able to capture more refined responses in support of an attribution bias, in that different masks and wearers tend to occupy different positions along a given social trait spectrum (e.g. the dominance or extroversion spectrum). I indeed observed clear attribution errors of the mask to the wearer across nearly all social and personality inferences. The consistency and strength of this pattern suggests that hyper-realistic face masks could offer a new and improved tool for studying connections between facial appearance and social inferences. I also argued that the observed patterns provide evidence of the automaticity with which viewers bind facial appearance to individuals, even when they know that facial appearance has been manipulated.

8.2. Advancement of the applied problem

To the best of my knowledge there is no previous research which involved hyper-realistic face masks in the field of human face perception. Hence the biggest contribution to scientific progress made by this thesis is merely putting hyper-

realistic face masks on the map.

Our pursuit was largely an applied one. The sometimes comical tone of media reports which covered hyper-realistic face mask crimes (Bernstein, 2010; Henderson, 2016, Stanton, 2015), and the outlets which covered them (e.g. Stanton, 2015; Cox, 2017; Raven, 2015) has often been geared towards sensationalised news that overlooks the more serious implications. By empirically assessing how hard hyper-realistic face masks were to detect, this thesis has introduced hyper-realistic face masks as an issue that needs to be taken seriously.

Face recognition and disguise

The literature that most closely approaches the effect of hyper-realistic masks on face identification concerns deliberate disguise. The few studies which considered the effects of disguise on identification performance mainly showed that identification performance deteriorates. These studies concerned partial disguises (e.g. use of hoodies, beards or glasses; Dhamecha et al., 2014; Kramer & Ritchie, 2016'Righi, Peissig, & Tarr; 2012; Terry, 1993; 1994;) which still allowed viewers to use the unveiled part of the face to aid recognition. Moreover, these disguises are often easy to spot. This thesis stands alone in addressing the effect whole-head disguise, which leaves only the eyes of the wearer uncovered to observers and some face shape cues – obscuring shape and texture cues to identity to a much greater extent.

As a more realistic approach to disguise in applied situations, I relate hyper-realistic masks as a disguise to work by Noyes and Jenkins (under review) addressed impersonation (trying to look like a specific other person) and evasion (trying to look as different from yourself as possible). They asked confederates to use props and make-up of their own choosing to impersonate and evade, and were highly effective in deteriorating performance for unfamiliar viewers. More strikingly, they showed that familiar viewer performance also deteriorated, which they argue approached "the limits of even familiar face identification" (Noyes &

Jenkins, submitted). The recognition task in Chapter 6 appears to exceed those limits, with the familiarity advantage for face recognition in case being entirely abolished with hyper-realistic face masks as a disguise. In fact, unfamiliar viewers outperformed familiar viewers in this particular study. We would need to replicate these findings to be certain, but in relation to the Noyes and Jenkins (under review) study, our observations suggest the possibility of an inverted U-shaped function for the familiarity advantage in identification, set out against increasingly obstructive facial disguise. In other words, it is possible that in the specific conditions tested here, the performance of familiar viewers does not plateau on par with unfamiliar viewers with increasingly obstructive disguise, but instead deteriorates in comparison to performance of unfamiliar viewers.

I argue that unlike regular familiar viewer face recognition, with masked recognition a viewer's familiar face processing mechanism – largely based in holistic processing – can no longer be relied upon. *Knowing* that you are a familiar viewer, may increase reliance on holistic processing. Holistic processing incorporates all available cues: a few valid cues to the wearer and many non-valid cues (e.g. the mask's appearance, or the local base rate). This approach may leave familiar viewers more exposed to bias than unfamiliar viewers, who may home in on a smaller subset of facial features (e.g. eyes and face shape).

It seems clear that the challenge of identifying individuals through hyper-realistic face masks exceeds the limits of the standard face familiarity advantage. This is an unusual finding, with theoretical implications for our understanding of the familiarity advantage. It also means that – unlike nearly every other face identification task – face familiarity cannot improve identification in case of recognition through hyper-realistic face masks.

Face detection and visual discrimination

As well as recognition through a disguise, I also examined detection of the mask, as a first step towards overcoming the disguise and identifying the wearer.

I expected that hyper-realistic face masks would trip the basic mechanisms we use to detect regular faces, as they imitate the kinds of facial colouration and configural information that have been shown to be important in successful face detection (Bindemann & Burton, 2009).

I argued that distinguishing realistic masks from real faces would rather be a process *following* the belief that a face has been detected, akin to discriminating stimuli in visual search, reliant on top-down serial searching rather than specific features guiding attention (Wolfe & Horowitz, 2004). I didn't study mask/face discrimination using a visual search paradigm in this thesis, but our data certainly supports serial search pattern. In fact, in the Array Challenge in Experiments 1 and 2, nearly 3 real faces were mistaken for masks in arrays of just 19 faces and 1 mask on average. This suggests that even close inspection was not enough to solve the mask detection problem. Experiments 4 and 5 outline the contrasts in discrimination approach of realistic masks/real faces to that of regular masks/real faces. Reaction times in Low realism mask discrimination from real faces, which – being brightly coloured and different in shape – suggest that they rely on a feature search. It took just one second per decision under unlimited viewing conditions, with near perfect accuracy. Not having such features available to aid discrimination of hyper-realistic mask, this task requires nearly a second longer, and left one in five realistic mask images to be mistaken for real face images.

Performance enhancement

Feature search

It is clear that diagnostic features of hyper-realistic masks do not naturally jump out to most viewers. However, mask detection in Chapter 2 informed us that high spatial frequency must be important, as spontaneous detection increases from the Far to the Near viewing condition (Experiment 3). The image analysis in Chapter 4 further revealed a region immediately under the eyes that distinguished high performers from low performers in this task (Image Analysis). Much as radiographers may use features in X-ray images that non-experts do not notice

(Drew, Vo and Wolfe, 2013; Miles-Worsley, Johnston & Simons, 1988), it is possible that guiding attention to facial cues that separate masks from real faces could build expertise and enhance feature-based searching. It would be interesting to attempt to identify more such features, and in particular to assess whether these could lead to higher accuracy rates. I propose that visual search paradigm would be ideally suited to this purpose.

Uncanny Valley cues and congruency

I also suggested that the '*uncanny'* cue (Mori, 2012; official translation from Mori, 1970) might be exploited to improve performance. Spontaneous reports in Experiment 1-3 suggest that hyper-realistic face masks may be perceived as *uncanny* in this sense. Despite their close resemblance to faces, and despite remaining undetected, they seemed to cause some viewers some unease (see Appendix 2.1 and 2.2 for example comments, and 2.3 for social characteristic rating differences between masks are real faces). I noted the uncanny valley response is likely triggered because the layer of silicone on the face eliminates some subtle facial movement and attenuates others, resulting in an impression of blunted animacy. I also anticipated that inanimacy alone would not be a reliable diagnostic, as high false alarm rates in the array challenge of Experiment 1 and 2 suggest that this cue may also encourage false positives for low-animacy real faces. To pursue the potential of this cue for improving performance, I 1) suggest comparing facial expressiveness of confederates with and without mask to establish a baseline and 2) obtain expressiveness judgement for the guided detection (Experiments 1 and 2) and in the Turing test images to compare across hits, misses, false alarms and correct rejections rates (Experiments 4 and 5).

The findings in Chapter 5 suggest that congruency between different social signals could provide a complementary cue (Campanella & Belin, 2007; Johnson, McKay, & Pollick, 2011; Meeren, Van Heijnergen, & Gelder, 2005; Montepare & Zebrowitz-McArthur, 1988; Van den Stock, Righart, & De Gelder, 2007). In Experiments 8, body cues appeared to influence perception of the wearer beneath the mask, especially for gender and age, leading to above-chance profiling

performance for those demographic traits. We attribute these effects to hands, body shape and eyes providing usable cues. Whether incongruency can actually be detected spontaneously is an interesting question. Much previous work on incongruency detection and face perception concerns auditory and visual cues in the context of crossmodal integration (e.g. McGurk & MacDonald, 1976). This literature suggests that a viewer will attempt to integrate incongruent information, hence it could only serve as a cue to detection if differences are large enough. This uncanny incongruence (e.g. black facial appearance with white hands, a sprinting senior, or a male with high pitched voice) could boost mask detection.

In fact, one could argue that in the long run, incongruence, whether in terms of auditory or visual cues, is the most promising route to improving performance. The quality of realistic face masks will presumably continue to improve, overcoming uncanny valley triggers such as inanimacy (e.g. by using thinner and more durable silicone ensuring a closer seal to the face) and improving where the mask breaks for the wearer's eyes (e.g. by thinning silicone gradually, blending skintone towards the edges of the mask). In fact, these changes are already being pursued by silicone mask producers (see spfx.com; facecompositeeffect.com in their most recent generation of masks).

Congruency effects, on the other hand, will always remain. These effects may become more subtle as masks get better, but if the purpose of the mask is to drastically change appearance, the mask's appearance will inherently be different from that of the wearer, including movement, vocal cues, and even gender. The experiments in Chapter 7 provide evidence of such subtle cues being detected, as personality and social inference judgements allowed viewers to separate between two masked wearers, even when they were unaware that different individuals were beneath the mask (Experiment 10 & 11, Unaware paradigm).

Individual differences

Finally, I looked to see whether we could exploit individual differences to aid detection performance. It comes as no surprise to see individual differences in

mask detection, as nearly all tasks in face perception show some variability in performance. Previous research has captured a broad range of performance in recognition studies, from 'super-recognisers', who rarely make errors (Bobak, Hancock, & Bate, 2016; Robertson et al., 2016; Russell, Duchaine, & Nakayama, 2009), to people developmental prosopagnosia who rarely exceed chance performance (Behrmann & Avidan, 2005; Duchaine & Nakayama, 2005). We observed a very similar range of individual differences in mask detection performance, but we cannot tell from the data whether we are tapping into the same skill set.

Research has shown that strong performance on one type of task does not necessarily predict strong performance on another (Noyes, Hill & O'Toole, 2018). For example, some super-recognisers are good at recognition faces from memory, whilst others are better at face matching. Face matching relies on fine discrimination principles. I established that hyper-realistic mask detection does too. It seems plausible that super-recognisers who excel at facial image comparison might also be especially good at realistic mask detection.

At the same time, some aspects of our data suggest mask detection could be a separable skillset. Previous research has shown that recognition (Towler, White, & Kemp, 2014, 2017; White, Kemp, Jenkins, & Burton, 2014), and similar stimulus discrimination (Wolfe & Horowitz, 2004) is very hard to train. That we identified a physical marker used by high-end performers in mask/real face discrimination in Chapter 4, suggests that the possibility for training could be different for realistic mask detection.

Chapter 6 too I saw that there was a strong relationship between recognition of one confederate with recognition of another, whilst we saw no relationship for familiarity with each confederate and recognition of that confederate. This pattern too suggests that there are high-end performers who must be doing something *other* than using regular recognition.

One suggestion is that the skills underlying successful mask detection and successful recognition of the wearer are related. A plausible explanation is that both scenarios require a clear separation between what is the wearer and what is

the mask, e.g. attention to unusual seams in the face or which components of the face carry more animacy cues than others. It would be interesting to pursue this research by comparing performance across different hyper-realistic mask tasks, and regular face perception tasks to home in on the position of hyper-realistic masks processing in relation to regular face processing.

8.3 Advancement of theoretical problems

So far in this discussion, I have considered advancement of our findings in relation to applied face recognition, face detection and visual stimulus discrimination, and performance enhancement, and discussed how theoretical advancement could contribute in these applied contexts. I also identify two additional theoretical contributions from this work.

Face Space Theory and Clustered Expertise

First, hyper-realistic face masks highlight that a face, non-face vector should be incorporated into Face space Theory (Valentine, 1991). Face space Theory is a multidimensional space within which all newly perceived faces or familiar faces of a new appearance (e.g. a new haircut, or weight loss) are stored and grouped according to identity. I used this theory as a framework for understanding the effects hyper-realistic face masks could have on facial identification. The model assumes that within-person variability in a face clusters in a predictable (lighting, viewing angle, change in appearance over time) and separable manner from between-subject variability. Noyes and Jenkins (under review) proposed how the impact of impersonation (attempting to invade another identity face cluster) and evasion (attempting to escape your own identity face cluster) affected face space theory. However, the regular disguises considered still expect certain boundaries to possible change. For example, effectively impersonating opposite gender or a different racial group is unlikely. Considering that hyper-realistic face masks are going undetected and allow much quicker

transformation into a person much more distinct (evasion), and allow more realistic impersonation of an individual with more precision than with other disguise types, they transform the potential of facial disguise. In sum, hyper-realistic masks overturn the basic assumption that identity is bound to facial appearance unless a highly refined face/non-face assessment vector – at least theoretically – is integrated into the model to ensure that hyper-realistic face masks do not infiltrate these clusters.

This is important because Face Space theory is not just used to understand face recognition and identification, it is also used to model self-learning algorithms (e.g. Moon & Philips, 2001; Furl, Philips, & O'Toole, 2002), which allow the extraction of identity clusters from a submission of face photographs, with potential for high impact, if used in automatic face recognition systems.

Second, I argue that hyper-realistic mask detection and recognition of the wearer provide support that face processing is organised according to clustered levels of expertise which advantage performance along separable dimensions – a key component of Face Space Theory (Valentine, 1991; Valentine & Endo, 1992). Whether each dimension can provide an added advantage is a contentious area, which is either explained according to the above perceptual expertise account (which suggests it can), or according to a socio-cognitive in-group/out-group labelling account, which suggests there is merely one layer of advantage (either you are in, or you are out).

I argued above that hyper-realistic face masks require a fine-tuned face/non-face dimension to be part of face space. Our individual differences in mask detection (Experiment 6 and 7) show that face/non-face discrimination is a cluster of expertise, as some participants were better at this task than others. Demographic cues carried by hyper-realistic face masks (age, gender and/or racial group) – as additional dimensions – then align or misalign with this expertise clusters in addition (see figure 8.1).

One could argue that the face/non-face discrimination advantage vector is different from, age, gender, and racial group advantages, because the latter are

all predicted by an observer-described proxy (e.g. my age predicts my ability to recognise individuals within my age group better), which we have not yet identified for mask/face discrimination. Considering that the Image Analysis (Chapter 4) identified a cue which is learnable, it is possible that there are such signature individuals. A plausible option is that individuals who work in special effects, as make-up artists or even produce types of silicone could be at an advantage over other people. Regardless of what may drive this cluster of expertise, there must be one from having seen high-end performers in Experiment 7.

In Experiment 4 and 5 I also saw an other-race effect. I argue that this is evidence of a second cluster of expertise, overlaying the prior. As Chapter 4 replicates, a very similar task to the one used in Chapter 3, we have no reason to expect the same individual differences would not replicate. In turn, I argue that the own-race expertise cluster provides additional, and dissociable advantage to the advantage we observed in for regular face/non-face discrimination in high-end performers.

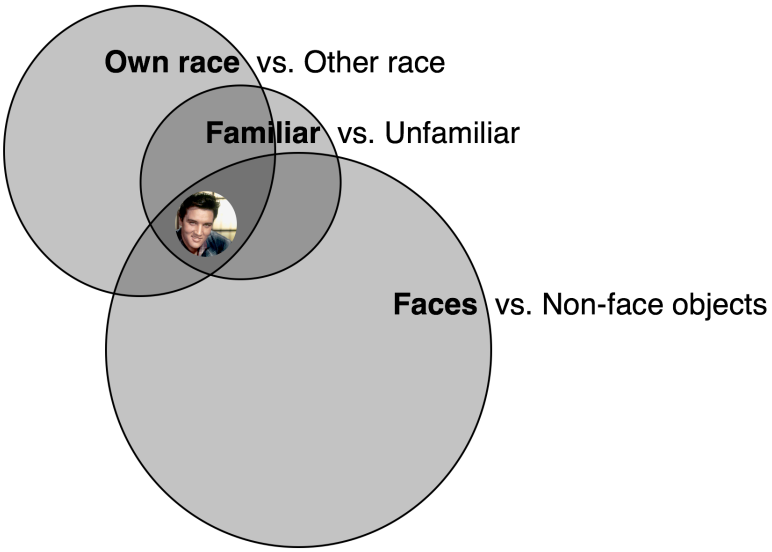

Figure 8.1. Illustration of clusters of expertise having overlapping benefit to recognition accuracy.

