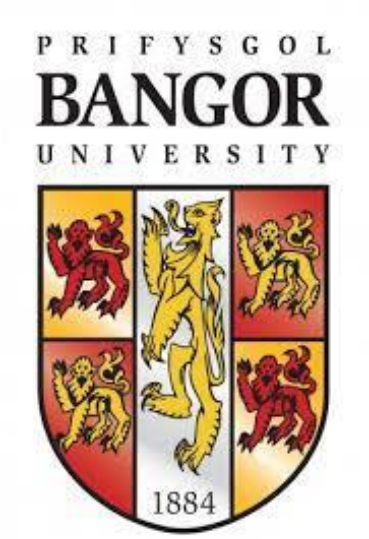
DOCTOR OF PHILOSOPHY

Incidental Impressions & Inconsistent Information

Newey, Rachel

Award date:
2022

Awarding institution:
Bangor University



Incidental Impressions & Inconsistent Information

Rachel Newey

Thesis submitted to the School of Psychology, Bangor University, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

September 2021

Declaration

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. All other sources are acknowledged by bibliographic references. This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree unless, as agreed by the University, for approved dual awards.

Funding

This work was part funded by the Economic and Social Research Council, under a 1+3 studentship, and part funded by the School of Psychology at Bangor University, UK.

For Paul. I wrote my damn thesis.

“Thank you for loving me. Thank you for being there.”

I would also like to thank my supervisors, Dr Kami Koldewyn & Dr Richard Ramsey. Kami, thank you for knowing what I was trying to say when I struggled to do so. But importantly, thank you for knowing that meetings are better with dogs.

Table of Contents

Chapter 1: Introduction

Overview

Detecting useful social cues and forming accurate impressions of people we encounter is crucial if we are to be successful in our social endeavours (Ames et al., 2011; Ames & Fiske, 2013; Capozzi & Ristic, 2018; Cuddy et al., 2008; Frith & Frith, 2007, 2011; Frith & Singer, 2008; Vogeley, 2017). Consequently, we form beliefs about people even when we are only paying cursory attention to them; in other words, we form incidental impressions. When forming these incidental impressions, we seek to form stable and coherent perceptions about the person's character (Ambady et al., 2000; Ames & Fiske, 2013; Moore, 2015; Vonk, 1994). However, people can be ambiguous, dynamic, and are prone to change their behaviour across time and context. This means that any initial impression we form may need to be updated if someone behaves inconsistently over time, rendering person perception an evolving and potentially complex process (Brambilla et al., 2019; Hughes et al., 2017; Mende-Siedlecki, et al., 2013; Zaki et al., 2010). However, to date, research on incidental impression formation has focused on the impressions we form when others behave consistently. The aim of the present work is to explore how experiencing others' inconsistent behaviour over time influences incidental impression formation and social decision-making.

Thesis

This thesis seeks to better understand how individuals form incidental impressions of others whose behaviours are inconsistent over time. Particularly, it considers whether incidental impressions that are formed when experiencing others' inconsistent gaze-cue behaviour are biased towards a person's early behaviours (i.e., primacy effects), later behaviours (i.e., recency effects), or a more complex amalgamation of all the information available (i.e., some form of averaging or valence based approach). In this way, the work examines how new, inconsistent behavioural information is incorporated, and impressions are updated, when impressions are formed incidentally.

Definitions

For the purposes of the present work, incidental impressions are defined as impressions formed when a perceiver (person forming the impression) has no social or other related goal, and the other person's behaviour does not comprise a core part of our task (i.e., if the task can be performed *without* reference to the other's behaviour). That is to say, if a perceiver experiences the other person's behaviours while they are attending to another task which neither references nor incorporates the other person directly, any social processing about the person or the meaning/intention of their behaviours occurs incidentally[1]. It is also important to clarify our use of "inconsistent" as behaviour, which if not strictly categorical (i.e., not wholly positive or negative), could be characterised as inconsistent. For our purposes, inconsistent behaviour is operationalised as behaviour that changes over time. As such, a person could display (wholly or mostly) positive social cues (e.g., helpful or cooperative) at the beginning of an encounter but later change to offer (wholly or mostly) negative social cues (e.g., deceitful or uncooperative).

Why study Incidental Impressions?

Holding a belief about what someone is like, or a prediction about what they might do is useful, as it allows us to plan our own behaviour (De Bruin & Van Lange, 1999; Fiske, 1992). As social beings, we encounter many different people under many different circumstances, and in such situations, we are often surrounded by multiple sources of social information. In order to operate successfully in the world, we need to detect, prioritise and make sense of potentially important information that others convey (Ames et al., 2011; Ames & Fiske, 2013; Capozzi & Ristic, 2018; Frith & Singer, 2008). This means we are typically very sensitive to social information, which we use to form impressions of others routinely and effortlessly (Birmingham & Kingstone, 2009; Fiske, 1993; Frith & Frith, 2008, 2011; Klapper et al., 2016; Moore, 2015; Todorov et al., 2015;

[1] This is to be distinguished from 'implicit' impressions, which, while also linked to impressions formed without awareness or intention, are often referenced as regards to how they are *revealed* (Uleman et al., 2005, 2008), not acquired. Implicit impressions are believed to be declaratively unavailable, meaning they are measured indirectly, using tools such as implicit association tasks (Greenwald & Banaji, 1995). The current work focuses exclusively on explicit impressions; those measured directly by asking what someone thinks about, or how they would act towards, a particular person.

Uleman et al., 1992, 1996, 2018; Uleman & Kressel, 2013; Vogeley, 2017). Indeed, social information does not even need to be the focus of our attention to influence our impressions of others; many of the impressions we form during everyday life are incidental in nature or the by-product of another task or goal (Carlson & Mae, 2003). Given our experiences with novel individuals can occur when they are not the focus of our attention, it is important to understand how incidental impressions are both formed and influenced. However, despite their ubiquity in our everyday lives, incidental impressions have received relatively little research attention.

What do we know about Incidental Impressions?

Incidental impression formation could be studied in many different ways. However, for the purpose of the current work, one specific approach was adopted; the use of gaze behaviour. We are highly sensitive to others' gaze behaviour (where and what they look at), which is usually an unconscious indication of where their attention lies, but which they can also use to send signals and communicate (Emery, 2000; Gobel et al., 2015). Changes in other's gaze direction have attentional effects, such that we reflexively orient our own visual attention to where they look if they suddenly shift their gaze (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999). This occurs even if we know the other person's gaze is not instructive, meaning others' gaze-based behaviour is capable of capturing and influencing our attention, even if we are not consciously aware of tracking their gaze behaviour and do not have a social goal. Beyond providing the source of their visual attention, where someone looks can also provide us with information about their internal mental state, such as their goals, desires, and emotions, which we can use to predict what they might do or to understand something they have already done (Argyle & Cook, 1976; Birmingham & Kingstone, 2009; Dalmaso et al., 2020; Doherty, 2006; Hamilton, 2016; Itier & Batty, 2009; Pfeiffer et al., 2013). We can use a person's gaze behaviour together with the context in which it is experienced to understand their intentions. Importantly, others' gaze behaviour can be used experimentally alongside of, rather than as part of, a task. This allows for social cues to be experienced without necessarily being the primary focus of attention, allowing for the investigation of incidental impression formation.

The use of gaze-cues to study incidental impression formation has been robustly explored and replicated (Bayliss & Tipper, 2006; Manssuer et al., 2016; Rogers et al., 2014; Strachan & Tipper, 2015; Strachan et al., 2016, 2017, 2020). Typically, perceivers engage in a visual attention paradigm where they are tasked with locating or categorising a target as quickly as they can when it appears. Alongside the task, centrally presented faces make eye contact with the perceiver before either looking towards (valid cue) or away from (invalid cue) the location where the task relevant target will imminently appear. As a consequence of unwittingly following the gazer's cue, the perceiver experiences faster (associated with valid cues) or slower (associated with invalid cues) performance when the target subsequently appears. While this means that others' behaviours are being processed at an attentional level, their presence is not explained, they are not related to the perceiver's primary goal (detect a target quickly), and the perceiver is attending to another task, meaning any additional social processing of the behaviours is both incidental and in competition with other attentional processes. Nevertheless, repeated exposure to a person's valid or invalid gaze-cue behaviour influences how we come to perceive them, with invalid cue providers being rated as less trustworthy than valid cue providers. Valid cue providers come to be associated with enhanced performance, whereas invalid cue providers are associated with poor performance, and in the context of the task, this could be interpreted as helpful or deceitful behaviour, respectively. This body of work demonstrates that we process others' behaviours at a social level even when they are not relevant to our goal and the focus of our attention is elsewhere.

What do we know about Incidental Impressions with Inconsistent Information?

To date, incidental impressions formed from exposure to other's gaze-cue behaviour have only been explored using consistent and wholly helpful or deceitful gaze-cue behaviours. This means that individual faces have only ever provided unwaveringly helpful or deceitful information. In real life, people and contexts are often more nuanced, adding complexity to impression formation processes and often requiring us to respond to new and potentially conflicting information (Brambilla et al., 2019; Mende-Siedlecki, 2018; Ross & Nisbett, 1991; Siegel et al., 2018). It is important, therefore, to understand how inconsistent information is incorporated and evaluated when we incidentally form impressions. As gaze is an important social cue which can convey a wide range of social information, it is also important to more deeply explore incidental impressions

formed from others' gaze-cue behaviour under different and more complex conditions. Wholly (in)valid gaze-cues likely facilitate the encoding of faces, as the relationship between cues (e.g., deceitful) and the outcome (e.g., negative) is congruent and never deviates (consistent). If inconsistencies or incongruities are introduced, the processing of the behaviours and resulting impressions may also be affected.

Thesis Questions

This thesis addresses three primary research questions, all of which explore incidental impression formation using gaze-cue behaviour to provide (in)consistent social information, though each approaches the subject from a different perspective. The questions, as they relate to empirical chapters, are stated and introduced below.

Chapter 2: How does inconsistent behaviour influence social judgements and decisions?

As stated above, prior work has not yet addressed how inconsistent gaze-cue behaviour is reflected in overall impressions. This first line of enquiry probes the question by extending previous research that used traditional gaze-cueing paradigms (Driver et al., 1999; Friesen & Kingstone, 1998) to study incidental impression formation involving consistent behaviour (e.g., Bayliss & Tipper, 2006), by incorporating inconsistent behaviour. Over a series of experiments, faces' overall levels of helpfulness (i.e., proportion of valid gaze-cues) are manipulated such that they either increase, decrease, or remain stable over time.