This general principle is not new. There is evidence that own-race, own-age and own-gender biases could have similar overlaying advantage (figure 8.1; see

Valentine, 1991; Valentine and Endo, 1992; and see Bernstein et al., 2007; Sporer, 2001 for opposing view). What is new is the potential benefit this could have in discrimination tasks for couterfeits, which suggests that this may be an interesting strand of research to pursue. Under this account race, gender and age group advantages could serve as a tool to supplement realistic mask detection. I have discussed the possibility of recruiting high-performing individuals to benefit detection performance (White et al., 2014; Robertson et al., 2016). This clustered account suggests that recruiting high performing individuals of the 1) same age, 2) same racial group and 3) same gender as the suspected mask wearer could significantly advantage detection performance. This could be remarkably useful, especially whilst we are still unclear on what is driving high-end performers to perform so well.

Attribution bias and social inferences

I also used hyper-realistic face masks to address whether there may be overextension of disguise properties to the wearer. This seemingly unusual question started from an applied perspective, but with theoretical implications. I argued that if the demographic profile or identity of the mask is attributed to that of the wearer, we need to be aware that this attribution may bias profiling and recognition. Across Chapters 5-7 I find support for this attribution bias, where the demographic and social information of the mask indeed influenced demographic, identity and social inferences of the wearer beneath the mask.

From a theoretical perspective, I argue that this is a new variety of bias to add to the family Attribution Theory biases. Attribution theory concerns “how the social perceiver uses information to arrive at causal judgement” (Fiske & Taylor, 1991), and the observed effect most closely aligns to being a Fundamental Attribution Error (FAE). The FAE is the tendency to explain perceived behaviour disproportionately through dispositional or situational factors (Malle, 2006). The FAE of realistic face masks fits best in this category, although we rather see attribution of situational factors (mask) to dispositional factors (wearer). In addition, the FAE of realistic face masks seems to be rooted in a perceptual rather

than social attribution, with the mask providing visual cues, not cues to social interaction.

I suggested that this addition to the Attribution Theory family could contribute to the debate on the 'fundamentality', and 'error' (Harvey, Town and Yarkin, 1981; Malle, 2006) of the FAE. Showing that an FAE is apparent in this perceptual task highlights how fundamental the FAE must be to cognitive processing, and in turn to social interaction. Similarly, the attribution of the mask to the wearer, is unarguably an error (participants are aware that the mask is highly irrelevant to their task of judging the wearer).

As an extension of this effect, I inspected whether attribution of the mask to the wearer could be used to study social inferences in face perception, their accuracy, and their consequences. Our results show that the appearance of the mask is consistently attributed to the social and personality inferences made of the wearer (Experiment 10 and 11).

Testing for the attribution bias of hyper-realistic face masks allows a merging of these two fields of research. A key component of FAE research is that social actors are made aware of all the factors which might affect their judgement, but are then nonetheless influenced by the manipulation at hand. As outlined in Chapter 7, social inference research is currently limited by correlational data, preserving observer ignorance (assuming that the research setting does not affect them) and unrealistic manipulations. I think that our findings in Chapter 7 illustrate that the use of hyper-realistic face masks, combined with the standard FAE research protocol, might be able to help the social inference research to progress in determining accuracy and consequence of facial social inferences.

Finally, attribution of the mask to the wearer serves as new evidence of the inherent binding between the face and the person's character. I argued that the clarity in what is an 'error' in this task, allows for an exceptionally stringent test of this observer's perceived relationship. Based on the data in Experiments 8-11, we can infer that observers are unable to judge an individual without their facial appearance affecting that judgment. We can conclude that the face biases human perception of a person's identity.

8.4 Applied future directions

The fact that hyper-realistic masks are accepted as real faces, at least part of the time, strains the connection between facial appearance and identity, and may, if left unchallenged, require adjustments to our legal (Bruce, 1988; Burton et al., 1999; Davis & Valentine, 2009; McCaffery & Burton, 2016; Wells & Olson, 2003; White et al., 2014;) and social identification systems (e.g. Gosselt et al., 2007; Vestlund et al., 2009). Currently, hyper-realistic masks are relatively uncomfortable to wear and also expensive (see facecompositeeffects.com; Bhattacharjee & Marcel, 2017). For the time being, these considerations probably limit widespread use. I expect that their quality and comfort will only improve as materials develop. This will possibly affect the demand, which will drive down the cost, and increase their market. In fact, since the first reported cases in 2010, new masks have become lighter weight – as a proxy for material improvement; see facecompositeeffects.com) and have reports of their use in criminal settings have become more frequent (see Appendix 1.1).

Currently spontaneous human detection of hyper-realistic masks is highly unreliable (Experiments 1-3), and even guided detection (Experiments 4-7) is poor. The limitations of profiling and recognition of the wearer are also very clear: these processes are systematically biased and give rise to many errors (Experiment 8-11). All of these limitations leave plenty of scope for improving hyper-realistic mask detection and recognising the wearer beneath the mask. This thesis has identified a number of possibilities, which, if validated, could form the basis of a training program for security personnel. The most promising possibilities include: 1) body congruency check, 2) vocal congruency and articulation check: e.g. ask a question 3) inanimacy check: e.g. ask a question/make them smile; 4) seams immediately under the eye region. Although all these areas need further testing to be certain.

Although training in facial image comparison has generally had disappointing results (Towler, White, & Kemp, 2014; 2017; White, Kemp, Jenkins, & Burton, 2014), training in detection of synthetic items (e.g. money, drug, gems and signatures) has been comparatively successful (Fernandez, Green & Newton,

2008; Jonker et al., 2006; WHO, 1999; Mitchell, 1934) and cost effective (Green & Newton, 2008; Jonker et al., 2006;).

It should be noted that hyper-realistic mask detection, and especially recognition of the wearer will be much harder if only captured footage is available (e.g. photographs and video), as single, static photos present much less information than dynamic, live faces (Jenkins & Burton, 2011. Nonetheless, some of the detection techniques suggested above could also apply in such scenarios (e.g. eye seams and inanimacy cues).

In addition, benefits of a crowd analysis should be considered (Galton, 1907; Krause et al., 2011; White, Burton, Kemp, & Jenkins, 2013). In applied contexts, this may point towards working on mask detection and recognition of the wearer in pairs or small groups (Dowsett & Burton, 2015). Ensuring diversity in these pairs/small groups could moreover counteract the applied consequences of other-race (Experiment 4-5), and possibly other-age and other-gender biases, by diversifying the untrained expertise captured amongst observers, and could reduce effects of some of heuristics we discussed to affect recognition of the wearer in Chapter 5.

I also suggested recruiting high performing individuals for both mask detection and recognition of the wearer could benefit performance. Alternatively, one could test amongst already recruited personnel whom may best suited to detect masks using our Turing test (Chapter 3) and/or recognise wearers using our identification task (Chapter 5). This would have the added bonus of allowing experimenters to test whether performance advantage on these two tasks are related (Megreya & Burton, 2007; Vokey & Read, 1992).

Mask detection may also benefit from non-human processing. A recent paper by Manjani et al., (2017) has examined mask detection from a computer vision perspective, albeit using mask images collected under much less stringent conditions (see SMAD image database; Manjani et al., 2017). They considered an algorithm to discriminate realistic images from real faces, and have provided a starting point for an automatic detection algorithm.

Thermal imaging based on infrared cameras has been suggested as complementary technology (Bhattacharjee & Marcel, 2017). These methods are currently used by some border agencies (e.g. Japan and US; Kong et al., 2005) to discriminate between animate and inanimate objects (see figure 8.2). Although costly, research shows these developments have also been effective in combatting smuggling (Andreas, 2003) and may therefore serve multiple purposes. In fact, a recent study by (Bhattacharjee & Marcel 2017), considered the use of infrared and thermal imaging as a simpler alternative to an algorithmic solutions 3D mask attacks (not just hyper-realistic face masks), and suggests that they both serve a promising route towards simplifying detection.

Figure 8.2. Regular (left) and infrared (right) view of the same image, illustrating effective differentiation between animate (skin) and inanimate objects (glasses) through heat reflection.

Finally, sales restrictions through licensing may reduce the availability of hyper-realistic masks for criminal use. Sales restriction has been effective in some other categories. For example, the use of firearms in crime is lower in the UK (with stricter policies) than the US (with less strict policies; Kates & Mauser, 2006). An Australian case study has shown that introducing legislation to restrict sales to those unfit to own a weapon (e.g. with a criminal record), has effectively reduced firearm misuse (Warner, 1999). Similar results might be expected with silicone masks, since their demand is lower than firearms in the first place.

8.4 Theoretical future directions

I discussed numerous theoretical aspects of mask perception throughout this discussion, mainly with the aim of improving understanding of mask perception by recruiting understanding of face perception. I also discussed potential avenues for attribution error and social inference research to continue.

Here I finish with the most important theoretical advance this thesis offers to the scientific community: synthetic faces have been shown to pass for real faces most of the time. This provides research opportunities for live manipulation of the face in any field that is affected by facial appearance.

In Chapter 7 I validated hyper-realistic masks as a useable tool for advancing research in a variety of directions. This could involve manipulations of facial expression, demographic traits, character traits, personality traits and even identity.

For example, it could be used to dissociate the effect of character and facial appearance on the behaviour of the observer in live test situations. Facial appearance could be systematically manipulated towards different social, emotional or personality traits and the observer response could be measured systematically in laboratory settings (e.g. gaze; Gobel et al., 2015) or in real-world situations (e.g. sentencing, job applications etc.). This could be highly complementary to the laboratory work, and correlational work that has already been done to show these effects.

One applied strand of work that makes use of this possibility is a new training program for nurses called Mask-Ed (see figure 8.2; Reid‐Searl, Eaton, Vieth, & Happell, 2011), where the use of hyper-realistic face masks to increase realism of their patient interaction training (Reid-Searl, Levett-Jones, Cooper, & Happell, 2014). The program has expanded to 3 univeristies, as has led to significant improvements in nurse performance in student confidence, and their performance in engagement with real patients (Curtis et al., 2016; Mainey, Dwyer, Reid-Searl, & Bassett, 2018).

Figure 8.3. The Mask-Ed interacts with a nursing student (Left); The Mask-Ed educator removes the mask to begin the debriefing process (Right). Images retrieved from Reid-Searl et al., (2014).

Finally, an exciting opportunity to use hyper-realistic face masks would be to explore contentious areas in social psychology. This area is currently at the heart of the replication crisis, and hyper-realistic mask could in some cases provide for a strong manipulation which was previously unavailable. Take the example of embodied cognition. Embodied cognition relies on the argument that humans are highly attentive to facial information, and use facial appearance to make inferences about the psychological states of others (e.g. Hareli et al., 2002). Our findings on the attribution error showed that hyper-realistic masks follow this trend. The contentious component of the theory is whether this relationship is bidirectional. To provide a simple example: when we are happy we smile (non-controversial), but when we smile it also makes us happy (e.g. Strack et al., 1988) This effect has been replicated in numerous forms (e.g. power posing: Carney et al, 2010), but not always with the same results (Adam & Galinsky, 2012; Bargh, Chen and Burrows, 1996; Dijksterhuis & Van Knippenberg, 1998; Williams and Bargh, 2008; Frank & Gilovich, 1988) and 'miss' (Wagenmakers et al, 2016; Doyen et al., 2012; Lynott et al., 2014; Ranehill et al., 2015; Womack et al., 2016) .Failures to replicate may not be surprising when findings rely on small, frail manipulations (e.g. holding a cup of hot coffee result in 'warmer' responses; Williams and Bargh, 2008). This is a complex theoretical issue where hyper-

realistic face masks could provide for the drastic manipulations in facial appearance – e.g. rather than contracting a few facial muscles, a change in age, gender and racial group – could truly advance the field. Hyper-realistic masks could be used to test for a range of effects of perceived gender, age, race, emotion and character on psychological state and behavior, and ultimately even assess the therapeutic and potential of such effects.

For now, I have introduced hyper-realistic face masks to the field of face perception, and shown how their realism raises new questions for applied face identification as well as new possibilities for research. As synthetic faces become more realistic, I expect that they will further strain the relationship between facial appearance and identity, with consequences in social and legal domains.

Appendix 1.1 Typology of hyper-realistic mask uses in criminal settings

Case	Year	Country	Crime	Wearer demographics	Mask demographics	Mask Detected?	Culprit Caught?	Link
1	2009	U.K.	Jewellery heist	Young white males	Young white males	No	No	https://bit.ly/2LbsJIY
2*	2010	U.S.A.	6 bank robberies	Young white male	Young black male	No	Yes	https://dailym.ai/1eq9vwa
3*	2010	Hong Kong /Canada	Illegal border crossing	Young Asian male	Old white male	No	Yes	https://bit.ly/2uFrHEr
4*	2011	U.S.A.	16 bank robberies	Unknown, profiled	Old white male	Yes	No	https://bit.ly/2O7A04c
5	2011	U.S.A.	11 bank robberies	Young white male	Young white male	No	Yes	https://cbsloc.al/2L9XRIT
6*	2012	U.S.A.	Bank robbery	Young black males	Young white males	No	Yes	https://dailym.ai/2mzGXOR
7	2012	U.K.	14 bank robberies	Young black male	Young white male	Unknown	Yes	https://bit.ly/2LuZx8K
8	2013	U.S.A.	False imprisonment	Young white male	Old white male	No	Yes	https://bit.ly/2Lf8GDh
9	2014	U.S.A.	Bank robbery	Unknown, profiled	Old white male	Yes	No	https://bit.ly/2LbakFU
10	2014	U.S.A.	Bank robbery	Unknown, profiled	Old white male	Yes	No	https://bit.ly/2A0XIvS
11	2014	U.S.A.	Bank robbery	Young white male	Old white male	Unknown	Yes	https://bit.ly/2JlJooH
12	2014	U.S.A.	Bank robbery	Young white male	Young Asian male	No	Yes	https://bit.ly/2LBrSY1
13	2014	U.S.A.	Bank robbery, assault, murder	Young black male	Old white male	No	Yes	https://bit.ly/2LuKOLc
14	2015	U.S.A.	Bank robbery	Young black male	Old white male	No	Yes	https://bit.ly/2uGdsPD
15	2015	U.S.A.	Bank robbery	Unknown, profiled	Old white male	Yes	No	https://bit.ly/2uRHW0s
16	2015	Canada	3 bank robberies	Young black male	Young white male	No	Yes	https://bit.ly/2JGVZvp
17	2015	France	Jewellery heist	Unknown, not profiled	Old white male	Yes	No	https://bit.ly/1cg2HaO
18*	2015	Brazil	Prison escape	Young Hispanic male	Old white female	Yes	Yes	http://bit.ly/2bnHbo7
19	2015	U.K.	Jewellery heist	Young white male	Young white male	Yes	Yes	https://bit.ly/2myhKEf

20	2016	U.S.A.	5 bank robberies	Young white male	Old white male	No	Yes	https://bit.ly/2O5lF72
21	2016	U.S.A.	Fugitive, Drug trafficking	Young white male	Old white male	Unknown	Yes	https://bit.ly/2mweNUR
22	2016	U.S.A.	2 bank robberies	Unknown, profiled	Old white male	Yes	No	https://bit.ly/2zWJjRc
23	2016	Australia	ATM scam	Unknown, not profiled	Unknown	Yes	No	https://bit.ly/2uV5S31
24*	2016	Sweden	Kidnapping	Young white male; Young white female	Old white male; Old white female	No	Yes	https://bbc.in/2O7oAxl
25	2017	U.S.A.	Bank robbery	Unknown, profiled	Old white male; Old white male	Yes	No	https://bit.ly/2OazB0S
26	2017	U.S.A.	Bank robbery	Young white male	Old white male	No	Yes	https://bit.ly/2LBV94O
27	2017	U.S.A.	2 bank robberies	Young Asian male	Old black male	Unknown	Yes	https://bit.ly/2O7ZUES
28	2017	U.S.A.	2 bank robberies	Unknown, profiled	Old black male	Yes	No	https://on-ajc.com/2uElFnp
29	2017	U.S.A.	Store robbery	Unknown, profiled	Old white male	Yes	No	https://bit.ly/2Lfl0TR
30	2017	U.S.A	Conspiracy, kidnapping	Young white female	Old white male	No	Yes	https://bit.ly/2JH7G5c
31	2017	Israel	7 robberies	Young white male	Old white male	No	Yes	https://bit.ly/2uPMUul
32	2017	U.S.A.	Murder	Old white male	Old white male	No	Yes	https://bit.ly/2JHVctW
33	2018	U.S.A.	2 bank robberies	Unknown, not profiled	Old white male	Yes	No	https://bit.ly/2LERFOL

***** High profile criminal cases with detailed case descriptions in Appendix 1.2

Table 1: Typology of criminal cases where hyper-realistic face masks were used based on non-exhaustive search terms combining: 'silicone, latex, realistic, mask, Hollywood, old man, old woman, black man, white man, white woman, disguise, crime, robbery, border control, plane, asylum seeker' from 'Google News'. Note 1: Web Searches were only performed in English. Note 2: For realistic masks to be covered by a media outlet, they had to have been detected or caught. This is not representative of the number of cases that have gone undetected and uncaught.

Appendix 1.2 Descriptions of high profile media reports on hyper-realistic mask use in crime

Conrad Zdrierak (Case 2)

White male Conrad Zdrierak targeted 4 banks and a pharmacy wearing a black male face mask produced by mask company SPFX. Eyewitnesses confirmed that the perpetrator was black, and the surveillance footage showed a black male (Figure 1). The investigators initially held a black man resembling Zdrierak's masked appearance in custody for several weeks prior to Zdrierak's arrest. Some eyewitnesses even identified the black male as the perpetrator from a security photograph. The media even reports that a mother of an Afro-American male identified her son as the culprit from a circulated surveillance photograph of Zdrierak wearing the mask. Zdrierak was caught because his girlfriend found two masks in his possession and handed him in (Gardner, 2010; Damani 2014).

Figure 1. Security footage of Conrad Zdrierak wearing black male face mask produced by mask company SPFX (right) and at his hearing prior to his arrest (left). Image retrieved from: https://bit.ly/2mxbQmV

Korean Refugee (Case 3)

A Korean refugee (Figure 2, left) managed to board a flight from Hong Kong to Vancouver using an elderly male mask (Figure 2, middle) with a real passport. The mask wearer passed several identity checks at Hong Kong airport, only to be discovered as he took the mask off mid-flight. One passenger did notice that the man had unusually young hands for his age. This suggests that border security, employed specifically to inspect passenger identity, did not manage to detect the mask in real time. Hong Kong airport acknowledges their failure to detect the mask, but also confirmed not to have made any mistakes. Border service officers reported that he resembled and behaved like an old man when he wore the mask (Telegraph, 2010).

Figure 2. Korean refugee (left), wearing realistic mask (middle and right) upon arrival at the Canadian border. Image retrieved from: https://bit.ly/2dX5jzw

‘Geezer Bandit’ (case 4)

An unknown male robbed at least 16 banks in the California Bay area wearing an old, white male hyper-realistic face mask, aliased the ‘Geezer Bandit’ he remains on the FBI most wanted list in 2018. Their wanted poster stated that the demographic profile of the robber was that of a ’60-70 year old male’ (see Figure 3). At the 16th robbery a dye pack exploded and video footage shows the robber run out of the bank. Perhaps this cued the FBI, whom subsequently, with the help of a silicone mask maker adjusted this estimate to a ‘20-40 year-old’ male (Weisman, 2018).

Figure 3. A wanted poster for the 'Geezer Bandit', issued by the FBI in 2010. Image retrieved from: https://bit.ly/2Lzlws8

'Mac the Guys' (Case 6)

Three black males Akeem Monsalvatge, Derrick Dunkley, and Edward Byam dressed as white police officers using variations of the 'Mac the Guy' mask produced by CFX to burgle a cash-checking store in Queens, New York City (Figure 4). Upon questioning, eyewitnesses of the crime reported collectively that the perpetrators where white (Epstein, 2012; David, 2014; Algar, 2013). The investigators eventually traced the perpetrators through means unrelated to the masks.

Figure 4. Security footage from cash-checking store burglary in Queens, New York City (left) displaying black males dressed as white police officers using the 'Mac the Guy' masks produced by CFX (right). Image retrieved from: https://nyp.st/2uSwmSN

Clodoaldo Antonio Felipe (case 18)

Clodoaldo Antonio Felipe, a 44-year old male, attempted to escape Coronel Odenir Guimaraes Prison in central Brazil wearing a realistic mask of an elderly female (Figure 5; Stanton, 2015). Felipe managed to pass multiple checkpoints, before he was stopped just after his exit. A guard noted that he never admitted any elderly females for visitation. Despite officers watching Felipe pass in real time (see Figure 5, middle image), this case shows that only once officers suspected something amiss did they detect the mask. Unlike other realistic mask cases described, viewers of the mask would have had more time to inspect the mask wearer, under stress-free and good lighting conditions.

Figure 5. Old female mask (top left) was worn by Clodoaldo Antonio Felipe (right) to attempt escape from Coronel Odenir Guimaraes prison in central Brazil. Middle image shows Felipe passing multiple unsuspecting guards. Image retrieved from: http://bit.ly/2bnHbo7

‘Swedish Fritzl’ (Case 24)

A swedish doctor, under the alias ‘Swedish Fritzl’ was found to have kept a woman as a sex slave in a home-made bunker, whom he – before going on a second date – drugged, raped and kidnapped a women in her thirties, using two realistic face masks: an old female face mask for her and a bearded old male mask for himself and matching identification (see Figure 6). He transported his victim in a wheelchair to his car, wearing the mask to make the 350 mile drive from his flat in Stockholm to a remote farm where he planned to keep her for several years. There are no cases of report on any unusual sightings. The masks were only discovered upon his arrest (Malm, 2016; Henderson, 2016).

Figure 6: Old female mask without its wig (right) and bearded male mask with matching identity card and driver’s license (left) worn by Trenneborg and the victim of his kidnapping. Image retrieved from: https://ind.pn/2A8aCbF

Appendix 1.3 Production and typology of hyper-realistic mask

Realistic silicone masks are high quality face masks made of flexible and durable silicone, cast over a fitted head-form. These masks are also called: latex or rubber masks, hyper-realistic masks (Gucci, 2014; Bernstein, 2010) realistic disguises (Becker, 2010; SPFXmasks.com) or real flesh (Bernstein, 2004; RealFleshmasks.com). The masks were originally made to replace hour-long make-up sessions for movie-production purposes (Gucci, 2014; Mazzuki et al., 2015). As production lines improved and demand grew, they became available to the wider market through online sales. They are now freely available to buy for between £400 to £2000 depending on the company and services requested (Gucci, 2014; Bernstein, 2010; Mazzuki et al., 2015).

Types of masks

Various manufacturers in North America (SPFX, CFX, RealFlesh, Studio135, Hyperflesh, Immortal masks, Odditymall and Real-F) and in Asia (Guangzhou Angel Company; Trxmask, Guangzhou Usilicone.,Ltd, XIXILI, YiRong EYUNG and RealFace) produce different masks. Masks allow natural hearing, vision and breathing. Each mask is finalised by hand, with highly detailed permanent paint works. Upon request they also come with strand-by-strand hair punching of real human hair or an attached wig. Masks can be separated into 3 types. Most masks cover the head and base of the shoulders. This type of mask fits closely around the face of the wearer and allows visible muscle movement through the mask. They come in two sizes (53.98-59.69 cm or 51.44-55.25 cm), but can be adjusted with a beanie cap for padding if the fit is uncomfortable (Mazzuki et al., 2015). They come with an integrated nose bridge, eyeholes that tuck seamlessly under the eyelids and a ‘mouth cupping system’ allowing the mask to fit around the lips into the mouth. Companies producing these masks include Realflesh, SPFX and

Composite Effects. A few other companies produce masks that cover the face only (RealFace), or the head only, in a skull-like shape (Hyperflesh). This type tends to be slightly larger in size and is not as fitted to the skin. Arguably these are comfortable for longer periods of time, yet tend to have a constant facial expression. It is also possible to obtain silicone mask pieces, allowing various configurations to be attached to the face directly using silicone glue (Studio135). This type requires make-up appliance *after* putting on the mask pieces, whilst the other masks do not. Although the latter requires more effort and make-up experience, it is likely more comfortable than the prior two, and could allow more visible muscle movement.

Of masks that are available for sale, the majority appear to be white males, aged 40 and over. For the six top selling companies (SPFX, CFX, Real flesh, Ganzhou Angel and TrxMasks) approximately 75% of masks appear Caucasian, 8% appear Afro-American and 16% appear Asian. Only 16% of masks appear female. Approximately 36% of masks appear to be over the age of 65. It should be noted that the repertoire of masks on sale is constantly updated and expanded.

Masks can also be modified in multiple ways. Firstly, at least half of companies offer customisation of skin colour and texture. They also offer the addition of personalised characteristics such as scars, moles, and hair colour, hairstyle, facial and chest hair (see Figure 1). The same companies also offer the services to manufacture a customised mask from scratch. This was done for a 2013 episode of the pop-science television show Mythbuster's, who found their personalised masks to perform strikingly well in a quasi-experimental detection study for unfamiliar and familiar viewers (see https://vimeo.com/59708532). Secondly, personalisation of the face can be done at home using silicone glue (Mazzuki et al., 2015). Thirdly, the masks interact with the face structure of the mask wearer (Figure 2). Finally, these masks can be used as a base for regular disguises with hats, wigs, sunglasses and clothing to assume a variety of characters. CFX/Composite Effects mask manufacturers found that in 3 minutes an experienced mask wearer was able to make 21 mask changes (average 8.57s per mask; Mazzuki et al., 2015) See http://bit.ly/2d293zD for a demonstration.

In sum masks can vary in an indefinite numbers ways, allowing the adoption of an even wider variety of characters within high time constraints.

Figure 1: Same mask with different hairstyles (left and middle) and different skin tones and adjustment of the nose (middle and right).

Figure 2: Masks look different on different wearers. In order from left to right: Wearer 1 mask 1, wearer 2 mask 1, wearer 1 mask 2, wearer 2 mask 2.

Experience of realistic silicone mask wearing

Based on social media networks on Facebook, Twitter and Instagram the population of users is estimated to be between 5,000-10,000 (popular forums include: Realistic Silicone Masks, Silicone mask sickos, Silicone Mask Addicts, Silicone Mask Community, Silicone Maskerade, #compositeeffects, #realisticmasks, #siliconemasks). Another estimate, based on production Figures quoted by mask manufacturers SPFX (Becker, 2010) and Google Search hits for other companies, estimated sales are around 2500-3500 masks per year worldwide.

A book by Mazzuki et al. (2015) is the only known publication to discuss the use of hyper-realistic masks. The book describes a social experiment conducted using CFX silicone masks:

> *"For an extended time, one of the authors wore a full-head human-like silicone mask while sitting in a dental office waiting room with a magazine in hand. Nobody using the room looked upon him with curiosity" (pp.134)*

Mask users on social media echo this experience, where videos and pictures of mask wearers in every day locations such as supermarkets (http://bit.ly/2czUkHF), hairdressers (http://bit.ly/2d28qFV) or outside (e.g. http://bit.ly/2cCR8OU). None of the bystanders show any sign of awareness of the use of realistic masks despite some abnormal behavioural display. This suggests that there is community of realistic mask wearers who seem to be going undetected under every day circumstances.

Realistic masks seem to be used by three demographics. Professionally, Mazzuki et al. (2015) highly recommends them to undercover investigators. Other professional uses include quick Hollywood makeovers (e.g. as used for popular television series Game of Thrones; http://bit.ly/2cpFTLT) and a recent report describes realistic masks to be used for interactive teaching of nursing students (Burns, 2015). Secondly, realistic masks are used by individuals comparable to a cross-dressing population (Styles, 2014) or for Halloween, party and cosplay dress up (Gucci, 2014) to adopt a different character. Thirdly, realistic masks are used as a disguise of identity in criminal settings. This thesis focuses on the use of masks by this third demographic only.

Appendix 2.1 Responses to open questions UK

Responses to open question ('What do you think of the faces you saw?') and prompted question ('Did you notice anything unusual?') in Experiment 1

P.No	Open response	Prompted response
2	i am female so for me, when i saw female faces, i think they are more trustworthy. and male with normal or smiling faces are also trustworthy. and one' s eys can reveal whether they are trustworthy or attractive.	one, why there were no children faces. two, it is odd that one female had freckles in her face, but i can see she had done some makeup. third, two of the old men looked very similar but just different facial expressions, one smiled one didn' t, which the former of looked trustworthy and attractive and the latter one reversely.
3	They all had similar expressions	No
4	Some of them were fairly easy to rate, facial expressions weighted some of my judgement. I have noticed that there was more older people picture displayed than that of younger people.	Some of the pictures looked like the colouring has been changed or was just a bit different to normal when the photo was taken. I didn' t notice any particular case where I would find a face unusual
5	There was a wide range of people to rate regarding their attractiveness and age.	No
6	people between 50 to 80 seem to be more trustwprthy. those who are smiling seem to be more attractive and trustworthy. people with a cold face seem to be dominant.	some faces with too many freckles and wrinkles may affect judgment directly.
7	they all looked like typical members of the public you would encounter on the street	quite a few of them were older
8	the faces seemed relatively trustworthy, i didn' t find any of the faces	no
9	they were a range of ages, gendars and ethnicities. Smiling ones looked friendlier so more trustworthy. hard to guess their age	Angry ones tended to be on a darker background
10	i did not notice anything unusual	
11	The majority were quite elderly so I rated their looks mostly according to their eyes which tend to be the most attractive features in elderly people. I tend to find people who are smiling naturally or behaving calmly to be trustworthy so I rated faces wish such attributes accordingly. I did not see much in the way of diversity when looking at the faces and I tend to find European faces generally less attractive than other ethnicities.	Many were elderly which I was not expecting in an experiment asking us to rate attraction. I saw little in the way of ethnic diversities.
12	they were very varied and i found it quite hard to judge the older peoples ages	There was one face that I thought looked about mid 30' s and was bald, his face didnt look fully human, it had a plastic aspect to it
13	those people have different facial expressions which may influence my judgement about their age or attractiveness. most of the faces are young people or senior people, the middle aged are rare. overall, those faces are ordinary.	there is a girl with dimple on her face which is unsual to me

14	most photos made the person come across as very friendly and kind, therefore most faces seemed trustworthy	no
15	A lot of them were quite old and with some it was difficult to decide because they were pulling faces	Not really, they all looked quite normal
16	they were very different from one another	there were a lot more older faces
17	i thought one of them was a young person in old people' s make up	yes, i noticed that one of the images of a bald man was probably a younger man dressed in make up so that he looks older. i typed his age as 40 although the make up suggested he was older
18	varied expressions, mostly young adults or elderly, no middleaged faces. varied in approachable look due to smiling or stern expressions. most seemed to be happy, smiling or content	one face obscured by shadow, angle of faces varied e.g more stern looking tilted head upwards to narrow eyes.
19	found the more attractive faces to generally appear less dominant, and the older faces to appear more trustworthy	nothing unusual
20	They were diverse	Some people were making faces in them
21	older people are more trustworthy, young people are more attractive.people who have beautiful faces are more dominant.	a girl who has freckles is not attractive.
22	v	not really
23	A mix of many faces of moderate attraction, most seemed nice but some seemed less nice	Not really, other than some faces had odd expressions
24	a range of expressions from smiling and warm, to cold and glaring.	Only that some pictures were taken with flash, nothing about the faces themselves
25	The faces were composed of a variety of ages, gender and expressions, so the people smiling looked more friendly and therefore more trustworthy.	Some of the faces were in darker lighting so they looked more untrustworthy
26	They varied a lot	Some of them looked as though they were posing for a photo but a couple looked as though the photo was taken in the moment
27	not many faces are smiles	the texture of the skin, some faces have beautifual eyes
28	The faces seemed to either express happiness or anger. They seemed to be either young, middle aged or old. All faces were of one race.	One picture was very pixelated which made the face look like a drawing.
29	I thought that they were, on the most part, having their faces measured by the expression that they had on their face at the time of the photograph as this affects my perception of trustworthiness etc.	I did not notice anything particularly unusual about the faces themselves, no.
30	very wide range of faces from all different age groups, the hardest part was trying to work out their age particularly the elder participants.	one man had extremely small eye, unaware as to how this may have been caused.