If someone changes how they act towards us, we may need to adjust our impression of them and subsequent behaviour towards them. Whether and how we do so is likely to be impacted by several factors, including the context of the encounter and our goal (Ames & Fiske, 2013; Jones & Goethals, 1987). Any change in behaviour (cue validity) during a gaze-cueing task can be expected to have a concomitant effect on the perceiver's performance in the task. For example, if a face changes from providing helpful (valid) to deceitful (invalid) gaze-cues, the perceiver's performance will subsequently suffer. This is because we reflexively follow faces' gaze shifts (Friesen & Kingstone, 1998; Langton & Bruce, 1999) and the attentional effect itself is impervious

to which face is giving the cue (Frischen & Tipper, 2004), meaning the cue, not the specific face per se, determines the perceiver's outcome. It is not known, however, how a change in gaze-cue behaviour will be incorporated into incidental impressions formed about the people behind the cues.

Broadly, there are three ways in which new, conflicting information (e.g., a once very helpful face now helps much less) could affect or interact with initial impressions. First, new information could replace and update earlier evaluations, as how someone has treated us most recently is the best predictor of how they will behave in the (near) future, and more likely reflects how they feel about us in the moment (Axelrod & Hamilton, 1981; King-Casas et al., 2005). Second, as updating an existing impression may require more effort than initially forming one (Erber & Fiske, 1984; Neuberg & Fiske, 1987), intentional (i.e., task related) impression formation and updating could involve different neural processes (Mende-Siedlecki, 2018). Therefore, if perceivers are not motivated to continually track others' behaviours or hold accurate beliefs about them, as seems plausible in a situation where social signals are received implicitly, it is possible impressions may not get updated (Ames & Fiske, 2013; Hendrick, 1972). Third, because all the information is received during a single, continuous session of brief interactions, the behaviours could be combined and averaged in some way (Hogarth & Einhorn, 1992) rather than one impression being formed, then that impression being updated following a change in behaviour. However, simple averaging is not often how people combine information about others (Jones et al., 1968; Soll & Larrick, 2009). Each of these options, however, ignores the valence (positive and negative) of initial behaviours/impressions and the direction of their change (e.g., provide more or less help), which are likely weighted (Anderson, 1965; Baumeister et al., 2001; Fiske, 1980; Skowronski & Carlston, 1989) and updated (Mende-Siedlecki, et al., 2013; Reeder & Coovert, 1986) differently. After all, while a person who helps more at the beginning than the end helps the same amount overall as someone who helps more at the end than the beginning, experiences with the two will differ qualitatively, with the former making a negative and the latter a positive change towards us.

Whether our initial impression of someone is positive or negative likely affects how much we adjust it in response to new, inconsistent information. For instance, when we learn about people's past morality-revealing behaviours, if our initial impression is particularly negative, new positive

information may not be sufficient to change our opinion much, whereas even a single piece of new, highly negative information may be enough erase an initial positive impression almost entirely (Mende-Siedlecki, et al., 2013; Reeder & Coovert, 1986; Richey et al., 1975). In the current work, inconsistencies in behaviour occur after both positive and negative behaviours have been experienced, meaning impressions can be adjusted bi-directionally; positive-negative and negative-positive. This is useful because experimental designs involving direct experience of a person's behaviour can sometimes only provide for the downgrading of positive impressions (e.g., Delgado et al., 2005; Fareri et al., 2012). It is important to also study the updating of negative impressions because in real life, people may not ordinarily continue interactions with people they disapprove of, thereby removing the opportunity for a negative impression to be changed and potentially making negative impressions more stable (Denrell, 2005). However, there is evidence to suggest that initial negative impressions are less certain than positive impressions (Siegel et al., 2018), which could mean that a negative impression can be revised when enough counter evidence is received.

Chapter 3: Do children form incidental impressions from gaze-cue behaviour?

To date, the influence of gaze-cue validity on impression formation has only been studied in adults. Using a traditional gaze-cueing task, this second line of enquiry examines whether the social trust learning effect is also present in adolescents (11-15 years of age). Most adults do not explicitly detect the faces' cue-outcome contingency, suggesting the relationship between the faces' behaviour and a participant's outcome is learned implicitly (Rogers et al., 2014). Those studies that have explored the development of trust learning across adolescence have used more explicit measures, such as interactive economic games (e.g., van den Bos et al., 2011) and it is not known when the socio-cognitive skills required for implicit trust learning come 'on-line'.

The effect of gaze-cue validity on performance (response times) in isolation is present in adolescents (van Rooijen et al., 2018) and younger children (though here it is often examined as a contrast to autistic children's performance) (e.g., Kylliäinen & Hietanen, 2004; Senju et al., 2004; Swettenham et al., 2003). Eight month olds have been shown to track the reliability of faces' (and arrows') cues (Tummeltshammer et al., 2014) and by around five years of age, children can use a

person's gaze behaviour to detect if they are verbally lying about the location of an object (Freire et al., 2004), demonstrating that we have a mentalistic understanding of others' gaze behaviour from a young age (Doherty, 2006). However, it is not known whether the ability to encode and infer the meaning behind a person's visual behaviour, all while completing another visual task, is present in children. This is especially pertinent given that in the gaze-cueing task, the others' behaviours are not referenced at all, such that gaze-information has no explicit relationship with the task or the child's goal. A face's presence and, more importantly, its motives and goals, may not be noticed or tracked by children as they are in adults. The implicit nature of the social information in gaze-cueing tasks is unlike most developmental studies in impression formation, where children focus on and learn about others who behave within a predefined set of rules (e.g., derivations of the trust game) (e.g., Lee et al., 2016; van den Bos et al., 2010, 2011). Further, in such scenarios, children are encouraged and motivated to think about the other person, both by the experimenter and by the nature of the interaction itself. By using a gaze-cueing task, we can assess if (and when) children form impressions incidentally.

Adolescents were selected as a suitable age group to start an exploration of the development of implicit impression formation as we felt we could use the same task we had already tested with adults without much alteration, and the age group also represented one developmental "step" back from adulthood. If we saw no evidence of an effect in an adolescent age-group, then stepping back further into childhood wouldn't seem necessary or interesting. During adolescence, many social skills and aspects of social cognition, including trust related behaviours, are still developing and undergoing refinement (Blakemore & Mills, 2014; Choudhury et al., 2006; Crone & Dahl, 2012; Evans et al., 2013; Kilford et al., 2016; Lee et al., 2016; van den Bos et al., 2010). Indeed, there is evidence that younger adolescents (~10-12 years of age) may not yet make distinctions between different interaction partners based on their contrasting, explicit sharing behaviours (Güroğlu et al., 2014; van den Bos et al., 2011). This would suggest that younger adolescents might not show the (implicit) trust learning effect and that even younger children would be highly unlikely to show effect. Interestingly, sensitivity to social stimuli, such as other's intentions and emotional expressions, and well as response to social feedback, has been shown to peak in mid-adolescence (Crone & Dahl, 2012), which could mean that the older adolescents in our study (14-15 years of

age) could be even more sensitive to the (perceived) meaning behind, and differences between, the faces' behaviours than are adults.

Chapter 4: Are incidental impressions and social decisions driven by the inferred social meaning of the other's behaviour or the perceiver's own outcome?

In many impression formation studies, a perceiver's outcome is of the same valence as the other person's intention; if they don't mean well, we don't do well. The congruent nature of this relationship presents no problem for us as we learn about someone and make decisions about how to respond to or act towards them, as their behaviour and our outcome are aligned. In a traditional gaze-cueing task, the social meaning attributed to a face's cue behaviour (i.e., invalid cues are deceitful) is congruent with and determines the perceiver's own outcome (i.e., invalid cues impair performance). This concomitant relationship also conflates impressions based on the other's behaviours with those that might be based on the perceiver's outcome. That is to say, when the other person's social behaviour and our goal-related outcome share the same valence (good or bad), it is not known whether the valence of the *behaviour* or the *outcome* drives our impression. In this final line of enquiry, we attempt to break the congruity between behavioural cue and perceiver outcome to explore whether social judgements and decisions are best predicted by the (perceived negative) social meaning behind a person's gaze behaviour, the perceiver's (actually positive) outcome, or both, depending on the situation.

In gaze-cueing tasks, the other person's behaviour is not explained or referenced, this means their cues are not directly interpretable and any social meaning or intention attributed to them must be inferred. Conversely, the outcome – which *is* experienced in their presence – provides actual knowledge about the relationship between their cues and our outcome, and this relationship can be learned associatively (Balliet et al., 2011; Behrens et al., 2008). It is not known, however, whether we base our social judgements and decisions on the inferred intention behind their behaviour or on our predicted future outcome, though we likely care about both. When we evaluate people by referencing the inferred reason for their behaviours, those who intend positive/negative outcomes are rated more positively/negatively than those whose behaviours are unintentional (Singer et al., 2004; Wu et al., 2018; Young et al., 2007). Indeed, so long as an outcome fails to occur due to an external factor, the mere intention of an immoral act is sufficient for us to judge someone

negatively (Hirozawa et al., 2020). Further, when we learn about someone's contrasting trait generosity and reward value from their sharing behaviour, we often weigh generosity (proportion shared) more heavily than the outcome (amount shared) when rating someone's likeability (Hackel et al., 2020). These findings might suggest that if the person's gaze behaviour is evaluated negatively and inferred to be intentional, then the actual outcome resulting from their behaviour will not be of huge importance to us when we make judgements about the person or decisions about future interactions with them.

On the other hand, while judgements about a person are likely to be based on inferred intentions, decisions involving them could also be influenced by our outcome. This may be especially the case if our decision is not social, but material in nature, such that any utility in the behaviour may be more relevant than any meaning behind it. Therefore, even if we think someone is trying to deceive us, if we learn that their behaviour is actually of benefit to us (e.g., when they look at the wrong answer, they reveal the correct answer), we may prioritise the utility of their behaviour over what we think of them when making certain goal related decisions. To explore this, the congruity between deceitful gaze-cue behaviour and negative outcomes is separated in our design, so that negative social behaviour can come to be associated with positive (financial) outcomes. In this case, the other person's behaviour does not determine a perceiver's outcome, rather it serves as a cue which the perceiver can learn and use to guide their decisions and improve their outcome. In this way, learning becomes more explicit, as we actively use others' cues to guide our behaviour. Importantly, impressions of the faces are still incidentally formed, as while the other person's behaviour is more salient than in the traditional gaze-cuing task, it remains outside the scope of our goal.