31	they seemed to be either young or quite old	difference in age
32	most of them are just general public, some of them may be a bit of aggressive.	Some of the pictures are not good for ' judging', half of the face in shadow or in an angry expression or only three quarters of the whole.
33	I thought that the older people rated as more trustworthy and the younger people as more attractive but that was rather obvious	no not really
34	There was a very diverse and interesting collection of faces.	A couple of the faces were partially obscured by shadow. One girl had a vast amount of freckles on her face.
35	Most trusted either the older or kinder faces, faces that were either covered or at an angle were much harder to judge. Faces with glasses more trustworthy but only really shown on older people.	Faces with freckles was hard to give answers for other than that again covered or angled faces were hardest to judge
36	ranged in age nd gender, they varied in scale of trustworthy, attractive and dominant . elderly people seemed to look more trustworthy and men seemed more dominant	different expressions and different ages either young or old
37	They were very diverse, and interesting. Some faces were much harder to guess ages, eg older faces. I found that the younger faces seemed a lot more attective and most of the time more dominant than the older faces.	The only unusual things that I noticed were things like when people had no teeth, glasses or were a different ethinicity to previous faces. However i do remember noticing that one girl had a lot more frecles on her face.
38	different faces from different age groups	not really
39	smiling faces were generally more attractive, imperfections could be endearing and therefore make the face seem attractive, but sometimes work in the opposite way. a genuine smile also makes the person seem more trustworthy and less dominant	one man had very close together eyes, one woman had very freckly skin, one young man had quite bad acne, one of the older men had no teeth and one very pink shiny skin. one old lady had very wrinkly skin
40	they averaged from 20 to 80 years of age. they also varied with their attractiveness, dominion and trustworthy	no
41	Tend to rate older adults as more trustworthy. Hard to determine age of elderlies.	Some of the faces are not natural, e.g. photoshopped to change the brows, add more freckles, add more contrasts etc. One of the picture of a female elderly is highly altered e.g. added wrinkles, more contrasts, brightness and hues altered etc.
42	I think that it was difficult to judge the trustworthiness and dominance although age was ok. They are all different people. There weren' t many different ethnicities of people.	They weren' t all pulling the same face so this may have had an impact on my perception of their dominance and trustworthiness. The pictures were all tightly cropped to not allow anything other than their face impact my decision.
43	There was a wide range of faces varying in age, gender and ethnicity. I felt there was more old faces than younger faces. I felt also that my ratings depended on the facial expression they were displaying e.g. if the blonde woman was pulling a cross face she would have got a lower rating on trustworthiness than i gave her despite all her other features being the same.	Just that some of them were pulling faces. Also that one girl who had loads of freckles all over her face. I also felt that there was more older faces than younger faces. I felt that when they were older i was more likely to view them as trustworthy than if they were younger.

44	It was difficult to rate them without thinking about how they would look if they had different expressions on their faces. I tried to imagine them differently, especially for trustworthiness, which probably brought my answers more to the middle.I don' t remember seeing any noncaucasian faces. The older people made me smile more often, and I rated their attractiveness according to what I would think were I of the same age range.	Nothing unusual...
45	Mixed bag, no discernable characteristics on an aggregate level	Nope
46	the attractiveness of the face can affect the trustworthy of this person. the more the face looks nice, the easier the people trust he or she. old people with smile also are easier to be trusted.	one of them has lots of black spots on her face, it makes people feel scarey. it makes her less trustworthy.
47	they were mostly of older people and the faces showed emotion and personality to a small extent	some faces were especially fierce looking and intense
48	the most of the faces i saw are old people and they seems more trustworthy than young people.	most of the faces i saw are usual but one face with freckle is bit unusual because it is too much.
49	some of the faces are very attractive with a sunny smile, comfortable eye contact. faces that wear glasses and without smile seem to be more serious, dominant and trustworthy.both of the two kind of faces seem to be trustworthy.however, faces that carry strange expressions seem to be less reliable.	yes, some of the faces looks very comfortable while some faces wear expressions that look very sneaky.
50	Sample set seemed slightly skewed towards older ages. Lighting in one of the samples was obscured, potentially leading the viewer towards a certain response. It is difficult to ponder on the dominance and trustworthiness of others based on a snapshot alone.	See previous answer about obscurity. Also, one or two samples appeared as though they had suffered from some disease or other, melanoma and other cancers perhaps. As such, I tried to reduce my age prediction based on this likely making the sample appear older than usual.
51	a good mix of male and females, mostly smiling. Old people always seem sweeter and nicer and happier somehow. very few middle aged people though. Mostly young or old. That, or I cannot tell people' s age.	brighter photos tended to be happier and darker photos tended to be frowning or angry looking. Not sure whether this may have affected trustworthiness.
52	A large range of different faces, some of the expressions made me a bit uncomfortable so that I wanted to answer questions quicker. I noticed that there was a bigger range of ages than there was of races.	Some of the old people were quite scary, and people who were looking straight at the camera seemed more dominant. The less attractive people were less trustworthy.
53	They were very varied, most of them were trustworthy and there were a lot of elderly people.	There were a lot of older faces, and a few had a specific expression, whereas most were neutral or just simiLing
54	The elderly faces were more difficult to determine the age	no not really
55	the younger faces were easier to guess the age than older faces, facial expression on the pictures influenced my responses in all 3 areas	none of the young faces looked very happy, only the elderly people looked happy. Older faces were of similar ethinicity, younger faces were more diverse, with different facial expressions too

56	i thought they were all were different ages and that they all pulled similar expressions for the camera.	one picture of an old lady had a different sort of colouring to the others.
57	i may have responded differently to the faces that were pulling an aggressive expresion if they had a more neutral one.	no
58	They were all showing some sort of mood expression and had generally quite distinctive features. The majority seemed to be relatively old.	Some of the faces were what society would think were normal, and some were not. For example an elderly women with bad teeth and a younger person with freckles. However there was nothing particularly unusual which I could notice.
59	they seemed very generic faces all what you would imagine someone to look like given a descriotion none had unsual characteristics	no
60	All very different when it comes to mood, clothing, hair colour, but most of them look like nice persons, except from one which looked like a drawing.	One of the faces looked like a drawing, and looked like a criminal.
61	There was a wide range of faces, but it was hard to compare one' s trustworthiness to another for example because of different expressions and situations.	There seemed to be a few very posed faces which may influence people' s opinions of dominance etc. These contrasted a lot with the more nautral photos. Also the differing quality in the photos used. Nothing unusual directly about the faces per se.

Appendix 2.2 Responses to open questions Japan

Responses to open question ('What do you think of the faces you saw?') and prompted question ('Did you notice anything unusual?') in Experiment 2. Translated from Japanese (original response in brackets).

P.No	Open response	Prompted response
1	Facial expression and atmosphere (*hyoujou to huinki*)	Nothing (*nanimo*)
2	Whether they smile or not (*Egaokadouka*)	Nothing (*Tokuninasi*)
3	I felt that they smiled a lot (*Egaogaooinatokannjita*)	I haven't noticed anything particularly (*Tokunihakidukanakatta*)
4	Nothing (*Tokuninanimo*)	English written by participant: No
5	Age, whether they are dominant or not, and whether they are attractive or not. (nenrei,shihaitekimiryokuteisinraidekirukadouka)	
6	I think Eastern people looked younger than Western people did, so I was considering this when I guessed age (*touyoukei no hito wa seiyoukei no hito yori wakaku mieru keikou ni aru to omou node sonokoto wo kouryo sinagara nennrei wo suitei sita*)	Nothing (*toku ni nakatta to omou*)
7	Whether I have a good impression about her or him (*koukanwomoterujinbutsuka*)	I did not realise anything (*kigatsukimasendesita*)
8	Whether there were face lines, or the size of the face. How sagging the face was. Facial expression. Or the sincerity of facial expressions. For age, I looked at face lines and how sagging the face was. For attractiveness, dominance and sincerity, I judged based on my sense of beauty and intuition (*kao no siwa no umu, ookisa . kao no tarumi. hyoujou. mata hyoujou no seijitusa. watashiha nennreini tuiteha kaono siwa ya tarumi no katachi de, miryoku ya sihaido, seijitusa si tuite ha watasi jisinn no biteki kankaku ya chokkan de handanshimashita*)	Not at all. I did not realise anything (*not at all. mattaku kidukimasendeshita*)
9	What kind of people the (presented) person was	Some photos were processed

	(*donna hito nanoka*)	(*syashin ga kakou sareteiru mono ga aru)*
10	Impression at a glance (pattomitekanjitainsyou)	Nothing (tokuninasi)
11	The identity (something translated as: 'true self') of the person (*Sonojinnbutunosujyounado*)	I did not realise anything (*Kidukanakatta*)
12	Older people looked less dominant than younger people did (*Toshiwototteiruhougashihaitekidehanasasou*)	Nothing (*Tokuniarimasenn*)
13	Facial expressions and atmosphere (*kao no hyoujou ya funiki*)	I did not notice anything at all (*mattaku ki ga tsukanakatta*)
14	The first impression (*Saisyonoinnsyou*)	Some people have a freckled face (*Sobakasugatakusannaruhitogaita*)
15	Age and occupation (*nenrei to syokugyou*)	Nothing (*Nashi*)
16	Whether I would trust this person or not. Something like this, I judged based on my criteria (*jibundattarakonohitowosinnraisurudarouka,nadoto,jibunwo kijunnisitekangaemasita*)	I thought some photos were processed. (*mononiyotteha,kakousareteirunodehanaika,tokanjita.*)
17	Whether facial expressions were calm. I think I judged people's impressions based on the sharpness (?) of eye stare (h*youjougaodayakasoukadouka,metukinosurudosadetanin nnoinnsyouwohanndannsiteiruyouniomou*)	There was no child photo (*Kodomonosyasinganakatta*)
18	Facial expression, impression (*hyouzyou innshou*)	There were two age populations (*Nennreisougahutatsuniwakareteiru*)
19	If the person were a teacher, I can trust or rely on what the person said etc. (*moshi sonohitotachi ga sensei dattara to kateisite ,sonohito no iukoto wa shinnyoudekiruka toka izonshiteshimauka nado wo*)	Nothing (*Tokuninashi*)
20	How I feel (*Jibunngadoukanjitaka*)	There were a lot of weird facial expressions (*kimyounahyoujougaookatta*)
21	I looked at people's eyes and judged when I evaluated each person (*Hitonohyoukawokudasutokimazumewomitekanngaeta*)	There were only Western faces and no Asian ones (*Seioujinnnokaobakarideajiakeinokawoganakatta*)
22	Mainly, I judged with wrinkles.	I did not notice anything

	(omonishiwawomitehandanshita)	(kidukimasendeshita)
23	The older the person was, the more s/he looked dominant. To such extent, they get less attractive. (*oiruhodosihaitekinimie,soredakemiryokutekidenai*)	Only a few faces had neutral expressions (*Magaodeututteirumonogasukunai*)
24	I guessed people's personality based on their impression (*sono hitono innshoni motoduite sono hitono seikakuwo yosou shimashita*)	I did not notice anything. (*tokuni kidukimasenn deshita*)
25	Young woman's face was blotchy (wakaijyoseinokaogashimidarakedatta)	
26	I imagined their expressions in daily life. (*Hudannnohyouzyouhadonnnakannjidearuka*)	I did not notice anything. (*Tokuninanimokidukanakatta*)
27	Facial expressions and the texture of the skin (*Hyoujouyahadasitu*)	I did not notice anything. (*Nanimokidukanakatta*)
28	I thought what I felt was affected by facial expression, viewing angle, and their clothes. (hyoujyouyamirukakudo,matakiteiruhukunadoniyottekanjikatagakawarunatoomoimashita)	I guessed the order of the question is controlled by faces. (kaoniyottetoinojyunbanwosousasiteirunatoomoimashita)
29	Nothing (*Tokuninasi*)	I did not notice anything. (*Tokunikidukazu*)
30	English written by participant: how were their teeth and eyes is important	English written by participant: there are pictures of same personss different ages
31	What kind of occupation, what they (*donoyounashokugyounanoka nanioshiteiruhitonanoka*)	I did not notice anything. (*Tokunikidukanakatta*)
32	Their actual age (*Zissainonennrei*)	Nothing (*tokuninasi*)
33	How beauty a face is (*Kaonoutukusisa*)	Contrast, freckle (*konntorasuto,simi*)
34	Eyes (*Metuki*)	English written by participant: Yes
35	People who are smiley are credible and attractive. (*egao no hito wa shinrai dekirushi miryokuteki*) English written by participant: The person who was smiling looked much more reliable and attractive than ones who were not.	There was one Asian person. (*Hitori no Asia jin.)* English written by participant: There was one Asian lady, and others are mostly White.
36	I thought there were many men with a had a shaved (or bald?) head (*Bouzugaooitoomoimashita*)	The hairstyle of each man was relatively unique (e.g., a baldhead, white hair). (*Bouzuyasiraganadokamigatokutyoutekinamonogahikakutekiookatta*)

37	Atmosphere and impressions from eyes (*Hunikitomekaratutawaruinsyou*)	I did not feel strange at all. (*Nanimoiwakanhaarimasendesita*)
38	How the person is like (*Donoyounazinnbutuzouka*)	There were no middle-aged people (*Tyunennnohitogainai*)
39	The number of face lines, the colour of people's hair (*shiwanokazu,kaminoiro*)	Some faces had many freckles (*Kaogamadaranahitogaita*)
40	How kind and scared facial expressions are and how many face lines they have (*hyoujounoyasasisayakowasa,mata,kaonosiwanooosa*)	I did not realise any unusual thing. (*ijounihakigatukanakatta.*)
41	Face lines, skin (*siwa, hada*)	Some people did not show neutral faces. (*Magaozyanaihitogairu*)
42	Facial expressions (Kaonohyoujyou)	Facial expressions were not natural. (Sizennahyoujyoudehanai)
43	Whether people are likely to lie and whether I can assign a job to him/her (*uso wo tsukisouka douka to sigoto wo makaseraresouka douka*)	The colour of photos were processed and it was different from the actual colour. (*sikisai ga wazato jissaino shashin to kotonaruyouni shiteattakoto*)
44	If people in a photo are those who are around me, I imagined how I would think and how I would feel. (*mosi syasinn no hitotati ga jibunn no minomawari no hito nara dou omouka, dou kannjiruka wo kanngaeta.*)	I noticed that some people had many freckles, but I thought it was just intended to be like that. I thought that some people deliberately showed weird facial expressions. (*kaojyuu ni simi no younamono ga hirogatteiru hito ga itakoto ha kiduitaga,tannnaru seisitu nanoka to omotta. suuninn teido wazato hennna kao wo siteita youna kimo sita.)*
45	Whether their eyes are warm or not. (megayasashiikadouka)	There were persons who had many age spots or prominent teeth. (ijyounisobakasugaooihitoya,ijyounideppanahitogaita)
46	Atmosphere (*hunniki*)	No (*Iie*)
47	I considered what kind of facial expressions they showed (*Donoyounahyoujouwositeirukawokanngaemashita*)	I thought that how people were taken photos, the amount of light and perspectives (perspective sensation) were different in each person (*Hitoniyotteshashinnnoutsurikatayahikarinoryouyaennkinnkangakotonarimashita)*
48	Nothing (*Tokuninanimo*)	Each background was completely different. (*Haikeigazennzenntigau*)

49	Whether the person looked good (kind) or not (*Hitogayosasoukadouka*)	I haven't noticed anything (*Tokuninanimokidukimasenndeshita*)
50	I looked at each facial expression carefully. I thought whether they were kind or scary. (*hyoujyouwoyokumimashita.yasashisouka,kowasoukanadowo,kanngaemashita.)*	There were many young and old people. Other than this, I did not notice anything. (*wakaihitoto,roujinngaooitoomoinashita.soreigaiha,nanimokidukimasenndeshita*.)
51	It was hard to guess the age of old people. *(nenpainohitononenreihawakarinikui*.)	I did not notice anything. (*tokunikidukanakatta.)*
52	Facial expressions, the colour of hair, face lines and backgrounds (Haikwi does not mean anything in Japanese, but it might be "Haikei" = background) (*hixyoujixyou,kaminoiro,siwa,sonohitonohaikwi*)	I did not notice anything. (*Nanimokidukimasenndesita*)
53	How people would show facial expressions in various situations (*Sonohitogairoironajoukyounioitedonnnakaowosuruka*)	All people were foreigners. (*Subetegaikokujinn*)
54	Interesting (sometimes translated as: funny) (Omoshiroi)	No (*Iie*)
55	Age and gender Nennreitoseibetu	I did not realise anything (*Kiduknakatta*)
56	Even though people looked old, I thought the actual age would be different from their appearance. (*kao ga fukete irukara to itte jissai no nenrei ga sou toha kagiranai nodewa naidarouka to kanngaeta*)	Nothing (*toku ni nanimo*)
57	There are many old people (*roujin ga ooi*)	It seemed that the colour of a photo was dark and that the experiment would manipulate the impression of each person (*iroai ga kurakattari insyou wo sousa siteiru youni mieta*)
58	Atmosphere (*Hunniki*)	I did not realise anything (*Kidukanakatta*)
59	Overall, the gender and age of each person were old or young *(Seibetuyanennreigadaitaiwakaikatosiwototteiruka*)	I did not know (*Wakaranakatta*)

60	I thought that those who were dominant were similar to those who were credible (*sihaitekinahitoto,sinraidekiruhitoganiteirukigasita*.)	I was impressed by those who were young and had many spots in their face. (*wakakute,kaonihantengaaruhitogainshountekinikanjitta*)

Appendix 2.3 Within-subject social rating differences between realistic mask and real face images for A) pooled, B) British, C) Japanese participants

A) Pooled British (exp. 1) and Japanese (exp. 2) participants. Comparison only include the 40 participants (df =39) that rated one of the three masks and relative comparison real-face images.