It is possible that changing how gaze-cues impact our outcome may also affect how they are processed. In the traditional task, deceitful (invalid) cues have negative effects, meaning any person who provides them will come to be associated with negative outcomes (Balliet et al., 2011; Behrens et al., 2008). In the study reported in Chapter 4, the same "negative" cue behaviour yields positive outcomes. When people (Sims et al., 2012) or non-social stimuli (Cox et al., 2005) are associated with positive outcomes, they acquire a reward value, however, it is not known whether a person who displays negative social behaviours will be learned about positively or if they will

remain ‘negative’. This is relevant to incidental impression formation, as the rewarding nature of our (goal related) outcome may be more salient than the inferred meaning behind a person’s gaze behaviour.

Finally, in this paradigm, inconsistent behaviour will not have the same impact on our outcome as it does in the traditional gaze-cueing task or other interaction based learning situations, such as iterative trust games (Campellone & Kring, 2013). If both positive (veridical) and negative (non-veridical) social cues can be used to our advantage, a change in someone’s behaviour will not have the same concomitant change to our outcome, as we can simply update our rule for that person and continue to do well. How we update our impressions, however, is an open question that we address in Chapter 4.

Literature Review

Owing to their importance in our social world, how we form impressions and the effects of inconsistent information have been the subject of empirical enquiry for many decades (for summaries and reviews see: Ames et al., 2011; Asch, 1946; Brambilla et al., 2019; Erber & Fiske, 1984; Fiske, 1980; Goodwin, 2015; Goodwin et al., 2014; Hogarth & Einhorn, 1992; Lee & Harris, 2013; Mende-Siedlecki, 2018; Moore, 2015; Reeder et al., 2004; Rozin & Royzman, 2001; Skowronski & Carlston, 1989; Todorov et al., 2015; Uleman et al., 2018; Uleman & Kressel, 2013)[2]. Inevitably, then, there is far more research that could be discussed here than there is reasonable space available to do so. As such, I have adopted a constrained approach to reviewing this vast literature. The purpose of this review is to 1) introduce what we already know about impression formation when information is inconsistent and impressions are formed intentionally or overtly, and 2) briefly review how others’ (consistent) gaze behaviour can provide social information that impacts perceivers’ behaviour and drives incidental impressions. To situate these two areas in the wider literature, the review begins with a brief overview of more general impression formation research, where information sources are singular or consistent. The literature

[2] For a review on implicit impression formation see Uleman and colleagues (2008) and for a recent review on the updating of implicit impressions see Ferguson and colleagues (2019).

investigating impression formation from inconsistent information will then be reviewed, where gaps and open questions relevant to the current work will be highlighted, followed by the literature looking at incidental impressions formed from gaze-cues.

Initial impression formation

Before we can update an impression, we must first form an initial impression. When we meet someone, whether we think they are a good or a bad person (i.e., their morality) is particularly important to us (Brambilla et al., 2021; Cuddy et al., 2008; Goodwin, 2015; Lammers et al., 2018), as it helps us to predict how they might behave towards us, allowing us to plan our own behaviour accordingly so as to increase our own chance of success (Ames et al., 2011; Cuddy et al., 2008; De Bruin & Van Lange, 1999; Fiske, 1992). We will feel positively about someone if we think they will be good to us and negatively if they may hurt us (Rozin & Royzman, 2001), and we make these appraisals effortlessly and automatically (Engell et al., 2007; Klapper et al., 2016; Oosterhof & Todorov, 2008; Todorov et al., 2015; Uleman et al., 1996, 2008; Vogeley, 2017). Seeing a person's face (Pakrashi et al., 2009; Willis & Todorov, 2006), reading about their past behaviour (Reeder & Coovert, 1986; Winter & Uleman, 1984) or simply experiencing their direct gaze (Kaisler & Leder, 2016) is sufficient for us to form an impression, demonstrating that we can base our judgements on both minimal and varied information. Though we can make many trait-based evaluations, initial impressions are most reliably and readily formed about a person's trustworthiness (Dotsch et al., 2017; Olivola et al., 2014)[3]. Given how important trust can be when interacting with people, holding a belief about their trustworthiness allows us to infer their intentions or predict their behaviour, and to distinguish between those who have positive intentions towards us and those who might intend to harm us or interfere with our goal (Ames et al., 2011; Cuddy et al., 2008; Engell et al., 2007; Oosterhof & Todorov, 2008; Todorov et al., 2008).

[3] We are also concerned with a person's competence as together with trustworthiness, they tell us what a person's intentions are and whether they have to ability to follow them through (Cuddy et al., 2008). In terms of first impressions, however, a person's trait warmth (specifically morality – which maps onto trustworthiness judgements) is most important and influential when we form impressions (Brambilla et al., 2021; Li et al., 2021).

Information sources

Broadly, there are two sources of information we can use to form trust-based impressions: what a person looks like (their appearance) and what a person does (their behaviour). To explore appearance based initial impressions, studies often use faces, which we rapidly evaluate for trustworthiness (for a review, see Todorov et al., 2015), even when we have no impression formation goal (Klapper et al., 2016). These judgements also guide trust based decisions, where more money is invested (and risked) with trustworthy than untrustworthy looking faces (Chang et al., 2010; Rezlescu et al., 2012; van 't Wout & Sanfey, 2008). However, while facial appearance may generate rapid, implicit judgements of trustworthiness that influence our behaviour, they do not actually show whether a person *is* trustworthy, which we can learn only from what they do. Indeed, how a face is perceived can be modulated by both explicit (Bliss-Moreau et al., 2008; Falvello et al., 2015; Todorov & Olson, 2008) and implicit (Bayliss & Tipper, 2006; Heerey & Velani, 2010) knowledge of the person's positive and negative behaviours.

Knowledge of people's behaviours also allows us to infer something about their stable disposition (Frith & Frith, 2007; Jones & Davis, 1965), with immoral (negative) behaviour being more diagnostic of a person's character than moral (positive) behaviour. Bad people can do good things, but only bad people do bad things (Baumeister et al., 2001; Fiske, 1980; Skowronski & Carlston, 1989). The step between observation of behaviour and judgement about personality necessitates inferring something about their intentions and goals. When we observe someone's behaviours, therefore, rather than immediately judging their stable personality traits, we first make inferences about why they are doing something and what they are trying to achieve (Ames et al., 2011; Malle, 2004; Malle & Holbrook, 2012). If we know someone's goal or intentions, we can better predict their behaviour and adjust our own accordingly (Frith & Frith, 2006, 2011; Malle, 1999). Further, when we evaluate people, we are particularly sensitive to the intention behind their behaviours, and such intentions often carry more weight than the result of the behaviour (Hackel et al., 2020; Levine & Schweitzer, 2014; Singer et al., 2004; Young et al., 2007). In other words, we often judge others more on their perceived intentions than on the actual outcome of their actions, especially when making judgements about their stable disposition.

There are two ways we can get behavioural information about a person; indirectly (through second-hand information or from actions that affect other people) or directly (experienced first-hand) (Zarolia et al., 2017). Indirect information about a person's past (im)moral behaviour or socially relevant actions influences judgements of trustworthiness and initial trust-related decisions. People who are reported to have committed negative acts are trusted less than those who have performed good deeds (De Bruin & Van Lange, 1999; Delgado et al., 2005; Fouragnan et al., 2013; Maurer et al., 2018; Zarolia et al., 2017). If we have no information about a person's present intention and no direct experience of them ourselves and yet have to make a decision regarding them, it makes sense that we interpret 'bad to other people' as 'will be bad to me' (De Bruin & Van Lange, 1999). However, while this is the best approximation we can make, given the circumstance, how a person treated someone else in a different context does not actually tell us how they would behave towards us in the present context. Further, indirect experience cannot give us feedback about the accuracy of our decision or the opportunity to learn about the person, which we can only do if we experience them directly over time (Zarolia et al., 2017).

Direct Experience

We often learn about people directly through experience (Behrens et al., 2008) and repeated exposure to how someone behaves allows us to build up knowledge about them, their goals, and what they are likely to do next, or even how they might act in different situations. We can learn about a person's trustworthiness directly over time through a series of negative or positive exchanges, where we receive feedback on our decisions (Chang et al., 2010; Fouragnan et al., 2013) and can infer their intentions from the combination of their behaviour and the outcome (Malle, 2011). Direct interactions are often studied using economic games, where behaviour is frequently not predicted by economic theory, as people have social motives and emotions which can interact/interfere with 'rational' decision making processes (Camerer, 2003). When we engage in these reasonably simple economic interactions (i.e., two people each make a decision), our success often depends upon our ability to consider the other person's motives and likely behaviour (Rilling & Sanfey, 2010; Sanfey, 2007). This is because our outcome is often interdependent with their decision, meaning we need to consider what they might do when we make our own decision(s). In these scenarios, the decisions the other person makes can clearly indicate their

intention, as they are rule based and often either directly provide financial benefit or cost to the other player. Over time, by associating a person's behaviour (shares with us vs steals from us) with an outcome (we do well vs we do badly), we can use reward-learning mechanisms to learn about their trustworthiness and predict their future behaviour (Balliet et al., 2011; Behrens et al., 2008).

When we have no prior (indirect) information about someone's trustworthiness, we can learn about them directly during iterative trust games (Berg et al., 1995). We can learn about their likely future behaviour and intentions by using the feedback we get from our decision (was I right to trust them?) to guide our future behaviour (should I trust them again?). If we make the wrong decision, we experience a prediction error (i.e., the difference between an expected and an actual outcome) which we can use to adjust our belief and update our behaviour (Niv & Schoenbaum, 2008). In this way, social decision making can be examined using reinforcement learning models and neural data, which show that the same regions active for non-social reward learning (principally striatal regions) are also recruited for social learning (for examples see: Chang et al., 2010; Fareri et al., 2012; Harris & Fiske, 2010; Lee & Harris, 2013; Mende-Siedlecki, 2018; Rilling & Sanfey, 2010). As we learn about someone in this trial-and-error way during an iterative trust game, we initially attend more to the feedback stage of the interaction, where we process the rewards/losses (i.e., their decision and our outcome) resulting from our decision in an associative manner. Over time, this reward-related processing shifts to reflect a decision to trust, suggesting that people make predictions that previous sharers will share again in the current trial, rather than waiting for the feedback given at the end of the trial (King-Casas et al., 2005). Interestingly, reward centres in the brain are activated when a person who we have learned we can trust reciprocates our trust, but not when someone who ordinarily steals from us shares with us instead (Phan et al., 2010), suggesting a learned (negative) association can also impact how reward information is processed. When we learn about people this way, our explicit impressions of their trustworthiness follow how they have behaved towards us, with those who share infrequently being rated as untrustworthy while those who (almost) always share are seen as trustworthy (King-Casas et al., 2005; Phan et al., 2010).