Age

		M	SE	t	p
Old female	Mask	79.41	2.33	2.28	.028
	Real face	75.88	2.29		
Old male	Mask	63.94	1.78	-4.45	<.001
	Real face	70.81	1.87		
Young male	Mask	37.86	.94	10.26	<.001
	Real face	28.18	.43		

Attractiveness

		M	SE	T	p
Old female	Mask	2.00	.19	-7.58	<.001
	Real face	3.18	.20		
Old male	Mask	2.33	.18	-5.59	<.001
	Real face	3.18	.20		
Young male	Mask	2.11	.18	-8.82	<.001
	Real face	3.65	.10		

Trustworthiness

		M	SE	T	p
Old female	Mask	3.23	.23	-7.78	<.001
	Real face	5.11	.16		
Old male	Mask	3.67	.25	-5.44	<.001
	Real face	4.85	.12		
Young male	Mask	2.84	.19	-4.20	<.001
	Real face	3.57	.13		

Dominance

		M	SE	T	p
Old female	Mask	3.75	.29	2.97	.005
	Real face	2.92	.16		
Old male	Mask	4.69	.22	2.91	.006
	Real face	4.04	.15		
Young male	Mask	4.76	.24	-.89	.378
	Real face	4.93	.15		

B) British participants (exp. 1). Comparisons only include the 20 participants (df=19) that rated one of the three masks and relative comparison real-face images.

Age

		M	SE	t	p
Old female	Mask	83.67	2.20	3.66	.002
	Real face	76.21	1.25		
Old male	Mask	61.79	3.01	-3.91	.001
	Real face	71.91	1.27		
Young male	Mask	38.00	1.26	7.44	<.001
	Real face	28.73	.48		

Attractiveness

		M	SE	t	p
Old female	Mask	1.43	.16	-5.82	<.001
	Real face	2.63	.22		
Old male	Mask	1.53	.16	-3.98	.001
	Real face	2.38	.25		
Young male	Mask	1.89	.25	-6.54	<.001
	Real face	3.59	.13		

Trustworthiness

		M	SE	t	p
Old female	Mask	3.19	.32	-7.10	<.001
	Real face	5.60	.14		
Old male	Mask	3.26	.42	-4.38	<.001
	Real face	4.83	.19		
Young male	Mask	2.53	.27	-3.40	.003
	Real face	3.42	.18		

Dominance

		M	SE	t	p
Old female	Mask	3.71	.37	1.632	.118
	Real face	3.13	.25		
Old male	Mask	4.68	.34	1.623	.122
	Real face	4.17	.24		
Young male	Mask	5.26	.34	-.562	.581
	Real face	5.41	.18		

C) Japanese participants (exp. 2). Comparisons only include the 20 participants (df =19) that rated one of the three masks and relative comparison real-face images.

Age

		M	SE	t	p
Old female	Mask	76.60	3.91	.38	.028
	Real face	75.75	3.23		
Old male	Mask	66.10	2.15	-2.83	<.001
	Real face	70.25	2.45		
Young male	Mask	37.79	1.42	7.02	<.001
	Real face	27.54	.69		

Attractiveness

		M	SE	t	p
Old female	Mask	2.55	.29	-4.88	<.001
	Real face	3.83	.28		
Old male	Mask	3.00	.22	-4.14	.001
	Real face	3.90	.17		
Young male	Mask	2.32	.24	-5.89	<.001
	Real face	3.70	.16		

Trustworthiness

		M	SE	t	p
Old female	Mask	3.25	.32	-4.67	<.001
	Real face	4.63	.23		
Old male	Mask	4.10	.25	-3.27	.004
	Real face	4.85	.15		
Young male	Mask	3.16	.25	-2.49	.023
	Real face	3.72	.20		

		Dominance			
		M	**SE**	**t**	**p**
Old female	Mask	3.85	.44	2.65	.016
	Real face	2.75	.20		
Old male	Mask	4.81	.29	2.65	.015
	Real face	3.80	.22		
Young male	Mask	4.26	.30	-.68	.506
	Real face	4.46	.19		

Appendix 7.1 Analysis of personality social inferences

Openness. Openness scores were submitted to a 2 × 2 x 3 mixed ANOVA to test for an effect of Mask (OFM, OMM, YMM), Wearer (Mladen, Florence) and difference between Paradigm (Unaware, Aware, Ignore). This analysis also only showed a main effect of Paradigm [$F(2, 86) = 16.24$, $p < .001$, partial $\eta^2 = .27$], and no main effect of Mask $F(2, 86) = .52$, $p = .594$], Wearer [$F(2, 86) = .83$, $p = .367$] or interactions. Participants produced significantly lowest Openness scores for the Aware paradigm (M = 16.0% yes responses; SE = .44, CI = -.07 – 16.6 [vs. Unaware: $p = .029$; vs. Ignore: $p < .001$]), followed by the Unaware (M = 29.1% yes responses; SE = .40, CI = 21.2 – 36.9), then followed by the Ignore paradigm (M = 48.6% yes responses; SE = .40, CI = 40.6 – 56.6 [vs. Unaware: $p = .001$]). These results suggest that knowledge of the wearer underneath the mask highlights the contrast in appearance, exaggerating scores from the Unaware and Aware paradigm.

To compare these scores to the wearer's actual appearance we ran a 4 x 2 repeated measures ANOVA of Mask (OFM, OMM, YMM, No mask) and Wearer (Mladen, Florence) for Aware paradigm only. As expected, the analysis shows a main effect of Mask [$F(3, 84) = 27.38$, $p < .001$, partial $\eta^2 = .49$], no effect of wearer [$F(1, 28) = 1.46$, $p = .110$], but an interaction effect ([$F(3, 84) = 3.52$, $p = .018$, partial $\eta^2 = .11$]. Simple main effects analysis shows that the interaction is due to a difference in between Mladen's and Florence's Openness scores without the mask, which is obstructed when they both wear the same masks. Most relevant to our recognition of the wearer and attribution error hypotheses, pairwise comparisons reveal that No mask Openess scores (M = 51.3% yes, SE = 3.5, CI = 44.1 – 58.4) are higher from all three masks (OFM: M = 17.2% yes, SE -2.7, CI = 11.6 – 22.7; OMM: M = 15.5% yes, SE = 3.4, CI = 8.6 – 22.4%; YMM: M = 17.2% yes, SE = 3.5, CI = 7.1 – 27.2; $p < .001$ for all comparisons).

As for Unaware and Aware paradigm, there is no trend in Openness scores

for the wearer underneath the mask (Ignore paradigm) [main effect of mask: F(2, 56) = .200, p = .819]. As Attractiveness, Openness scores do not provide evidence for an attribution error. We are not able to say whether this is due to the lack of diversity in openness appearance between hyper-realistic masks, or the lack of attribution to the wearer. Next we compare the averaged mask scores for each wearer in Ignore paradigm with the No mask scores for both wearer collected in the Aware paradigm in a 2 x 2 ANOVA between subject ANOVA. We see that despite a main effect of Wearer [F(1, 57) = 4.53, p = .036, partial η^2 = .07], there is no difference in scores between the two paradigms [main effect of Paradigm: F(1, 57) = .182, p = .672]. In other words, viewers are able to judge the wearer's openness without being affected by the mask's appearance.

Conscientiousness. We submitted Conscientiousness scores to the same mixed design ANOVA of Mask (OFM, OMM, YMM), Wearer (Mladen, Florence) and Paradigm (Unaware, Aware, Ignore). There was a main effect of Mask [F(2, 86) = 5.87, p =.004, partial η^2 =.12], main effect of Paradigm[F(2, 86) = 9.45, p <.001, partial η^2 =.18] and interaction effect between Wearer x Paradigm [F(2, 86) = 10.69, p <.001, partial η^2 =.20], but no main effect of Wearer [F(2, 86) = .53, p =.469]. Participants produced significantly lowest Conscientiousness scores for the YMM (M = 31.5% yes responses; SE = 2.9, CI = 25.8– 37.3 [vs. OMM: p = .002; vs. OFM: p = .001]), followed by the OMM (M = 39.9% yes responses; SE = 2.6, CI = 34.7– 45.2 [vs. OFM: p = .082]) and the OFM (M = 43.8% yes responses; SE = 3.0, CI = 37.9– 49.8). As with the Openness judgements, participants produced significantly lowest Conscientiousness scores for the Aware paradigm (M = 27.6% yes responses; SE = 3.9, CI = 19.8– 35.5 [vs. Unaware: p =.139; vs. Ignore: p <.001]), followed by the Unaware paradigm (M = 35.9% yes responses; SE = 3.9, CI = 19.8– 35.5 [vs. Ignore: p =.139]) and Ignore paradigm (M = 43.8% yes responses; SE = 3.0, CI = 43.7– 59.7). These results suggest that knowledge of the wearer underneath the mask highlights the contrast in appearance, exaggerating conscientiousness scores from the Unaware and Aware paradigm towards both further extremes.

To compare these scores to the wearer's actual appearance we ran a 4 x 2

repeated measures ANOVA of Mask (OFM, OMM, YMM, No mask) and Wearer (Mladen, Florence) for the Aware paradigm only. As expected, the analysis shows a main effect of Mask [$F(3, 84) = 13.39$, $p < .001$, partial $\eta^2 = .32$], no effect of Wearer [$F(1, 28) = .64$, $p = .429$], but also an interaction effect ([$F(3, 84) = 12.44$, $p = .001$, partial $\eta^2 = .30$]. Simple main effects analysis show that the interaction is due to a difference between Mladen's and Florence's Conscientiousness scores without a Mask (with Florence appearing significantly more conscientious), which is the opposite when they both wear the same masks. Florence in any mask looks much less conscientious than Mladen does. Most relevant to our recognition of the wearer and attribution error hypotheses, pairwise comparisons reveal that No mask Conscientiousness scores (M = 54.8% yes, SE = 3.2, CI = 48.2 – 61.4) are higher from all three Masks (OFM: M = 28.5% yes, SE = 4.7, CI = 18.8 – 32.2; OMM: M = 31.7% yes, SE = 4.4, CI = 22.8– 40.6%; YMM: M = 22.7% yes, SE = 4.9, CI = 12.7 – 32.7; $p < .001$ for all comparisons).

Although there is a trend in the same direction as observed across masks in Paradigm, there is no main effect of masks on Conscientiousness scores for the wearer underneath the mask (Ignore paradigm) [$F(2, 56) = 1.69$, $p = .194$]. As attractiveness and Openness, Conscientiousness of the mask does seem to be attributed to the wearer. This cannot be explained by a lack of diversity between mask appearance, visible from Unaware and Aware paradigm. We do see a main effect of Wearer [$F(2, 56) = 4.67$, $p = .039$, partial $\eta^2 = .14$], with perceivers scoring Mladen through the mask significantly lower in Conscientiousness (M = 43.9% yes, SE = 5.9, CI = 31.9 – 55.9) than Florence through the Mask (M = 59.5% yes, SE = 6.2, CI = 46.9 – 72.2), indicating that viewers can in fact distinguish between Wearers. Comparing the average across mask scores for each wearer's in Ignore paradigm with the No mask scores for both wearer collected in the Aware paradigm in a 2 x 2 ANOVA between subject ANOVA, we again see the main effect of Wearer [$F(1, 57) = 13.63$, $p = .001$, partial $\eta^2 = .19$], and no difference in scores between the two paradigms [main effect of Paradigm: $F(1, 57) = .281$, $p = .598$]. In other words, viewers are not only able to make a distinction between Wearers, but are also – with the right instruction – able to

judge the wearer's conscientiousness without being affected by the mask's appearance.

Extroversion. We also submitted Extroversion scores to a mixed design ANOVA of Mask (OFM, OMM, YMM), wearer (Mladen, Florence) and Paradigm (Unaware, Aware, Ignore). There was a main effect of Mask [$F(2, 85) = 5.97$, $p = .004$, partial $\eta^2 = .12$], main effect of Wearer [$F(2, 85) = 4.41$, $p = .039$, partial $\eta^2 = .05$] and a main effect of Paradigm [$F(2, 86) = 22.86$, $p < .001$ partial $\eta^2 = .35$], but no interaction effects. Participants produced significantly highest Extroversion scores for the YMM (M = 33.7% yes responses; SE = 2.6, CI = 28.5 – 38.9 [vs. OMM: p = .001; vs. OFM: p =.026]), followed by the OFM (M = 27.1% yes responses; SE = 2.2, CI = 22.6– 31.5 [vs. OMM: n.s p=.017]) and the OMM (M = 26.6% yes responses; SE = 2.4, CI = 21.8– 31.4). As with the Openness and Conscientiousness judgements, participants produced significantly lowest Extravesion scores for the Aware paradigm (M = 14.2% yes responses; SE = 3.6, CI = 7.0 – 21.3[vs. Unaware: p =.035; vs. Ignore: p <.001]), followed by the Unaware paradigm (M = 25.1% yes responses; SE = 3.6, CI = 17.9– 32.2 [vs. Ignore: p <.001]) and Ignore paradigm (M = 48.2% yes responses; SE = 3.7, CI = 40.9– 55.4). These results suggest further support that knowing the wearer underneath the mask highlights the contrast in appearance, exaggerating extravesion scores from Unaware and Aware paradigm.

To compare these scores to the wearer's actual appearance we ran a 4 x 2 repeated measures ANOVA of mask (OFM, OMM, YMM, No mask) and wearer (Mladen, Florence) for the Aware paradigm only. As expected, the analysis shows a main effect of Mask [$F(3, 84) = 18.69$, $p < .001$, partial $\eta^2 = .68$, an effect of Wearer [$F(1, 28) = 6.02$, $p = .02$, partial $\eta^2 = .17$], and an interaction effect ([$F(3, 84) = 5.88$, $p = .003$, partial $\eta^2 = .40$]. Simple main effects analysis show that the interaction is due to difference between Mladen's and Florence's Extroversion scores without the masks (with Mladen appearing significantly more extravert), which is not apparent when they are rated in the mask. Most relevant to our recognition of the wearer and attribution error hypotheses, pairwise comparisons reveal that No mask Extroversion scores (M = 44.2% yes, SE = 3.4, CI = 37.2 –

51.2) are higher than all three masks (OFM: M = 12.7% yes, SE = 3.3, CI = 6.0 – 19.3; OMM: M = 12.8% yes, SE = 3.6, CI = 5.5 – 20.2; YMM: M = 17.0% yes, SE = 3.7, CI = 9.4 – 24.6; $p < .001$ for all comparisons).