As well as being able to distinguish between people based on their reward value (i.e., our outcome), we also care about whether people are acting intentionally (making their own choices) or unintentionally (following instructions). Intentions and goals ordinarily guide actions, meaning

why someone behaves in a certain way may be more important to us than the behaviour itself (Malle, 1999; Malle & Holbrook, 2012; Reeder, 2009). While we cannot learn much about people if they are following instructions, how they *choose* to act can tell us a lot about them (Frith & Frith, 2006). Reward value can be sufficient to learn about a person and guide our decisions during an interaction, but intentionality matters when it comes to judging and distinguishing between people. For instance, when faces are learned to be cooperators or defectors in an iterative prisoner's dilemma game, self-reported emotional responses (e.g., anger) to faces are higher for those who act with intent, and intentional defectors are rated significantly more negatively than those who also defect but do so without intent (Singer et al., 2004). Interestingly, faces who cooperate are rated the same for likeability, regardless of intention, however, those who intentionally cooperate elicit greater neural activity in social cognition and reward related regions (Singer et al., 2004). Findings suggest that not only do we learn to associate people's behaviours with outcomes so as to make predictions, we also process the meaning behind the behaviours, which then influences what we come to think about the person.

Inconsistent information & Impression Updating

When we encounter someone new, we use the information available to make dispositional inferences about their underlying character and we tend to expect them, guided by their disposition, to conform to our impression (Ambady et al., 2000; Ames & Fiske, 2015; Gilbert & Malone, 1995; Harris & Fiske, 2010; Ross & Nisbett, 1991). People, however, are complex and dynamic, with motives and goals that can change from context-to-context or moment-to-moment. This means that as we learn about someone, we could receive conflicting indirect information, or they could change how they behave towards us as we interact with them directly. Accordingly, we need to detect and potentially respond to inconsistent or expectancy violating information if we are to form accurate impressions and make optimal decisions (Brambilla et al., 2019; Erber & Fiske, 1984; Mende-Siedlecki, 2018; Mende-Siedlecki, et al., 2013). Depending on the circumstance, new information may prompt us to update or adjust an initial or evolving impression.

If we receive conflicting or inconsistent information about someone, we can essentially do one of two things; we can ignore it or we can incorporate it into our judgement of them. What we actually

do likely depends on a myriad of factors, such as the consequence of the change, our motivation or goal, our attention, and so forth. Further, the inconsistency itself can vary, as original and new information could be meaningfully inconsistent (e.g., initial and new information are both morality based) or more generally inconsistent (e.g., initial information is morality based, new information is about intelligence). There is evidence that updating an impression when new information is meaningfully inconsistent recruits distinct brain regions as compared to when information is only generally inconsistent, as we are considering and revising a specific trait based impression in the first case and updating general 'person knowledge' in the second case (Mende-Siedlecki & Todorov, 2016). How new information relates to the original information, therefore, likely impacts if *and* how we update our belief about a person. If we find out that the initial information was incorrect (e.g., the person did not commit a bad act), we will update our impression in light of the new information or mistake (Ecker & Rodricks, 2020; Wyer, 2010). If we receive new information that changes the meaning of initial information (e.g., behaviours originally perceived negatively are learned to have been done with positive intent), we re-evaluate the original information in the new context and readily update our impression (Cone et al., 2017; Mann & Ferguson, 2015). In these circumstances, revising the impression is the appropriate thing to do, as the new information over-rides the initial information in some meaningful way. If, however, the new information is meaningfully inconsistent with old, but is only inconsistent in the sense that it differs in valence (positive vs negative), and does not necessarily replace or invalidate original information (e.g., we know someone contributes to charities, then we learn that they cheat at poker), it is less clear how that new information should be incorporated into our impressions of them.

The issue of how we process inconsistent social information has been the subject of empirical scrutiny for decades. Indeed, the study perhaps considered to be the first impression formation study (Asch, 1946) actually concerned (generally) inconsistent information, finding that the order in which information was received influenced overall impressions. Though findings from traditional 'order effects' research often reflected attentional processing (i.e., memory effects for trait-based information) to a greater degree than they did social cognition (i.e., mentalising about a particular person) (Fiske, 1993), they paved the way for research that identified the kinds of social nuances, rules and heuristics that we use when forming impressions of people (Anderson, 1965; Baumeister et al., 2001; Cuddy et al., 2008; Fiske, 1980; Hogarth & Einhorn, 1992; Landy

et al., 2018; Skowronski & Carlston, 1989). The emerging field of 'impression updating', where an impression of a specific person is formed and then subjected to inconsistent information, has received less attention (Mende-Siedlecki & Todorov, 2016). While order effects findings are subject to a number of caveats (see 'Order effects' below), they may be of relevance to the current work, which focuses on (incidental) impressions formed when we have no social or 'mentalising' goal. To this end, the remainder of this section briefly summarises the main findings from (and limitations of) traditional order effects research, before discussing (intentional) impression updating based on both indirect information and direct experience.

Order effects

The simplest way to approach how inconsistent information is processed is order effects, which refer to situations where the order in which we receive information influences how we remember or evaluate the information (Hogarth & Einhorn, 1992). As a classic example, when iteratively learning about a person's performance in a test, if they do better at the beginning than at the end (i.e., descending performance), we think of them as more intelligent, and as having scored more accurately overall, than those who either do worse at the beginning (i.e., ascending performance) or who perform randomly (Jones et al., 1968). This happens in spite of the fact that the person gets the same number of answers correct in all scenarios, suggesting that rather than continually tracking a person's behaviour and continually updating our impression, we form an initial impression and do not tend to adjust it, even when we receive evidence to the contrary. This example illustrates the primacy effect, whereby information learned early on both predicts final impressions and is better remembered than later information (Anderson & Barrios, 1961; Asch, 1946; Tulving, 2008). A brief perusal of traditional impression formation and order effects literature leads one to believe that impressions are prone to primacy effects and resist being updated. As Nisbett and Ross (1980) put it:

"*although order of presentation of information sometimes has no effects on final judgment, and recency effects sometimes are found, these are the exception; several decades of psychological research have shown that primacy effects are overwhelmingly more probable*" (p172)

This conclusion, however, has been contradicted by several lines of subsequent enquiry. From a methodological perspective, it was noted that order effects may be the product of task parameters (Fiske, 1980; Forgas, 2011). When people are given trait-based information (in the form of adjectives) and asked to form an opinion and provide it only at the end, or they are not expecting to have to recall the information at all, primacy effects are often observed. Alternatively, if impressions are captured step-by-step, or people expect to be asked to recall the information, recency effects are often observed (Anderson & Hubert, 1963; Dreben et al., 1979; Stewart, 1965). This suggests that attention plays a role in order effects, such that when we are not continually monitoring information, we attend to later information less carefully than earlier information (Crano, 1977). This is supported by the fact that we spend less time on each piece of information as time passes (Belmore, 1987) and is consistent with the attention decrement (Hendrick & Costantini, 1970), discounting (Rabin & Schrag, 1999) or cognitive miser hypotheses (for a review see North & Fiske, 2012).

Primacy effects may also be influenced by how motivated perceivers are to attend to and process relevant social information. If we have no reason to track information or update our impression, why would we bother? A 'cognitive miser' could be seen as an efficient processer when considered in the context of failing to continually monitor or adjust an inconsequential belief. However, if our outcome is dependent on the other person in some way, we are more likely to attend to inconsistent information, as we are more motivated to form accurate and up-to-date impressions (Erber & Fiske, 1984; Neuberg & Fiske, 1987). This has been confirmed at a neural level, where consistent information receives more attention if our experience is not outcome dependent (i.e., we ignore inconsistent information if it is not relevant to us), and inconsistent information receives more attention if we think we will later have to interact with or rely on the other person (Ames & Fiske, 2013). Together, these findings show that our attention and motivation can interact with the process of impression formation as they change the relative weight given to early vs. late information.

The effects observed in updating studies also depends to some extent on the type of information used. Earlier studies used lists of trait adjectives (e.g., industrious, envious, determined, intelligent) to describe people. As a result, they tended to capture memory or attentional effects (what we

remember about someone) more than they did person perception (what we think about someone) (Fiske, 1993). Remembering someone as intelligent because the first word we learned about them was 'intelligent' is qualitatively (and likely cognitively) different to forming an impression that the person is intelligent based on knowledge or observational evidence of some kind. Relatedly, earlier studies often neglected the content or meaning of the information that was provided. Fiske (1980 – see also for a review of earlier impression formation and order effects research) noted that lists of trait adjectives are not how we experience people in real life and that these context-lacking words ignored any extremity (i.e., salience) or valence (i.e., positive or negative) attached to the information. The type of information we receive and whether it portrays someone positively or negatively likely effects what we come to think of them. Further, context is also important for understanding intent and evaluating unfolding events (Jones & Goethals, 1987). As such, early impression formation research lacked many aspects of perceiving people and forming impressions about specific individuals (Fiske, 1980; Zajonc, 1980). Later studies, therefore, started to make use of more complex stimuli by introducing richer descriptions about a person's good and bad behaviours (i.e., morality based information) allowing for the person and their behaviours to be evaluated rather than simply recalled.

When we judge someone based on their behaviours, valence matters. When equal amounts of positive and negative descriptive information are received, overall impressions are not neutral (i.e., information is not mathematically averaged), but rather negative, suggesting we do not treat/weigh positive and negative information the same (Anderson, 1965). Indeed, negative information can exert such a powerful effect that a single piece of negative information can negate five equally positive pieces of information, and the resulting impression of a person described in this way is only marginally better than that of someone who has been described with equal bad and good information (Richey et al., 1975). A single negative act can, therefore, undermine the presumption of goodness or morality, so it receives more weight (Baumeister et al., 2001; Fiske, 1980). Research consistently finds that behaviour reflecting a person's (im)moral character dominates impression formation and updating processes, when compared with information on other factors such as sociability or competence (Brambilla et al., 2019, 2021; Goodwin et al., 2013; Landy et al., 2018; Wojciszke et al., 1998). Relatedly, negative (immoral) information is regarded as more informative and diagnostic of a person's character than positive (moral) information, as anybody

can do something good, but only bad people are believed to do bad things (Skowronski & Carlston, 1989). As such, while the order in which descriptive information is presented can affect our impressions, it is not as important as the evaluative content of the information. In other words, we are much more likely to update our impression of someone we initially think is 'good' when we receive negative information than we are to update our impression of someone we think is 'bad' when we receive positive information about them.

Indirect Inconsistent Information

Building on the findings that moral information dominates impression formation and negative information weighs more than positive information, studies have gone on to explore how we update impressions of a person's morality[4] through indirect information. Starting with its simplest form, Reeder and Coovert (1986), provided participants with one piece of moral/immoral information about a person, had them rate their impression, then provided another piece of information that was inconsistent with the first, before having participants give their impression again. Both positive and negative impressions were adjusted following the inconsistent information, however, there was a greater change to impressions when new information was negative (moral-to-immoral change) than positive (immoral-to-moral changes). Additionally, people tended to spend less time deciding their final impression when an initial impression was negative and new information was positive. This asymmetrical updating effect is also present when more pieces of information are received and impressions are recorded step-by-step (Mende-Siedlecki, Baron, et al., 2013; Mende-Siedlecki, et al., 2013; Mende-Siedlecki & Todorov, 2016). These findings suggest that new positive information is not enough to reverse a negative impression (bad people can do good things), whereas new negative information receives greater attention and a favourable impression can be greatly reduced (only bad people do bad things). This supports the notion that negative information weighs more than positive, regardless of presentation order, and indicates that negative impressions may be harder to update than positive impressions.

[4] Though other types of information have been explored, the discussion focuses on morality based information as it maps closely onto judgements of trustworthiness and has been found to be the dominant source of information used when we update our impressions (Brambilla et al., 2019).

How we perceive people based on their indirect behaviour is important to understand, as it can guide both our judgements and our own behaviour when we actually encounter those people ourselves. However, because the encounter is observational in nature, any mental states we ascribe to the other person cannot affect us and we cannot respond to them (De Jaegher & Di Paolo, 2007; Pfeiffer et al., 2013; Redcay & Schilbach, 2019). In this way, observational encounters tend to measure 'offline' cognition, where we collect information about a person from a distance, rather than 'online' cognition, where we gain knowledge in person as an interactor (Pfeiffer et al., 2013). This may be an important distinction, as interacting with someone (where we can act and where they can personally affect us) is likely to be processed and experienced differently than when we are merely passively observing someone (Schilbach et al., 2013). Further, indirect information is often provided as discrete pieces of information which lack context or coherence, whereas in real life, we often encounter people or receive information embedded in a rich, dynamic context (Redcay & Moraczewski, 2020; Smith & Semin, 2004). In order for others' behaviours to be experienced within a context and for perceivers to be affected by and/or respond to the behaviours, encounters need to take place in real-time environments, where the perceiver can experience behaviours as they happen (Redcay & Moraczewski, 2020).

Direct Inconsistent Information

In the real world, when we interact with someone directly, they may behave contrary to our expectation or change how they behave over time. To maintain an accurate impression of someone, we thus need to detect any inconsistencies in their behaviour or intentions and flexibly update our impressions and behaviour. As previously discussed, the impressions that result from repeated direct interactions have been explored experimentally using iterative trust games, though inconsistent behaviours are rarely portrayed during the course of the direct interaction. Instead, perhaps due to its demonstrated ability to generate impressions, indirect information (e.g., descriptive moral vignettes) is often used to generate initial impressions, and inconsistent information is introduced during a subsequent direct interaction. As such, impression updating is often explored from the perspective of how we update our behaviour and impressions when our direct experience runs contrary to an initial indirect impression.

Behaviour based initial impressions

To explore how indirect behavioural information may interfere with learning during a direct encounter, Delgado and colleagues (2005) presented participants with vignettes that described interaction partners' characters prior to the direct interaction. Three characters were set up as either morally 'good', 'bad' or 'neutral'. Participants were then asked to rate each character's trustworthiness. The design had the desired effect of matching ratings with characters (i.e. the morally good character was rated as the most trustworthy). Participants then engaged in iterative (24 rounds) trust games with each character (while being scanned in an MRI machine), where all characters shared 50% of the time (i.e., behaved randomly and were non-predictive), meaning that the good character did not display much positive behaviour during the interaction. Unsurprisingly, initial investments followed character descriptions. However, over time, participants still chose to invest more often in the good character than the other characters, despite such characters' untrustworthy actions. As found in a previous study (King-Casas et al., 2005), participants did not initially attend to the outcome phase of the interactions with 'good' people, whereas they did attend during trials associated with neutral people, who they were still trying to learn about (Delgado et al., 2005). Notably, however, post interaction trustworthiness ratings revealed that trust ratings of the good character had significantly dropped and now matched that of the other two characters, demonstrating that while prior information can interfere with learning, impressions of the person are updated to reflect how a person has treated us rather than how they were initially depicted.

The Delgado (2005) study revealed how our initial impressions can influence both if and how new information is processed, as well as how it impacts our behaviour, such that inconsistent information is not automatically detected or acted upon if we have a strong prior impression of someone. However, there were two aspects to the design which need to be addressed. First, all the characters behaved in a non-predictive fashion during the interaction, making it impossible to associatively learn about them. In such a situation, it is not clear how an impression should be adjusted, as individual's intentions may not be clear (are they acting with a specific intention?) and there is no way to distinguish between individuals, as they are all behaving similarly (do any of them know what they're doing?). Second, unlike in studies using all indirect information, there

was no opportunity for a negative impression to be updated here, as the new information was neither positive nor negative.

Recent studies have followed a similar approach to explore how impressions are updated when subsequent direct experience is not non-predictive. Initial indirect information interferes less with learning about others when their behaviour in direct interactions is more clearly (un)trustworthy. For example, one recent study used the same initial moral vignettes as the Delgado (2005) study to generate positive or negative impressions, however, during the following iterative trust game, characters either shared or stole 75% of the time, which was either consistent or inconsistent with the initial impression (Zarolia et al., 2017, exp.2). Though learning rates were a little slower for people who were depicted as bad but who actually shared (bad-to-good change), by block 2 (rounds 6-10) investment decisions were based on peoples' reciprocity rates, not the initial negative information, and remained reasonably stable until the end of the interaction (after 20 rounds). As expected, pre-interaction trustworthiness ratings matched the valence of the initial indirect information, with ratings of bad people being particularly low. Interestingly, after the interaction, ratings were based (by and large) solely on characters' reciprocity rates, with very little difference between those who were depicted as morally different but who behaved the same way in the trust game. In these circumstances, initial negative impressions from indirect information were able to be updated through positive direct experience.

In a similar study (Maurer et al., 2018), participants received biographical (job related) information about four people, two of whom were depicted as morally good (e.g., works for doctors without borders) and two morally bad (e.g., scams pensioners and underpays employees). In the trust games (10 rounds), players were either consistent or inconsistent with initial impressions. Initial investments were again influenced by initial impressions, however, participants soon updated their behaviour to reflect how the others shared with them during the game. Using more clearly trustworthy and untrustworthy direct behaviour, these studies demonstrate that when our experience is inconsistent with our initial impression, we update our behaviour accordingly (albeit sometimes gradually) and adjust our impressions in the direction of our own experience, and their most recent behaviour.

Appearance based initial impressions

As already mentioned, we easily form initial impressions from someone's appearance. Yet we all consciously know that such appearance-based impressions may not accurately reflect how people actually act (Rule et al., 2013). Assessing how such appearance-based impressions are updated is complex, however, as while face-based impressions do lack substantive behavioural evidence, they are based on our own physiological responses (i.e., how the face makes us feel), which is a form of experience (Engell et al., 2007; Mattavelli et al., 2014; Olivola et al., 2014). As such, they may continue to exert an influence on our impressions even after we obtain opposing information during a direct interaction. Using 2-by-2 designs, studies have provided initial information on appearance (trustworthy or untrustworthy) and direct sharing behaviour (share ≥80% or steal ≥80%) to explore impression updating when initial impressions are appearance-based (Chang et al., 2010; Yu et al., 2014). Unsurprisingly, initial behaviour is influenced by faces' appearance (more money is invested in faces that look trustworthy), however, reasonably quickly, investments to those faces who mostly steal drop significantly, and after twenty five rounds of direct experience, investments are based only on sharing behaviour (Yu et al., 2014). Using reinforcement learning models to explore how participants' behaviour could be best predicted across trials, it was found that trust is best explained by a dynamic belief model that is initially influenced by appearance and is updated over time to reflect the other person's behaviour (Chang et al., 2010). Post interaction ratings of trustworthiness found no effect of face-type in one study, with differences being based only on behaviour (Yu et al., 2014), while another found that the trustworthy looking face that could be trusted to share was rated more highly than the untrustworthy looking face that displayed the same behaviour (Chang et al., 2010). Having a trustworthy looking face may, therefore, enhance our opinions of someone we learn about directly, even when it carries no corresponding enhancing information.

Studies looking at the relative influence of both faces and moral statements on trust have also reported 'face premium' effects. A trustworthy looking face retains a premium over a non-trustworthy looking face even when (more diagnostic) behavioural information is available that depicts the people equally (Li et al., 2017; Rezlescu et al., 2012). Pragmatically, behavioural evidence ought to override appearance-based impressions (Engell et al., 2007) and these studies

(Chang et al., 2010; Yu et al., 2014) demonstrated that trust related behaviour is indeed quickly adjusted when faces' behaviour does not conform to initial expectations. However, when faces are later rated, appearance can continue to impact judgements of trustworthiness, suggesting that despite ourselves we continue to place weight on a trustworthy looking face even after we have directly learned that they behave just the same as someone else.