When scoring wearers underneath the mask (Ignore paradigm), we only see a main effect of masks on Extroversion scores [$F(2, 56) = 4.55$, $p = .015$, partial $\eta^2 = .14$] in the same trend as Unaware and Aware paradigm. This is evidence of an attribution error of the mask to the wearer. Comparing the averaged mask scores for each wearer's in Ignore paradigm with the No mask scores for both wearer collected in the Aware paradigm in a 2 x 2 ANOVA between subject ANOVA, we again see the main effect of wearer [$F(1, 57) = 13.55$, $p = .001$, partial $\eta^2 = .19$], and no difference in scores between the two paradigms[main effect of Paradigm: $F(1, 57) = .493$, $p = .483$]. In sum, viewers are able to distinguish between mask wearers, but are also influenced by the mask's appearance through the mask.

Agreeableness. Next we submitted the Agreeableness scores to the same mixed design ANOVA of Mask (OFM, OMM, YMM), Wearer (Mladen, Florence) and Paradigm (Unaware, Aware, Ignore). There was only a main effect of Mask [$F(2, 86) = 29.42$, $p < .001$, partial $\eta^2 = .26$] and a main effect of Paradigm [$F(2, 86) = 14.92$, $p < .001$ partial $\eta^2 = .26$], but no main effect of Wearer [$F(2, 85) = 1.46$, $p = .230$] or interaction effects. Participants produced significantly highest scores for the YMM (M = 25.6% yes responses; SE = 2.7, CI = 20.3 – 30.9) followed by the OMM (M = 35.3% yes responses; SE = 2.7, CI = 30.3– 40.5) and the OFM (M = 43.4% yes responses; SE = 3.2, CI = 37.1– 49.7, $p < .001$ for all comparisons). Participants produced significantly higher Agreeableness scores for the Ignore paradigm (M = 53.5% yes responses; SE = 4.4, CI = 44.8 – 62.2) than for the other two (Unaware paradigm: M = 29.7% yes responses; SE = 4.3, CI = 21.2– 38.3; Aware paradigm: M = 21.1% yes responses; SE = 4.3, CI = 12.5– 29.6). These results suggest a clear distinction between mask judgement and wearer judgement.

To compare these scores to the wearer's actual appearance we ran a 4 x 2

repeated measures ANOVA of Mask (OFM, OMM, YMM, No mask) and Wearer (Mladen, Florence) for the Aware paradigm only. As expected, the analysis shows a main effect of Mask [$F(3, 84) = 52.37$, $p < .001$, partial $\eta^2 = .64$], no effect of Wearer [$F(1, 28) = 3.08$, $p = .09$, partial $\eta^2 = .10$], and an interaction effect ([$F(3, 84) = 6.66$, $p < .001$, partial $\eta^2 = .19$]. Simple main effects analysis show that the interaction is due to difference between Mladen's and Florence's Extroversion scores without the masks (with Mladen appearing significantly more agreeable), which is not apparent when they are rated in the mask. Most relevant to recognising the wearer and attribution error hypotheses, pairwise comparisons reveal that No mask Agreeableness scores (M = 62.3% yes, SE = 3.2, CI = 55.7 – 69.0) are higher than all three masks (OFM: M = 32.5% yes, SE = 5.5, CI = 21.2 – 43.8; OMM: M = 21.5% yes, SE = 3.7, CI = 13.8 – 29.2; YMM: M = 9.2% yes, SE = 2.8, CI = 3.4– 14.9; $p < .001$ for all comparisons).

When scoring wearers underneath the mask (Ignore paradigm), we no longer see a main effect of masks on Agreeableness scores [$F(2, 56) = 2.46$, $p = .094$], despite it following the same trend as Unaware and Aware paradigm. Agreeableness judgements therefore do not provide evidence of attribution error of the mask to the wearer. Comparing the averaged mask scores for each wearer's in Ignore paradigm with the No mask scores for both wearer collected in the Aware paradigm in a 2 x 2 ANOVA between subject ANOVA, we do see a main effect of Wearer [$F(1, 57) = 10.30$, $p = .002$, partial $\eta^2 = .15$], and no difference in scores between the two paradigms [main effect of Paradigm: $F(1, 57) = .463$, $p = .483$]. In sum, viewers can distinguish between mask wearers, without being influenced by the mask's appearance through the mask.

Neuroticism. Finally, we submitted Neuroticism scores to the mixed design ANOVA of Mask (OFM, OMM, YMM), Wearer (Mladen, Florence) and Paradigm (Unaware, Aware, Ignore). There was only a main effect of Mask [$F(2, 86) = 13.35$, $p < .001$, partial $\eta^2 = .13$], a main effect of Wearer [$F(2, 85) = 9.93$, $p = .002$, partial $\eta^2 = .10$] and a main effect of Paradigm [$F(2, 86) = 13.45$, $p < .001$, partial $\eta^2 = .24$], or interaction effects. Participants produced significantly highest scores for the YMM (M = 54.2% yes responses; SE = 3.1, CI = 48.0 – 60.5 [vs. OFM: p <

.001; vs. OMM: p = .015) followed by the OMM (M = 46.5% yes responses; SE = 3.0, CI = 40.5– 52.5 [vs. OFM: p=.001]) and the OFM (M = 41.4% yes responses; SE = 3.1, CI = 35.7– 47.0). Participants produced significantly lower Neuroticism scores for the Ignore paradigm (M = 29.0% yes responses; SE = 4.6, CI = 19.9 – 38.2) than for the other two (Unaware paradigm: M = 51.2% yes responses; SE = 4.5, CI = 42.2– 60.2; Aware paradigm: M = 61.9% yes responses; SE = 4.5, CI = 52.9– 70.9). These results again suggest a clear distinction between mask judgement and wearer judgement.

To compare these scores to the wearer's actual appearance we ran a 4 x 2 repeated measures ANOVA of Mask (OFM, OMM, YMM, No mask) and Wearer (Mladen, Florence) for the Aware paradigm only. As expected, the analysis shows a main effect of Mask [$F(3, 84) = 27.00$, $p < .001$, partial $\eta^2 = .48$], no effect of wearer [$F(1, 28) = 8.26$, $p = .008$, partial $\eta^2 = .22$], or interaction effect ([$F(3, 84) = 1.93$, $p = .130$]. Most relevant to recognising the wearer and attribution error hypotheses, pairwise comparisons reveal that No mask Neuroticism scores (M = 27.8% yes, SE = 2.8, CI = 22.0 – 33.6) are significantly lower than all three masks (OFM: M = 57.8% yes, SE = 5.1, CI = 47.3 – 68.3; OMM: M = 58.7% yes, SE = 5.3, CI = 47.8 – 69.6; YMM: M = 69.2% yes, SE = 2.8, CI = 22.0– 33.6; $p < .001$ for all comparisons).

When scoring wearers underneath the mask (Ignore paradigm), we see a main effect of masks on Neuroticism scores [$F(2, 56) = 4.19$, $p = .020$, partial $\eta^2 = .13$], following the same trend as Unaware and Aware paradigm. This provides evidence that the neuroticism of the mask is attributed to the wearer. Comparing averaged mask scores for each wearer in Ignore paradigm with the No mask scores for both wearer collected in the Aware paradigm in a 2 x 2 ANOVA between subject ANOVA, we see no main effect of Wearer [$F(1, 57) = .622$, $p = .434$], and no difference in scores between the two paradigms [main effect of Paradigm: $F(1, 57) = 1.2$, $p = .278$]. In sum, neuroticism in masks was similar across wearers, and although viewers were able to approximate the wearer's neuroticism through the mask, their scores were also significantly influenced by the mask's appearance.

List of Abbreviations

2AFC. 2-Alternative Forced Choice

ANOVA. Analysis of Variance

FAE. Fundamental Attribution Error

DV. Dependent Variable

IV. Independent Variable

OMM. Old Male mask

OFM. Old Female mask

YMM. Young Male mask

YM. Young Male

References

Adam, H., & Galinsky, A. D. (2012). Enclothed cognition. *Journal of Experimental Social Psychology, 48*(4), 918-925.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current opinion in neurobiology, 12*(2), 169-177.

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ, 3*, e1184.

Algar, S. (2013) *Black robbers used $2,000 white masks to fool victims in $200,000 'Town'-style stickup, prosecutors say,* Retrieved from: http://nypost. com/2013/07/31/black-robbers-used-2000-white-masks-to-fool-victims-in-200000-town-style-stickup-prosecutors-say/, Accessed 21 Jul 2018

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Bornstein, B., & Carlson, C. (2014). Contribution to Alonga et al. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*(5), 556-578.

Anastasi, J. S., & Rhodes, M. G. (2005). An own-age bias in face recognition for children and older adults. *Psychonomic Bulletin & Review*, *12*(6), 1043-1047.

Anderson, B. L., Singh, M., & Fleming, R. W. (2002). The interpolation of object and surface structure. *Cognitive Psychology*, *44*(2), 148-190.

Andreas, P. (2003). Redrawing the line: borders and security in the twenty-first century. *International Security, 28*(2), 78-111.

Aviezer, H., Trope, Y., & Todorov, A. (2012a). Body cues, not facial expressions,

discriminate between intense positive and negative emotions. *Science, 338*(6111), 1225-1229.

Aviezer, H., Trope, Y., & Todorov, A. (2012b). Holistic person processing: faces with bodies tell the whole story. *Journal of Personality and Social Psychology, 103*(1), 20.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science, 21*(3), 372-374.

Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology, 36*(2), 147-168.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology, 81*(6), 1014.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Ssychology and Psychiatry*, *42*(2), 241-251.

Bartlett, F. C. (1932). *Remembering: An experimental and social study.* Cambridge: Cambridge University.

Bartlett, F. C., & Burt, C. (1933). Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology, 3*(2), 187-192.

Beer, A., & Watson, D. (2008). Personality judgment at zero acquaintance:

Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment, 90*(3), 250-260.

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: face-blind from birth. *Trends in Cognitive Sciences, 9*(4), 180-187.

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The cross-category effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological Science, 18*(8), 706-712.

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The cross-category effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological Science, 18*(8), 706-712.

Bernstein, S. (2010). *Masks so realistic they're arresting the wrong guy.* Retrieved from http://articles.latimes.com/2010/dec/08/business/la-fi-mask-20101209. Accessed 4 Oct 2017.

Berry, D. S., & Brownlow, S. (1989). Were the physiognomists right? Personality correlates of facial babyishness. *Personality and Social Psychology Bulletin, 15*(2), 266-279.

Berry, D. S., & Zebrowitz-McArthur, L. (1988). What's in a face? Facial maturity and the attribution of legal responsibility. *Personality and Social Psychology Bulletin, 14*(1), 23-33.

Bhattacharjee, S., & Marcel, S. (2017). What you can'" t see can help you--extended-range imaging for 3D-mask presentation attack detection. In Proceedings of the 16th International Conference on Biometrics Special Interest Group. (No. EPFL-CONF-231840). *Gesellschaft fuer Informatik eV* (GI).

Bindemann, M., & Burton, A. M. (2009). The role of color in human face detection. *Cognitive Science*, *33*(6), 1144-1156.

Bindemann, M., & Lewis, M. B. (2013). Face detection differs from categorization:

Evidence from visual search in natural scenes. *Psychonomic bulletin & review, 20*(6), 1140-1145.

Bindemann, M., Burton, A. M., Hooge, I. T., Jenkins, R., & De Haan, E. H. (2005). Faces retain attention. *Psychonomic Bulletin & Review, 12*(6), 1048-1053.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15*(10), 674-679.

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex, 82*, 48-62.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700.

Bond, Jr, C. F., Berry, D. S., & Omar, A. (1994). The kernel of truth in judgments of deceptiveness. *Basic and Applied Social Psychology, 15*(4), 523-534.

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2017). Can we detect cooperators by looking at their face? *Current Directions in Psychological Science, 26*(3), 276-281.

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72*(4), 691.

Bruce, V. (1988). *Recognising faces*. Lawrence Erlbaum Associates, Inc.

Bruce, V. (2012). Familiar face recognition. *Craniofacial Identification,* 1.

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

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207.

Burton, A. M., & Bindemann, M. (2009). The role of view in human face detection. *Vision Research, 49(*15), 2026-2036.

Burton, A. M., & Vokey, J. R. (1998). The face-space typicality paradox: Understanding the face-space metaphor. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(3), 475-483.

Burton, A. M., Bruce, V., & Hancock, P. J. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, *23*(1), 1-31.

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943-958.

Burton, A. M., Jenkins, R., Hancock, P. J., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256-284.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods, 42*(1), 286-291.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243-248.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26-48.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11(*12), 535-543.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London B:*

Biological Sciences, *363*(1493), 1001-1010.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21*(10), 1363-1368.

Chang, L. J., Doll, B. B., van't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87-105.

Chassin, L., Presson, C. C., & Sherman, S. J. (1990). Social psychological contributions to the understanding and prevention of adolescent cigarette smoking. *Personality and Social Psychology Bulletin, 16*(1), 133-151.

Chen, F. F., Jing, Y., & Lee, J. M. (2014). The looks of a leader: Competent and trustworthy, but not dominant. *Journal of Experimental Social Psychology, 51*, 27-33.

Chiao, J. Y., Bowman, N. E., & Gill, H. (2008). The political gender gap: Gender bias in facial inferences that predict voting behavior. *PLoS One, 3*(10), e3666.

Churches, O., Baron-Cohen, S., & Ring, H. (2009). Seeing face-like objects: an event-related potential study. *NeuroReport*, *20*(14), 1290-1294.

Clark, K., Cain, M. S., Adcock, R. A., & Mitroff, S. R. (2014). Context matters: The structure of task goals affects accuracy in multiple-target visual search. *Applied Ergonomics*, *45*(3), 528-533.

Cox, L. (2017) *Bank robbers stump the NYPD with life-like masks that made them look white even though they may have been black or Hispanic.* Retreived from: http://www.dailymail.co.uk/news/article-2108276/Bank-robbers-stump-NYPD-life-like-masks-look-white-black-Hispanic.html. Accessed 19 Jul 2018

Crouzet, S. M., & Thorpe, S. J. (2011). Low-level cues and ultra-fast face detection. *Frontiers in psychology*, *2*, 342.

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: face detection in just 100 ms. *Journal of vision*, *10*(4), 16-16.

Curtis, E., Ryan, C., Roy, S., Simes, T., Lapkin, S., O'Neill, B., & Faithfull-Byrne, A. (2016). Incorporating peer-to-peer facilitation with a mid-level fidelity student led simulation experience for undergraduate nurses. *Nurse Education in Practice, 20*, 80-84.

Daily Mail Reporter. (2011) *Police arrest passenger who boarded plane in Hong Kong as an old man in flat cap and arrived in Canada a young Asian refugee.* MailOnline. Retrieved from: http://www.dailymail.co.uk/news/article-1326885/Man-boards-plane-disguised-old-man-arrested-arrival-Canada.html, Accessed 21 Jul 2018

Damani, S. (2014) *The white robber who carried out six raids disguised as a black man (and very nearly got away with it)* Retrieved from: http://www.sebadamani. com/blog/the-white-robber-who-carried-out-six-raids-disguised-as-a-black-man-and-very-nearly-got-away-with-it, Accessed 21 Jul 2018

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. Philosophical Transactions of the Royal Society of London B: *Biological Sciences, 362*(1480), 679-704.

David, M.B. (2014) *Robbers Dressed As White Cops Got Away With $200,000 Until Sending 'Thank You' Letter To Mask-Maker,* Political Blindspot. Retrieved from: http://politicalblindspot.com/robbers-dressed-as-white-cops-got-away-with-200000-until-sending-thank-you-letter-to-mask-maker/, Accessed 21 Jul 2018

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*(4), 482-505.

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology, 30*(6), 827-840.

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior, 4*(4), 243.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory, *Law and human behavior*, *28*(6), 687.

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition, 126*(1), 87-100.