Emotions & Inconsistent behaviour

It is worth noting that in the studies that have been discussed so far, faces used in the described tasks have displayed neutral expressions. Thus, while the faces may have appeared more or less trustworthy, they did not express emotions which could be used to infer intention. Emotional expressions are a clear social cue that can change how we perceive and behave towards others during an interaction (Van Kleef et al., 2010). Someone who is scowling at us (i.e., does not mean well) may provide more contextual information than someone who provides more implicit information (i.e., has untrustworthy facial features but a neutral expression). Campellone & Kring (2013) explored trust learning when emotional expressions (smiling or angry) and sharing behaviour (share or steal) were incongruent. In one condition (behaviour first), participants first learned directly about four people (denoted only by a number) who were either trustworthy or untrustworthy. Midway, expressive faces were introduced to the game, displaying either congruent (e.g., trustworthy and smiling) or incongruent (e.g., trustworthy and angry) expressions, while behaviour continued as before. Incongruent expressions had no effect on perceivers' behaviour, suggesting that new inconsistent and potentially meaningful information may not affect our behaviour when it has no effect on our outcome. Congruent expressions, however, did influence behaviour during the interaction, with participants trusting smiling, trustworthy faces more, and untrustworthy angry faces less. This suggests that additional consistent information might increase confidence in our belief or confirm any inference we may have about their intention. When participants later rated how trustworthy they thought each face was, the two faces who displayed untrustworthy behaviour were rated similarly, whereas for the two trustworthy faces, the face who gained a smile was trusted slightly more. This is similar to the premium trustworthy looking faces can receive (Chang et al., 2010; Li et al., 2017; Rezlescu et al., 2012).

In a second condition (face first), smiling and angry faces were included alongside others' congruent and incongruent trustworthy behaviours from the beginning of the interaction, where again, participants' learned based on the faces behaviours, not expressions (Campellone & Kring, 2013). This study also included a change of behaviour during the trust game itself. Halfway through the game, the faces were removed and two of the players changed how they behaved by reversing how much they reciprocated during the second half. Perhaps unsurprisingly, when players did not change their behaviour, neither did perceivers, meaning the removal of an expression had no impact on decisions in the game. When players changed how they behaved, however, perceivers quickly updated their decisions to match the players' new behaviours. When subsequently judging players' (depicted displaying neutral expressions) trustworthiness, ratings were entirely predicted by how the player behaved in the second half of the game. This is interesting, as it suggests that more consistent positive behaviour does not result in increased trust over those who changed from being untrustworthy only at the end of the game, implying only recent behaviour matters. While the best predictor of someone's future behaviour might be how they have behaved in (recent) the past (Axelrod & Hamilton, 1981; King-Casas et al., 2005), it is surprising that consistently positive behavioural does not receive a bonus. In this scenario, a change in behaviour is more salient than a person's expression. Moreover, whereas a trustworthy looking face (which may represent a more enduring trait) may continue to influence our perception of someone, an expression (transient representation of current mood) apparently does not.

All behaviour directly experienced

So far, the studies discussed have all included some form of indirect information. We clearly also need to update our impressions and behaviour if we learn about someone directly and they later change how they treat us. One study examined whether learning about someone in one environment influenced behaviour in another. Fareri and colleagues (2012) directly introduced participants to other people using a Cyberball game, which included the participant and three other players (Williams & Jarvis, 2006). Here, other players varied in how much they included the participant in a ball tossing game, and participants learned whether they were nice (included them a lot), mean (did not include them much) or neutral (tossed the ball to them randomly). Post-game subjective ratings of players' trustworthiness matched subjective experience during the ball game.

Participants then engaged in (24 round) iterative trust games, where all players behaved in the same, non-predictive (50% share rate) fashion, meaning the nice player clearly behaved inconsistently to their initial impression, whereas the others did not. Initial investments mirrored subjective ratings and they continued to do so for quite some time. By the end of the interaction, however, investment rates more closely resembled the players' behaviour during the trust game. Final trust ratings revealed little to no change from initial impressions for the mean and neutral faces while trust in the nice face had decreased to match that of the neutral face. This suggests that while a final impression can be adjusted in response to inconsistent information, decision making during an interaction can continue to be influenced by directly acquired prior information for a much longer time than indirect information. Further, reinforcement learning models (using MRI data) suggest that initial impressions continued to be reinforced as people are more sensitive to behaviour that is consistent with initial impressions than they are to inconsistent behaviour (Fareri et al., 2012). This suggests that rather than attending to inconsistent information that could have helped them change their behaviour during an interaction where the other player influenced their outcome (Erber & Fiske, 1984; Neuberg & Fiske, 1987), people instead attend to information that confirms their initial impression (Rabin & Schrag, 1999).

As such, impressions formed directly may interfere with learning more than impressions formed indirectly. This makes sense, as we have gone through the process of learning about the person associatively, allowing us to accurately predict their behaviour across multiple encounters (Behrens et al., 2008; Niv & Schoenbaum, 2008). Consequently, if we have been successful and feel confident in our knowledge of the other, by the time we receive new and potentially conflicting information, we are no longer carefully attending to the outcome phase of the encounter (King-Casas et al., 2005; Phan et al., 2010), as we expect people to behave in an impression-consistent manner (Gilbert & Malone, 1995). When people behave ambiguously or randomly, however, it is hard to understand their intention, use their actions in a predicative manner, or learn new social information from their behaviour, which could be one reason why players were so slow to update their initial impressions, which had been formed from direct, predictable, valenced experience.

Gaps & Open questions

In each of the scenarios discussed where social cues were observed through direct experience, the other person's behaviour was experienced within a rule-based context (e.g., in a game) and that behaviour had a direct and overt effect on the perceiver. Such scenarios result in a person's intentions and goals being relatively transparent to the perceiver. Further, the other person was the focus of the perceiver's attention, who was also outcome dependent on the other's behaviour (Neuberg & Fiske, 1987). Together, these factors serve to highlight and accentuate the other person's behaviour and this increased salience may contribute to the effects that have been demonstrated. In real life, repeated interactions can vary considerably both in terms of context and how much the other person's behaviour can affect us (Rousseau et al., 1998). Additionally, people's behaviour may be more ambiguous or may need to be deciphered given the context before it can be understood or evaluated (Redcay & Moraczewski, 2020). Finally, our direct goal may not involve the other person, potentially making it that much harder to evaluate their behaviour, as it is not directly relevant or related to the context. This means that if encounters take place under more socially impoverished circumstances, inconsistent information may not be detected or processed in the same way. Accordingly, it is important to understand impression formation processes that occur in response to a variety of social cues, in different contexts, and while observers are pursuing other goals.

Incidental Impressions & Gaze

Impressions are considered incidental if they are formed in the absence of any social goal and while attention is focused on some other non-social task. This means that the observer must be able to detect the other person's behaviour when it is not part of the task (i.e., their presence or actions are not explained) and is not goal related (i.e., our goal relates to something unrelated to the person). As discussed, (static) faces are a source of information about a person and we evaluate them automatically (Klapper et al., 2016; Olivola et al., 2014). Faces can also signal behaviour, and our eyes, and their movements, play a key role in social cognition (Birmingham & Kingstone, 2009; Marotta et al., 2019; Shepherd, 2010). Eyes can provide us with a great deal of information about a person's emotions, attention, and intentions (Itier & Batty, 2009). Eyes also serve both

attention and communication functions, being capable of both sending and receiving signals (Emery, 2000; Gobel et al., 2015). Indeed, observing a shift in another's gaze can have an automatic orienting effect on our own visual attention. Thus, we can both extract meaning from a person's gaze behaviour (where and what they look at) and be affected by their gaze behaviour, making gaze a good candidate for impression formation research. Importantly, gaze-based behaviour can be provided covertly, as it does not need to be part of the task (or goal) to be detected, and its presence does not require any explanation or rules. For this reason, gaze can be used to study the formation of incidental impressions.

This section will briefly discuss gaze as a social and attentional source of information, before discussing findings from studies that have used consistent gaze-cue behaviour to explore incidental impression formation.

Gaze

Faces can provide a lot of social information; we are particularly sensitive to other people's eyes and gaze behaviour (where and what they look at) (for reviews see: (Birmingham & Kingstone, 2009; Dalmaso et al., 2020; Emery, 2000; Frischen et al., 2007; Hamilton, 2016), and we can use gaze behaviour to infer a person's mental state and intentions (Argyle & Cook, 1976; Doherty, 2006; Shepherd, 2010). Gaze plays a special role in our social world as it has multiple functions (Gobel et al., 2015). From an observer's perspective, seeing where someone else is looking and what they are looking at can reveal the current object of their attention. We can also decipher gaze behaviour, given the context, to predict their likely actions or to infer their intentions. Observing someone's gaze has been found to engage the same social regions in the brain as performing theory of mind tasks (Calder et al., 2002) and by four years of age, children can use an adult's gaze behaviour to infer their desires (Lee et al., 1998) and detect their lies (Freire et al., 2004). For example, if we are tasked with detecting objects in our environment and another person always looks at the wrong location, we are likely to think that they are deliberately trying to prevent us from reaching our goal. After all, if they always look at the *wrong* location, they must know the *right* location. In this way, by adopting a mentalistic understanding of a gaze based behaviour, we can 'see' the contents of other people's minds (Doherty, 2006).

However, if someone's gaze is directed at us, that makes us the subject of their attention, which can affect us in many ways, ranging from the physiological to the perceptual (for reviews see: Hamilton, 2016; Hietanen, 2018; Senju & Johnson, 2009). For instance, if someone continually looks directly at us during an initial encounter, we don't like them as much as someone who looks at us on a more occasional basis (Argyle et al., 1974). Even when experienced out of context, a relatively meaningless difference in a person's gaze, such as a looking at us (direct gaze) or looking away from us (averted gaze), can affect how we process information and perceive the person. For example, we are faster to categorise a person based on their gender if they make direct eye contact with us (Macrae et al., 2002), we remember people who look at us better than those that don't (Mason et al., 2004), and we rate people as more likeable (Kuzmanovic et al, 2016), attractive (Mason et al, 2005) and trustworthy (Kaisler & Leder, 2016) if they display direct rather than averted gaze. Additionally, a change in the direction of someone's gaze (i.e., direct to averted) can bring about attentional effects in the perceiver. When a person who is looking at us suddenly shifts their gaze to another location, we tend to reflexively re-orient our own gaze to where they are looking, and this can happen even if we know that the other person's gaze is not useful (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999). This automatic visual orienting effect can be useful in the real world, as it can help us to identify locations or objects of interest quickly. However, if the other person is not trying to cooperate with us, it can be used to deceive us, drawing our attention away from our goal or target (Emery, 2000). Together, this means that a person's gaze can influence both what we think about them and how we behave in response to them.