Department of Justice (2016) Bank crime statistics 2015. Retrieved from: https://www.fbi.gov/file-repository/stats-services-publications-bank-crime-statistics-2015-bank-crime-statistics-2015/view, Accessed: 21 Jul 2018

Department of Justice (2017) *Bank crime statistics 2016*. Retrieved from: https://www.fbi.gov/file-repository/bank-crime-statistics-2016.pdf/view, Accessed: 21 Jul 2018

Dhamecha, T. I., Singh, R., Vatsa, M., & Kumar, A. (2014). Recognizing disguised faces: Human and machine evaluation. *PloS one, 9*(7), e99212.

Dijksterhuis, A., & Van Knippenberg, A. (1998). The relation between perception and behavior, or how to win a game of trivial pursuit. *Journal of Personality and Social Psychology, 74*(4), 865.

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology, 106*(3), 433-445.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PloS one, 7*(1), e29081.

Drew, T., Võ, M. L. H., & Wolfe, J. M. (2013). The invisible gorilla strikes again sustained inattentional blindness in expert observers. *Psychological Science, 24*(9), 1848-1853.

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of Cognitive Neuroscience, 17*(2), 249-261.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*(4), 576-585.

Dumas, R., & Testé, B. (2006). The influence of criminal facial stereotypes on juridic judgments. *Swiss Journal of Psychology, 65*(4), 237-244.

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child Development, 82*(3), 793-811.

Dysart, J. E., Lindsay, R. C., & Dupuis, P. R. (2006). Show-ups: The critical issue of clothing bias. *Applied Cognitive Psychology*, *20*(8), 1009-1023.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science, 17*(5), 383-386.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, *1000*(1), 205-221.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124.

Ellis, H. D., Davies, G. M., & Shepherd, J. W. (1977). Experimental studies of face identification. *National Journal of Criminal Defense, 3,* 219.

Epstein, E.A. (2012) *Robbers who disguised themselves as white cops are caught... after they send polite thank-you letter to company that made their 'unbelievable' latex masks.* Daily Mail. Retrieved from: http://www.dailymail.co.uk/news /article-2192115/Robbers-disguised-white-

cops-caught--send-polite-thank-letter-company-unbelievable-latex-masks.html, Accessed 21 Jul 2018

Erdogmus, N., & Marcel, S. (2014). Spoofing face recognition with 3D masks. *IEEE transactions on information forensics and security*, *9*(7), 1084-1097.

Farage, M. A., Miller, K. W., Elsner, P., & Maibach, H. I. (2013). Characteristics of the aging skin. *Advances in wound care, 2*(1), 5-10.

Fecher, N., & Watt, D. (2013). Effects of forensically-realistic facial concealment on auditory-visual consonant recognition in quiet and noise conditions, *AVSP*, 81-86.

Fechner, G. T. (1860/1966) [Elemente der Psychophysik. 1860. Leipzig: Breitkopf und Hertel.] In H. E. Adler (Transl.), *Elements of psychophysics*. 1966. Holt, Rinehart, and Winston, New York.

Fernandez, F. M., Green, M. D., & Newton, P. N. (2008). Prevalence and detection of counterfeit pharmaceuticals: a mini review. *Industrial & Engineering Chemistry Research, 47*(3), 585-590.

Fink, B., Grammer, K., & Matts, P. J. (2006). Visible skin color distribution plays a role in the perception of age, attractiveness, and health in female faces. *Evolution and Human Behavior*, *27*(6), 433-442.

Fiske, S. T., & Taylor, S. E. (1991). Attribution theory. *Social Cognition,* 22-41.

Frank, M. G., & Gilovich, T. (1988). The dark side of self-and social perception: black uniforms and aggression in professional sports. *Journal of personality and social psychology, 54*(1), 74.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*(3), 490-495.

Frost, P. (1988). Human skin color: a possible relationship between its sexual

dimorphism and its social perception. *Perspectives in Biology and Medicine*, *32*(1), 38-58.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652.

Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science, 26*(6), 797-815.

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature, 75*(7), 450-451.

Gardner, D. (2010) *The bank rubber... Raider uses movie-style mask to disguise himself.* Daily Mail. Retrieved from: http://www.dailymail.co.uk/news/article-1268215/White-robber-fools-police-weeks-elaborate-African-American-Hollywood-mask.html, Accessed 21 Jul 2018

Gauthier, I., Behrmann, M., & Tarr, M. J. (1999). Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience, 11*(4), 349-370.

Gladwell, M. (2005). *Blink: The power of thinking without thinking.* New York: Little, Brown & Co.

Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition, 136*, 359-364.

Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48(*1), 26.

Gosselt, J. F., van Hoof, J. J., de Jong, M. D., & Prinsen, S. (2007). Mystery shopping and alcohol sales: Do supermarkets and liquor stores sell alcohol to underage customers? *Journal of Adolescent Health*, *41*(3), 302-308.

Graham, J. R., Harvey, C. R., & Puri, M. (2015). Capital allocation and delegation of decision-making authority within firms. Journal of Financial Economics, 115(3), 449-470. Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science, 19*(2), 109-111.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics, 95*(3), 537-557.

Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology, 52*, 15-23.

Hadjikhani, N., Kveraga, K., Naik, P., & Ahlfors, S. P. (2009). Early (N170) activation of face-specific cortex by face-like objects. *Neuroreport, 20*(4), 403.

Hall, E. T. (1966). *The Hidden Dimension*. Anchor Books, USA.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*(4), 845.

Harari, H. (1967). An experimental evaluation of Heider's balance theory with respect to situational and predispositional variables. *The Journal of Social Psychology, 73(*2), 177-189.

Hareli, S., & Rafaeli, A. (2008). Emotion cycles: On the social influence of emotion in organizations. *Research in Organizational Behavior, 28,* 35-59.

Harvey, J. H., Town, J. P., & Yarkin, K. L. (1981). How fundamental is" " the fundamental attribution error"? *Journal of Personality and Social Psychology, 40(*2), 346.

Henderson, E (2016) *'Swedish Fritzl' accused of kidnap and rape of woman kept in purpose built bunker 'just wanted a girlfriend*, Retrieved from

https://www.independent.co.uk/news/world/europe/swedish-fritzl-who-kidnapped-and-raped-woman-he-kept-in-purpose-built-bunker-for-six-days-just-a6820686.html, Accessed 19 Jul 2018

Her Majesty's Passport Office (2014) *Annual Report and Accounts 2013-14.* Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/332680/HMPO_ARA_2013-14_v5_1_FINAL_WEB.pdf, Accessed 21 Jul 2018.

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: a meta-analytic review. *Visual Cognition, 21*(9-10), 1306-1336.

Hirt, E. R., McDonald, H. E., & Markman, K. D. (1998). *Expectancy effects in reconstructive memory.*

Hockley, W. E., Hemsworth, D. H., & Consoli, A. (1999). Shades of the mirror effect: Recognition of faces with and without sunglasses. *Memory & Cognition, 27*(1), 128-138.

Holmes, O., Banks, M. S., & Farid, H. (2016). Assessing and improving the identification of computer-generated portraits. *ACM Transactions on Applied Perception (TAP)*, *13*(2), 7.

Home Office (2016) *Guidance on examining identity documents 2016.* Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/536918/Guidance_on_examining_identity_documents_v._June_2016.pdf

Ichikawa, H., Kanazawa, S., & Yamaguchi, M. K. (2011). Finding a face in a face-like object. *Perception, 40*(4), 500-502.

Ivcevic, Z., & Ambady, N. (2012). Personality impressions from identity claims on Facebook. *Psychology of Popular Media Culture, 1*(1), 38.

Jakobsen, K. V., Umstead, L., & Simpson, E. A. (2016). Efficient human face detection in infancy. *Developmental Psychobiology, 58*(1), 129-136.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 366*(1571), 1671-1683.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research, 2*(1999), 102-138.

Johnson, K. L., McKay, L. S., & Pollick, F. E. (2011). He throws like a girl (but only when he's sad): Emotion affects sex-decoding of biological motion displays. *Cognition*, *119*(2), 265-280.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*(1), 1-24.

Jonker, N., Scholten, B., van Emmerik, M., & van der Hoeven, M. (2006). Counterfeit or genuine: can you tell the difference (No. 121). *Netherlands Central Bank, Research Department*.

Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus and Giroux

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*(11), 4302-4311.

Kasra, M., Shen, C., & O'Brien, J. F. (2018). Seeing Is Believing: How People Fail to Identify Fake Images on the Web. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (p. LBW516). ACM.

Keil, M.S. (2009). 'I Look in Your Eyes, Honey': Internal Face Features Induce Spatial Frequency Preference for Human Face Processing. *PLoS*

*Computational Biology, 5(*3): e1000329.

Kelley, H. H. (1967). *Attribution theory in social psychology. In Nebraska symposium on motivation.* University of Nebraska Press.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11*(3), 211-222.

Kok, G., Den Boer, D. J., De Vries, H., Gerards, H. J. H., & Mudde, A. N. (2014). Self-efficacy and attribution theory. *Self-efficacy: Thought control of action,* 245-282.

Kong, S. G., Heo, J., Abidi, B. R., Paik, J., & Abidi, M. A. (2005). Recent advances in visual and infrared face recognition—a review. *Computer Vision and Image Understanding, 97*(1), 103-135.

Kose, N., & Dugelay, J. L. (2013). Countermeasure for the protection of face recognition systems against mask attacks. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on* (pp. 1-6). IEEE.

Kramer, R. S., & Ritchie, K. L. (2016). Disguising superman: how glasses affect unfamiliar face matching. *Applied Cognitive Psychology, 30*(6), 841-845.

Kramer, R. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology, 63*(11), 2273-2287.

Kramer, R. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods, 49*(6), 2002-2011.

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: diversity can trump ability. *Animal Behaviour, 81*(5), 941-948.

Kruglanski, A. W., & Ajzen, I. (1983). Bias and error in human judgment. European *Journal of Social Psychology, 13*(1), 1-44.

Kutys, J. M. (2012). *Juror Decision Making: The Impact of Attractiveness and Socioeconomic Status on Criminal Sentencing and an Examination of Motivated Reasoning in Mock Jurors.*

Lampert, C. H., Mei, L., & Breuel, T. M. (2006). Printing technique classification for document counterfeit detection. *In Computational Intelligence and Security, 2006 International Conference on (Vol. 1, pp. 639-644).* IEEE.

Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 905.

Langdridge, D., & Butt, T. (2004). The fundamental attribution error: A phenomenological critique. *British Journal of Social Psychology, 43*(3), 357-369.

Langton, S. R., Law, A. S., Burton, A. M., & Schweinberger, S. R. (2008). Attention capture by faces. *Cognition, 107*(1), 330-342.

Learning, S. (2003). Altruism and prosocial behavior. *Volume 5 Personality and Social Psychology,* 463.

Leder, H., Tinio, P. P., Fuchs, I. M., & Bohrn, I. (2010). When attractiveness demands longer looks: The effects of situation and gender. *The Quarterly Journal of Experimental Psychology*, *63*(9), 1858-1871.

Leikas, S., Verkasalo, M., & Lönnqvist, J. E. (2013). Posing personality: Is it possible to enact the Big Five traits in photographs? *Journal of Research in Personality, 47*(1), 15-21.

Lindsay, R. C., Wallbridge, H., & Drennan, D. (1987). Do the clothes make the man? An exploration of the effect of lineup attire on eyewitness identification accuracy. *Canadian Journal of Behavioral Science/Revue Canadienne des*

Sciences du Comportement, *19*(4), 463.

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology, 98*(1), 111-126.

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior, 28*(1), 18-27.

Little, A. C., Burt, D. M., & Perrett, D. I. (2006). What is good is beautiful: Face preference reflects desired personality. *Personality and Individual Differences, 41*(6), 1107-1118.

Lobmaier, J. S., & Mast, F. W. (2007). Perception of novel faces: The parts have it!. *Perception, 36*(11), 1660-1673.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology, 12*(3), 17-39.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280.

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., 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.

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, and Law, 7*(1), 98.

Mainey, L., Dwyer, T., Reid-Searl, K., & Bassett, J. (2018). High-Level Realism in

Simulation: A Catalyst for Providing Intimate Care. *Clinical Simulation in Nursing, 17*, 47-57.

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* Mit Press.

Malm, S. (2016) *'Sweden's Fritzl' used plastic masks to disguise his victim during the 350mile drive from her home to his specially made 'sex slave' bunker on their second date.* Retrieved from: http://www.dailymail.co.uk/news/article-3403420/How-Sweden-s-Fritzl-used-plastic-masks-disguise-victim-350mile-drive-home-specially-sex-slave-bunker-second-date.html, Accessed 18 Jul 2018

Manjani, I., Tariyal, S., Vatsa, M., Singh, R., & Majumdar, A. (2017). Detecting silicone mask-based presentation attack via deep dictionary learning. IEEE *Transactions on Information Forensics and Security, 12*(7), 1713-1723.

Mannetti, L., Brizi, A., Belanger, J., & Bufalari, I. (2016). All we need is the candidate's face: The irrelevance of information about political coalition affiliation and campaign promises. *Cogent Psychology, 3*(1), 1268365.

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the London Royal Society B, 258(*1353), 273-279.

Matsumoto, D., & Hwang, H. S. (2011). Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion, 35*(2), 181-191.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in cognitive sciences, 6*(6), 255-260.

McCaffery, J. M., & Burton, A. M. (2016). Passport Checks: Interactions Between Matching Faces and Biographical Details. *Applied Cognitive Psychology*, *30*(6), 925-933.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of

personality across instruments and observers. *Journal of personality and social psychology, 52*(1), 81.

McGrath, M., & Turvey, B. E. (2014). Eyewitness Identification: Uncertainty, Error, and Miscarriages of Justice. *Miscarriages of Justice: Actual Innocence, Forensic Evidence, and the Law*, 91.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11*(1), 8-15.

McNeil, J. E., & Warrington, E. K. (1993). Prosopagnosia: A face-specific disorder. The *Quarterly Journal of Experimental Psychology, 46*(1), 1-10.

Meeren, H. K., van Heijnsbergen, C. C., & de Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences, 102*(45), 16518-16523.

Megreya, A. M. (2018). Feature-by-feature comparison and holistic processing in unfamiliar face matching. *PeerJ, 6*, e4437.

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *The Quarterly Journal of Experimental Psychology*, *64*(8), 1473-1483.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3.

Meissner, C. A., Susa, K. J., & Ross, A. B. (2013). Can I see your passport please? Perceptual discrimination of own-and other-race faces. *Visual Cognition*, *21*(9-10), 1287-1305.

Mendoza, B.B. (2015) *Practical Disguise: The art of hiding in plain sight.* USA:

Steel Springs Press.

Merry, S. (2012). Eye See You: How Criminal Defendants Have Utilized the Nerd Defense to Influence Jurors' Perceptions. *Journal of Law & Policy, 21*, 725.

Mileva, M. (2017). *Within-Person Variability in Social Evaluation* (Doctoral dissertation, University of York).

Miller, F. D., Smith, E. R., & Uleman, J. (1981). Measurement and interpretation of situational and dispositional attributions. *Journal of Experimental Social Psychology, 17*(1), 80-95.

Montepare, J. M., & Zebrowitz-McArthur, L. (1988). Impressions of people created by age-related qualities of their gaits. *Journal of personality and social psychology*, *55*(4), 547.

Moon, H., & Phillips, P. J. (2001). Computational and performance aspects of PCA-based face-recognition algorithms. *Perception, 30*(3), 303-321.

Mori, M. (1970). Bukimi no tani [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.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychological Review*, *98*(2), 164.

Motoyoshi, I., Nishida, S. Y., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, *447*(7141), 206-209.

Neil, L., Cappagli, G., Karaminis, T., Jenkins, R., & Pellicano, E. (2016). Recognizing the same face in different contexts: Testing within-person face recognition in typical development and in autism. *Journal of Experimental Child Psychology*, *143*, 139-153.

Niedenthal, P. M., Brauer, M., Halberstadt, J. B., & Innes-Ker, Å. H. (2001). When did her smile drop? Facial mimicry and the influences of emotional state on

the detection of change in emotional expression. *Cognition & Emotion*, *15*(6), 853-864.

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications, 2*(1), 30.

Noyes, E. (2016). *Face Recognition in Challenging Situations* (Doctoral dissertation, University of York).