Gaze-cueing

The attentional effects of another person's gaze-behaviour on a perceiver have been examined robustly using gaze-cueing paradigms (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999) for a review see Frischen et al., 2007). In such tasks, the other person's behaviour is neither related to the task's rules nor the perceiver's goal, who is tasked with detecting a target as soon as it appears on screen. Thus, in such tasks, the gaze-cues are task-irrelevant and are present only as 'distractors'. Nevertheless, owing to the reflexive orienting effect brought about by a sudden shift from direct to averted gaze, gaze-cues interact with our performance during the task

by re-orienting our gaze just moments before our target appears. When a person's gaze-cue orients us to the imminent location of our target (i.e., a valid cue), our response times are facilitated, as our visual attention is already directed at the target's location when it appears. Conversely, when a gaze-cue orients us to the wrong location (i.e., an invalid cue), our response times are slowed, as we need to re-orient our attention to the other location to locate our target. The difference between response times for valid vs. invalid cues is known as the gaze-cueing effect.

The gaze-cueing effect has been examined under a number of different conditions (for reviews see (Birmingham & Kingstone, 2009; Dalmaso et al., 2020; Frischen et al., 2007). It is not modulated by whether a new face appears on every trial or the same face is used throughout the task (Frischen & Tipper, 2004) and the same cueing effect elicited by eyes is also found for arrows (e.g., Galfano et al., 2012; Hermens & Walker, 2010; Kuhn & Kingstone, 2009; Tipples, 2008), which suggests that the relationship between cue and outcome is not social but rather spatial or informational. However, while both eyes and arrows can influence our attention and goal-related behaviour in a similar way, that does not mean that we perceive or selectively attend to them equally. For example, when we are free to look where we choose, we attend to eyes more than arrows (Birmingham & Kingstone, 2009), objects previously cued (i.e., looked at) by faces, but not arrows, are preferred (Bayliss et al., 2006), and only (valid) gaze-cues enhance working memory (Gregory & Jackson, 2017). Arrows may, therefore, direct us towards locations, but only eyes can look at us and behave, which not only affects us but also grants us a window to the thoughts and intentions of the person behind the eyes (Colombatto et al., 2020; Marotta et al., 2018).

Incidental Impressions – Gaze

Gaze-based behaviour can be used to study incidental impression formation by exploiting the gaze-cue effect. While the cueing effect itself may be indifferent to the face making the cue, that does not necessarily mean that no information is processed about the face. After all, faces capture our attention and effortlessly generate impressions (Olivola et al., 2014), and we use gaze behaviour to extract meaning from the mind behind the eyes (Shepherd, 2010), suggesting that how an individual face behaves could affect how it is later perceived (Falvello et al., 2015). This idea was tested by Bayliss and Tipper (2006), who manipulated whether individual faces displayed wholly

helpful (valid – always cued the right location) or wholly deceitful (invalid – always cued the wrong location) gaze-cue behaviour while participants completed a gaze-cueing task[5]. After the task, participants were shown pairs of faces, one helpful and one deceitful, and asked to select who they thought was the most trustworthy. On average, faces who provided helpful gaze-cues were selected as the most trustworthy face. Additionally, most people did not explicitly recall the contingency and the gaze-cueing effect remained stable over the course of the task (Bayliss & Tipper, 2006). Together this suggests that while our own outcome in the task cannot be improved by any knowledge about, or learning of, the faces' behaviours, the faces themselves are nevertheless encoded in such a way that we form divergent impressions about their trustworthiness based on their behaviour. We learn that those who consistently and repeatedly mislead our gaze are not to be trusted.

This trust learning effect has been replicated across a number of similar paradigms (for a recent review, see Barbato et al., 2020). It has been found to extend to decision-making in one-shot trust games, where most people invest more money with faces that give valid cues than those that give invalid cues (Rogers et al., 2014). The effect on trustworthiness judgements remains when the 2-alternative forced choice trust task is substituted for scalar trustworthiness ratings of each face individually, and pre- and post-ratings of faces' trustworthiness are collected (Manssuer et al., 2015, 2016; Strachan et al., 2016, 2020; Strachan & Tipper, 2017). Faces used in these designs are pre-judged (out of sample) as being similarly neutral for trustworthiness, meaning pre- and post-task ratings can show whether the effect is specific to a decrease in trust for invalid faces, an increase in trust for valid faces, or both. While bi-directional effects have been observed, overall evidence to date tends to suggest that effects are driven by a decrease in trust towards deceitful faces (Strachan & Tipper, 2017). This fits with the notion that negative information about a person is more salient than positive information (Baumeister et al., 2001; Fiske, 1980; Rozin & Royzman, 2001; Skowronski & Carlston, 1989). Interestingly, this is observed despite the fact that any negative intention on the part of the gazer is entirely unknown and therefore must be inferred.

[5] Faces providing non-predictive (50% valid / 50% invalid) behaviour were also included in the task, but were not included in trustworthiness decisions.

That negative gaze-cue behaviour drives (negative) incidental impressions in this scenario also makes sense when considering the effect negative cues have on the perceiver's outcome. While valid cues do facilitate our response time, we would have located the target on our own, which may make their behaviour less overtly helpful. Whereas invalid cues bring about an involuntarily shift in our attention to the location where the target never then appears. Consequently, we make the wrong visual move and have to adjust our attention to complete our goal. As such, deceitful faces may come to be associated with 'error' trials (Behrens et al., 2008). We are tuned to detect negative or potentially threatening information (Bell & Buchner, 2012), and invalid cue givers may be even more noticeable when experienced in the presence of other faces who always provide helpful cues. This is supported by the finding that invalid faces are also better remembered than valid faces, suggesting they are better encoded (Bayliss & Tipper, 2006). Other studies have shown a memory advantage for untrustworthy looking faces (Rule et al., 2012) and cheaters (Tanida et al., 2003), and children remember people who are mean over those who are nice (Kinzler & Shutts, 2008). Together, this suggests that when we repeatedly experience someone whose behaviour negatively affects our outcome, we learn that they cannot be trusted.

This trust learning effect is particularly noteworthy for two reasons. First, faces' gaze-cue behaviour is presented alongside of, rather than as part of, the main (target detection) task. This means people are not directly attending to the faces as their attention is focused on another task, making impressions formed incidental (no social goal). Second, the faces are not referenced and are not behaving within any known or pre-defined rules. In order for us to work out the meaning behind a person's averted gaze, we need to understand the context in which it is made (Hamilton, 2016), which here must be deciphered or inferred. Together, these factors reduce both our attention *to* the faces and the evaluative content *of* the faces' behaviour, which ought to make any social processing about the people behind the behaviours more difficult. These studies have demonstrated, however, that we are capable of detecting and extracting social meaning from other people's behaviour even when its presence is ill-defined and we are otherwise engaged. However, in past work, gaze-cue behaviour has always been wholly helpful or unhelpful (i.e., wholly valid or invalid cue providers). This may facilitate the encoding of behaviours, as the difference between the two types of cues is categorically stark (i.e., 100% help vs. 0% help) and each individual's behaviour is entirely consistent (e.g., every encounter is helpful). Further, in real life, people may

not always be so reliable or may change how they behave towards us during the course of an initial encounter. It is important, therefore, to examine the influence of gaze-cue behaviour on incidental impressions under more varied and potentially more challenging circumstances. The current work contributes to this literature by exploring incidental impression formation when: people behave inconsistently over time (Chapter 2), perceivers are still developing their social skills (Chapter 3), and when others' social behaviours are incongruent with perceivers' outcomes (Chapter 4).

Chapter 2: *Inconsistent Information & Gaze*

Overview

This Chapter explores the influence of inconsistent gaze-cue behaviour on incidental impressions. After pilot work confirmed that people can form impressions of others even when their behaviour is not wholly (un)helpful (see **Appendix A: Chapter 2: Preparatory work**), the primary focus here was to examine how incidental impressions are impacted when others' behaviour changes over time. Across three Experiments, faces provided different levels of helping behaviour to begin with (i.e., varied in how many helpful cues they each displayed), then half way through the gaze-cueing task, some faces changed how much they helped (i.e., their ratio of helpful to deceitful cues increased or decreased). Based on previous research, it was expected that participants would form impressions of the faces during the first half of the task[6], but it was not known how a change, for better or worse, in helping behaviour would feed into overall impressions formed.

Inconsistent Information & Social Milieu

While considering and conducting the initial series of experiments, it was noted that not only could the consistent and straightforward nature of the relationship between each faces' cue validity and a trial's outcome facilitate the processing of individual faces, it could also interact with how the faces are perceived relative to one another. When wholly helpful faces are encountered in the presence of wholly deceitful faces, and vice versa, impressions of each could be influenced by the behaviour of the other; their social milieu. In other words, one face's negative behaviour could be highlighted or made more salient by another face's contrasting positive behaviour, and a contrast between faces may feed into differences relating to how they are individually perceived. To explore this idea, two additional Experiments were conducted to examine how the behaviour of the wider group impacts upon the perception of an individual encountered among the group.

[6] Impressions were not measured prior to any change in validity and were only measured once, after the gaze-cueing task. Capturing impressions mid-way through would have drawn explicit attention to the faces, something we wished to avoid, and altered participants' experience during the second half of the task.

Research Question

How does inconsistent behaviour influence social judgements and decisions?

Contribution to Research

This is the first gaze-cueing study to use inconsistent gaze-cue behaviour, and the first impression updating study to use gaze-cues to provide social behaviour. Results, therefore, contribute to both fields of research separately and bridge the gap between these hitherto disparate fields.

Results of all five experiments detailed in this chapter have been prepared for publication under the title:

> "*Incidental impression formation from visual social cues; the influence of inconsistent information and group context*" *(Newey, Koldewyn & Ramsey, under review)*

Work is presented here in its submitted form. Raw data and analysis scripts for each experiment are available from the project's Open Science Framework page.
(https://osf.io/bx3gj/?View_only=a696123ca1a5443abeb9b94b7a7777c4)

Contribution of Author

I conceptualised, designed[7], collected data for, analysed and wrote up the experiments detailed in this Chapter. The programming of experiments was done by Dr. Julia Landsiedel, to whom I am thankful. Experiments were written in Matlab and run in PsychToolbox.