Noyes, E. & Jenkins, R. (under review) *Deliberate disguise in face identification.*

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition, 165*, 97-104.

Noyes, E., Hill, M. Q., & O'Toole, A. J. (2018). Face recognition ability does not predict person identification performance: using individual data in the interpretation of group results. *Cognitive Research: Principles and Implications, 3*(1), 23.

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, *22*(2), 208-224.

Olivola CY, Eastwick PW, Finkel EJ, Hortacsu A, Ariely D, Todorov A. (2014). *A picture is worth a thousand inferences: first impressions and mate selection in Internet matchmaking and speed-dating*. Unpublished manuscript.

Olivola, C. Y., & Todorov, A. (2010a). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior, 34*(2), 83-110.

Olivola, C. Y., & Todorov, A. (2010b). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*(2), 315-324.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias

human choices. *Trends in Cognitive Sciences, 18*(11), 566-570.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092.

Parra, E. J., Kittles, R. A., & Shriver, M. D. (2004). Implications of correlations between skin color and genetic ancestry for biomedical research. *Nature genetics, 36*(11s), S54.

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory, 3*(4), 406.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social cognition, 24*(5), 607-640.

Peterson, M. Cox, I. Eckstein, M. (2008). The use of the eyes for human face recognition explained through information distribution analysis, *Journal of Vision, 8*(6):894, 894a,

Petrican, R., Todorov, A., & Grady, C. (2014). Personality at face value: Facial appearance predicts self and other personality judgments among strangers and spouses. *Journal of Nonverbal Behavior, 38*(2), 259-277.

Pezdek, K. (2012). Fallible eyewitness memory and identification. *Conviction of the innocent: Lessons from psychological research,* 105-124.

Phillips, M. R., McAuliff, B. D., Kovera, M. B., & Cutler, B. L. (1999). Double-blind photoarray administration as a safeguard against investigator bias. *Journal of Applied Psychology*, *84*(6), 940.

Pollick, F. E. (2009, December). *In search of the uncanny valley.* In International Conference on User Centric Media (pp. 69-78). Springer, Berlin, Heidelberg.

Porter, S., England, L., Juodis, M., Ten Brinke, L., & Wilson, K. (2008). Is the face

a window to the soul? Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 40*(3), 171.

Porter, S., ten Brinke, L., & Gustaw, C. (2010). Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime & Law, 16*(6), 477-491.

Rae, J. A. (2012). Will it ever be possible to profile the terrorist? *Journal of Terrorism Research.*

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the Robustness of Power Posing No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological Science,* 0956797614553946.

Raven, D. (2015) *Black bank robbery suspect 'wore white old man mask to con police*' Retrieved from https://www.mirror.co.uk/news/world-news/black-bank-robbery-suspect-wore-5321157, Accessed 19 Jul 2018

Reid-Searl, K., Levett-Jones, T., Cooper, S., & Happell, B. (2014). The implementation of Mask-Ed: Reflections of academic participants. *Nurse Education in Practice, 14*(5), 485-490.

Reid-Searl, K., Eaton, A., Vieth, L., & Happell, B. (2011). The educator inside the patient: students' insights into the use of high fidelity silicone patient simulation. *Journal of Clinical Nursing, 20*(19-20), 2752-2760.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS one, 7*(3), e34293.

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: a meta-analytic and theoretical review. *Psychological Bulletin, 138*(1), 146.

Richler, J. J., Floyd, R. J., & Gauthier, I. (2015). About-face on face recognition

ability and holistic processing. *Journal of Vision, 15*(9), 15-15.

Righi, G., Peissig, J. J., & Tarr, M. J. (2012). Recognizing disguised faces. *Visual Cognition, 20*(2), 143-169.

Robertson, D. J. (2018). Face recognition: security contexts, super-recognizers, and sophisticated fraud. *The Journal of The United States Homeland Defence and Security Information Analysis Center (HDIAC), 5*(1), 6-10.

Robertson, D. J., Jenkins, R., & Burton, A. M. (2017). Face detection dissociates from face identification. *Visual Cognition*, *25*(7-8), 740-748.

Robertson, D. J., Kramer, R. S., & Burton, A. M. (2017). Fraudulent ID using face morphs: Experiments on human and automatic recognition. *PloS one*, *12*(3), e0173319T

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by Metropolitan Police super-recognisers. *PloS one, 11*(2), e0150036.

Roesch, S. C., & Weiner, B. (2001). A meta-analytic review of coping with illness: do causal attributions matter? *Journal of Psychosomatic Research, 50*(4), 205-219.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Rosenthal, R. (1966). *Experimenter effects in behavioral research.*

Ross, L. (1977). The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process. *In Advances in Experimental Social Psychology, 10,* 173-220.

Rule, N. O., & Ambady, N. (2009). She's got the look: Inferences from female chief executive officers' faces predict their success. *Sex Roles, 61*(9-10), 644-652.

Rule, N. O., & Ambady, N. (2010). Democrats and Republicans can be differentiated from their faces. *PloS one, 5*(1), e8733.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology, 104*(3), 409-426.

Russano, M. B., Dickinson, J. J., Greathouse, S. M., & Kovera, M. B. (2006). Why don't you take another look at number three: Investigator knowledge and its effects on eyewitness confidence and identification decisions. *Cardozo Public Law, Policy and Ethics Journal, 4*, 355.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*(2), 252-257.

Sabawi, F. (2018) *San Antonio bank robber who wore "old man" mask learns fate in federal court,* Retrieved from https://www.mysanantonio.com/news/local/crime/a rticle/San-Antonio-bank-robber-who-wore-old-man-mask-12562652.php#item-85307-tbla-5, Accessed 19 Jul 2018

Sabini, J., Siepmann, M., & Stein, J. (2001a). The Really Fundamental Attribution Error in Social Psychological Research. *Psychological inquiry, 12*(1), 1-15.

Sabini, J., Siepmann, M., & Stein, J. (2001b). Authors' response to commentaries. *Psychological Inquiry, 12*(1), 41-48.

Said, C. P., & Todorov, A. (2011). A statistical model of facial attractiveness. *Psychological Science, 22*(9), 1183-1190.

Samochowiec, J., Wänke, M., & Fiedler, K. (2010). Political ideology at face value. *Social Psychological and Personality Science, 1*(3), 206-213.

Schlicht, E. J., Shimojo, S., Camerer, C. F., Battaglia, P., & Nakayama, K. (2010). Human wagering behavior depends on opponents' faces. *PloS one, 5*(7),

e11663.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: another look at the availability heuristic. *Journal of Personality and Social psychology, 61*(2), 195.

Sergent, J., & Signoret, J. L. (1992). Varieties of functional deficits in prosopagnosia. *Cerebral Cortex, 2*(5), 375-388.

Seyama, J. I., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and Virtual Environments*, *16*(4), 337-351.

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100*(2), 139.

Shaver, K. G. (2012). *The attribution of blame: Causality, responsibility, and blameworthiness.* Springer Science & Business Media.

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, *6*(12), 1317.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059-1074.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7*(1), 36.

Stanton, J. (2015) *Brazilian drug trafficker attempts prison break wearing a mask, wig and dress to disguise himself as an old woman,* Daily Mail, Retrieved from: http://www.dailymail.co.uk/news/article-3293688/Brazilian-drug-trafficker-attempts-prison-break-wearing-mask-wig-dress-disguise-old-woman.html#ixzz4KKuuY3Ba, Accessed 21 Jul 2018

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological science, 21*(3), 349-354.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of personality and social psychology, 54*(5), 768.

Sugita, Y. (2009). Innate face processing. *Current Opinion in Neurobiology*,*19*(1), 39-44.

Sui, J., & Liu, C. H. (2009). Can beauty be ignored? Effects of facial attractiveness on covert attention. *Psychonomic Bulletin & Review*, *16*(2), 276-281.

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105-118.

Sutherland, C. A., Rowley, L. E., Amoaku, U. T., Daguzan, E., Kidd-Rossiter, K. A., Maceviciute, U., & Young, A. W. (2015). Personality judgments from everyday images of faces. *Frontiers in Psychology, 6*, 1616.

Sutherland, C. A., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology, 108*(2), 397-415.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology Section A, 46*(2), 225-245.

Tanaka, J. W., & Gordon, I. (2011). Features, configuration, and holistic face processing. *The Oxford Handbook of Face Perception,* 177-194.

Telegraph (2010) *Hong Kong conviction over 'old man' plane disguise,* Retrieved from: http://www. telegraph.co.uk/news/worldnews/asia/hongkong/8832028/Hong-Kong-conviction-over-old-man-plane-disguise.html, Accessed 21 Jul 2018

Terry, R. L. (1993). How wearing eyeglasses affects facial recognition. *Current Psychology, 12*(2), 151-162.

Terry, R. L. (1993). How wearing eyeglasses affects facial recognition. *Current Psychology, 12*(2), 151-162.

Terry, R. L. (1994). Effects of facial transformations on accuracy of recognition. *The Journal of Social Psychology, 134*(4), 483-492.

Terry, R. L. (1994). Effects of facial transformations on accuracy of recognition. *The Journal of Social Psychology, 134*(4), 483-492.

Tham, D. S. Y., Bremner, J. G., & Hay, D. (2017). The other-race effect in children from a multiracial population: a cross-cultural comparison. *Journal of experimental child psychology, 155,* 128-137.

The Guardian (2014). *Italian's passport used to board flight MH370 was stolen in Phuket,* Retrieved from: https://www.theguardian.com/world/2014/mar/09/italian-passport-malaysia-airlines-flight-mh370-stolen-phuket

The Innocence Project (2016). http://www.innocenceproject.org/about/

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657-665.

Tingley, D. (2014). Face-Off: Facial Features and Strategic Choice. *Political Psychology, 35*(1), 35-55.

Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces [social sciences]. *IEEE Signal Processing Magazine, 28*(2), 117-122.

Todorov, A., Dotsch, R., Wigboldus, D. H., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass, 5*(10), 775-791.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding

evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*(12), 455-460.

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception, 43*(2-3), 214-218.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied, 23(*1), 47.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136.

Turing, A. M. (1950). Mind. *Mind*, *59*(236), 433-460.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

UK Border Agency (2013) *UK Border Agency Annual Reports and Accounts 2012-2013.* Retrieved from: https://www.gov.uk/government/publications/uk-border-agency-annual-report-and-accounts-2012-to-2013--2, Accessed 21 Jul 2018.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology Section A, 43*(2), 161-204.

Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *The Quarterly Journal of Experimental Psychology, 44*(4), 671-703.

Valla, J. M., Ceci, S. J., & Williams, W. M. (2011). The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology, 5*(1), 66.

Van den Stock, J., Righart, R., & De Gelder, B. (2007). Body expressions

influence recognition of emotions in the face and voice. *Emotion*, *7*(3), 487.

Van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796-803.

Verhoff, M. A., Witzel, C., Kreutz, K., & Ramsthaler, F. (2008). The ideal subject distance for passport pictures. *Forensic Science International, 178*, 153–156.

Vestlund, J., Langeborg, L., Sörqvist, P., & Eriksson, M. (2009). Experts on age estimation. *Scandinavian Journal of Psychology*, *50*(4), 301-307.

Vokey, J. R., & Hockley, W. E. (2012). Unmasking a shady mirror effect: Recognition of normal versus obscured faces. *The Quarterly Journal of Experimental Psychology, 65(*4), 739-759.

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20*(3), 291-302.

Wagenmakers, E.J., Beek, T., Dijkhoff, L. & Gronau, Q.F. (2016) Registered Replication Report: Strack, Martin, & Stepper (1988) Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*, 768-777.

Warner, K. (1999, March). *Firearm deaths and firearm crime after gun licensing in Tasmania.* In Third National Outlook Symposium on Crime in Australia, convened by the Australian Institute of Criminology, Canberra.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(3), 129-140.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20(*3), 273-281.

Watson, D. (1989). Strangers' ratings of the five robust personality factors:

Evidence of a surprising convergence with self-report. *Journal of Personality and Social Psychology, 57*(1), 120.

Weiner, S. (2017). *US-Israeli indicted for 'old man mask' heists*, Retrieved from https://www.timesofisrael.com/us-israeli-indicted-for-old-man-mask-heists/, Accessed 19 Jul 2018

Weisman, D. (2018). Move over D.B. Cooper for Geezer Bandit, Retrieved from: https://escondidograpevine.com/2018/04/13/move-over-d-b-cooper-for-geezer-bandit/, Accessed 21 Jul 2018

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54*(1), 277-295.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human behavior*, *22*(6), 603.

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2012). Eyewitness Identification Reforms Are Suggestiveness-Induced Hits and Guesses True Hits? *Perspectives on Psychological Science*, *7*(3), 264-271.

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied cognitive psychology, 27*(6), 769-777.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS one, 9*(8), e103510.

Wiese, H., Komes, J., & Schweinberger, S. R. (2013). Ageing faces in ageing minds: A review on the own-age bias in face recognition. *Visual Cognition*,*21*(9-10), 1337-1363.

Wiese, H., Wolff, N., Steffens, M. C., & Schweinberger, S. R. (2013). How experience shapes memory for faces: an event-related potential study on the own-age bias. *Biological psychology, 94*(2), 369-379.

Wiesel, D.L. (2007) *Bank Robbery - Center for Problem-Oriented Policing,* U.S.

Department of Justice.

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science, 322*(5901), 606-607.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science, 17*(7), 592-598.

Windhager, S., Hutzler, F., Carbon, C. C., Oberzaucher, E., Schaefer, K., Thorstensen, T., & Grammer, K. (2010). Laying eyes on headlights: Eye movements suggest facial features in cars. *Collegium antropologicum*, *34*(3), 1075-1080.

Windhager, S., Slice, D. E., Schaefer, K., Oberzaucher, E., Thorstensen, T., & Grammer, K. (2008). Face to face. *Human Nature, 19*(4), 331-346.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review, 1*(2), 202-238.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience, 5*(6), 495.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*(7041), 439.

Womack, R. (2016). Enclothed Cognition: The Effect of Attire on Attention Task Performance. Samford Undergraduate Research Journal, 94.ogel, C. M. 2012. *Annual Brook Reviews, 78*, 741-743.

World Health Organization. (1999). *Counterfeit drugs: guidelines for the development of measures to combat counterfeit drugs* (No. WHO/EDM/QSM/99.1). Geneva: World Health Organization.

Yan, X., Andrews, T. J., Jenkins, R., & Young, A. W. (2016). Cross-cultural differences and similarities underlying other-race effects for facial identity and expression. *The Quarterly Journal of Experimental Psychology*, *69*(7),

1247-1254.

Yang, G., & Huang, T. S. (1994). Human face detection in a complex background. *Pattern recognition, 27*(1), 53-63.

Yarmey, A. D. (1986). Verbal, visual, and voice identification of a rape suspect under different levels of illumination. *Journal of Applied Psychology, 71*(3), 363.

Young, A. W., Hellawell, D., & Hay, D. C. (2013). Configurational information in face perception. *Perception, 42*(11), 1166-1178.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, *63*(3), 271-313.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*, *9*(2), 1–27.

Zamost, S. (2010). *Exclusive: Man in disguise boards international flight.* Retrieved from http://edition.cnn.com/2010/WORLD/americas/11/04/canada.disguised. passenger/ index.html. Accessed 4 Oct 2017.

Zebrowitz, L. A. (2004). The origin of first impressions. *Journal of Cultural and Evolutionary Psychology, 2*(1-2), 93-108.

Zebrowitz, L. A., & Montepare, J. M. (2006). The ecological approach to person perception: Evolutionary roots and contemporary offshoots. *Evolution and social psychology,* 81-113.

Zebrowitz, L. A., Voinescu, L., & Collins, M. A. (1996). " Wide-Eyed" and" Crooked-Faced": Determinants of Perceived and Real Honesty Across the Life Span. *Personality and social psychology bulletin, 22*(12), 1258-1269.

Zhuang, Z., Landsittel, D., Benson, S., Roberge, R., & Shaffer, R. (2010). Facial anthropometric differences among gender, ethnicity, and age groups.

Annals of occupational hygiene, 54(4), 391-402.

Złotowski, J. A., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2018). Persistence of the Uncanny Valley. *Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids*, 163-187.