[7] Designed insofar as the experimental conditions and paradigms. The underlying experimental method was taken from Bayliss & Tipper (2006), who manipulated faces' gaze-cue validity to explore effects on trust learning.

Abstract

A variety of social cues, including gaze behaviour, are used to form impressions of others. For example, if another's eye-gaze reliably helps or hinders us while we complete a task, we incidentally form a positive or negative impression about them. In real life, people are rarely so consistent in their behaviour, and they are often encountered in dynamic group contexts. To date, however, it is not yet known how incidental impressions based on eye-gaze cues are affected by either changes in target individual's behaviour over time, or by the group's behaviour. To better understand how impressions are formed when social behaviours change valence over time, we manipulated helping behaviour both at the level of the individual (Experiments 1-3) and the wider group (Experiments 4 & 5). We found no evidence that impressions were driven by initial behaviour (primacy effects). Instead, people tended to form impressions based on the most recent behaviour, with some influence from the overall, average behaviour. In addition, we found that individuals' behaviours appear to be viewed more or less favourably, depending on the behaviour of the wider group. Overall, we demonstrate that impression formation based on eye-gaze cues is not dominated by a single process, but instead reflects a complex product of cognitive mechanisms that integrate average valence over time, the direction of behaviour changes, the recency of observed behaviour, and the group context in which the behaviour is observed.

Introduction

Impressions of others play an important role in social interactions. For example, whether we think someone is trustworthy or deceitful influences how we might interact with them. But since people's social behaviour is dynamic and changes over time, it is important to detect variations in a person's behaviour and update our impressions accordingly. In addition, many social cues are often experienced peripherally to our primary focus of attention, meaning any social inferences we make about a person are incidental to any explicit social goal. For example, during a team meeting with colleagues, we form impressions based on a whole host of social cues that are not task-relevant, including the person's tone of voice, body language and eye-gaze patterns. Although links between social cues and impression formation have been studied previously, the factors that determine how gaze cues that change over time impact incidental impression formation are not well-established. Therefore, the current series of experiments investigates how sensitivity to changing eye-gaze behaviour influences incidentally formed impressions.

To investigate how impressions are updated over time, researchers often explore how initial impressions are updated when new conflicting or inconsistent information is received (Campellone & Kring, 2013; Chang et al., 2010; Cone et al., 2017; Delgado et al., 2005; Ecker & Rodricks, 2020; Fareri et al., 2012; Mann & Ferguson, 2015; Maurer et al., 2018; Mende-Siedlecki, 2018; Mende-Siedlecki & Todorov, 2016; Rezlescu et al., 2012; Shen et al., 2020; Wyer, 2010; Zarolia et al., 2017). Some studies show that initial impressions continue to influence behaviour when new information is ambiguous, though final impressions are updated (e.g., Delgado et al., 2005; Fareri et al., 2012). Others find that initial appearance (e.g., Rezlescu et al., 2012) or morality (e.g., Zarolia et al., 2017) based indirect impressions continue to exert influence on final impressions when direct experience is available, while others have shown a rapid reversal when new information prompts a re-evaluation of initial information (e.g., Cone et al., 2017; Mann & Ferguson, 2015). These findings demonstrate that in some circumstances (but not all), impressions can be revised following changes in trait-diagnostic behaviour.

A feature of many of these studies is that they involve written descriptions of behaviour, which provides highly salient, valenced, and categorical social information about individuals who are

also the focus of attention. For example, observers may witness someone do something unequivocally bad followed by something good (e.g. Mann & Ferguson, 2015). Such designs provide a strong experimental manipulation, while also being quite unlike how most impressions are formed in daily life, which rarely involve such complete reversals of heavily valenced and trait-diagnostic behaviour or the depiction of people as 'saints' or 'sinners' (Siegel et al., 2018). Indeed, social information that speaks to a person's character is rarely so unambiguously diagnostic and in the real-world, social behaviour often needs to be tracked over time and "decoded" before it can be used to infer people's dispositions. Further, impressions are often formed outside of our awareness (Uleman et al., 2008) or more incidentally, as 'by-products' of our other goals (Carlson & Mae, 2003), meaning our attention is often focused elsewhere as we process people's social behaviour.

Instead of using explicit descriptions of behaviour to study impression formation, an alternative approach has made use of more implicit, non-verbal social signals, such as eye-gaze cues. Gaze is a social cue to which humans are particularly sensitive (for reviews, see Argyle & Cook, 1976; Dalmaso et al., 2020; Emery, 2000; Frischen et al., 2007; Hamilton, 2016), which we can use to infer other's mental states, attentional focus and intentions (Argyle & Cook, 1976; Colombatto et al., 2020; Doherty, 2006; Gobel et al., 2015; Ricciardelli et al., 2002, 2013). Using a gaze-cueing paradigm (Driver et al., 1999; Friesen & Kingstone, 1998; Langton & Bruce, 1999), modelled on a visual attention cueing task (Posner, 1980), task-irrelevant faces that always provide valid gaze-cues (i.e., look toward the correct location) are judged to be more trustworthy than those that always provide invalid gaze-cues (i.e., look toward the incorrect location) (Bayliss & Tipper, 2006; Manssuer et al., 2015; Strachan & Tipper, 2015; Strachan et al., 2017, 2020; Strachan & Tipper, 2017). The effects of gaze-cue behaviour also generalise to impact trust-based decisions in economic games, such that participants risk investing more money with valid (helpful) faces than invalid (deceitful) faces (Rogers et al., 2014). These findings demonstrate that even when there is no impression formation goal, the observed behaviour is irrelevant to task completion, and the salience of trait-diagnostic signals is relatively low, incidental impressions about an individual's character are nevertheless formed and influence subsequent social behaviour.

Although these initial studies provide strong evidence that we can learn about others' trustworthiness through (task-irrelevant) eye-gaze cues, many questions remain unanswered. For instance, prior work has only used wholly valid or wholly invalid gaze-cueing manipulations. In other words, a given individual always cues either the correct (i.e., always helps) or incorrect (i.e., always deceives) location. In real life, however, people are rarely so consistent in their behaviour and also may change their behaviour to be more (or less) helpful over time, meaning impressions may be harder to form or require updating over time (Brambilla et al., 2019; Mende-Siedlecki, 2018). As such, it is not yet known how less categorical (e.g., helps some, but not all of the time) and evolving (e.g., helps less over time) gaze-cuing behaviour may impact impression formation. In addition, social cognition does not end at the level of the individual; we also process the behaviour of the wider group when forming impressions (Carlston et al., 2015; Cline, 1956; Lammers et al., 2018; Lee & Harris, 2013; Newman & Uleman, 1990; Simpson & Ostrom, 1976), which may impact upon how we perceive individuals within the group. Although the social milieu is likely to be important for understanding how social cues are interpreted in the context of impression formation, to date, eye-gaze cues have not been studied in relation to the social behaviour of the wider group.

The current work, therefore, seeks to probe the factors that contribute to impression formation both when gaze-cue behaviour changes over time and in light of complex group dynamics. More specifically, we investigate how incidental impressions are affected when the helpfulness of individuals' behaviour changes over time (Experiments 1-3) and how changes in group context over time can affect impressions of individuals (Experiments 4 & 5). Understanding the kind of factors that affect impression formation in such dynamic contexts is important, given that incidental impression formation based on unattended social information is ubiquitous and operates in nearly all social exchanges.

General Methods

Transparency & Openness

For each of the main experiments, primary research questions and hypotheses, planned analyses, target sample sizes and exclusion criteria[8] were pre-registered. A link to each pre-registration is provided in its relevant experimental section. Raw data and analysis scripts for each Experiment are available from the project's Open Science Framework page (https://osf.io/bx3gj/?view_only=a696123ca1a5443abeb9b94b7a7777c4). All manipulations and measures are reported, and all exclusions are specified within individual experiments.

Replication and Pilot experiments

Prior to conducting the main experiments, we first wanted to replicate a key result from an experiment on which the current work is based (Rogers et al., 2014) to check that our design would result in similarly-sized effects of gaze-cue validity on social impressions, despite our paradigm being slightly different. Rogers and colleagues (2014) used an object categorisation task, where participants identify the category to which the target belongs (for example, kitchen vs garage tools). In these designs, although the eyes provide a cue regarding where a target will appear, once the target is seen the participant must still additionally identify the correct category. Therefore, the judgement is orthogonal to the location of the target. We opted to use a target location version of the task, in which participants simply respond to the location of the target. In this version, although the faces are still incidental to the task (i.e., not necessary to complete the task), they do help (or hinder) participants' performance by identifying (or lying about) the target's future location. We selected this version to make the faces more genuinely 'social' through displaying directly helpful or deceitful behaviour. This means that while faces' gaze-cues are no longer task irrelevant, impressions formed are still incidental. Happily, despite the change in task, and a few other minor

[8] Participants who liked or invested the same amounts with all faces were pre-registered as 'response outliers' and their data was to be removed for that individual measure. Upon later reflection, however, this may be inappropriate, as not all people should be expected to show an effect. Removing such participants may, therefore, remove valid data. Pursuantly, we re-ran all analyses without removing these outliers, which included analyses in Experiments 1 (three participants) and 2 (one participant). Including these data had a negligible impact, and the direction and meaning of results remained unchanged. The decision was made, therefore, to report results in accordance with the pre-registration.

differences (see **Appendix A1: Replication Study**, p.163 for full details), the replication experiment produced similar between-condition effect sizes, with wholly valid faces being trusted much more than wholly invalid faces (original units £1.33; t(14)=2.82, p=.014, CI[0.32, 2.35], dz=.0.73).

The planned design of our main experiments also differed from prior work by using a range of gaze-cue validities. In other studies, while distractor faces sometimes displayed non-predictive cues (i.e., cues were valid 50% of the time), target faces always provided either only valid cues (i.e., they always looked towards the side where the target would appear) or only invalid cues (i.e., they always shifted their gaze to the side opposite of where the target would appear). Our design required validities that varied both among faces and over time. In addition, gaze-cue validities that vary between 100% and 0% better reflect social behaviour in the real world, where people are rarely fully consistent. We piloted our paradigm to confirm that people are sensitive to a range of gaze-cue-validities and, in general, found that people's impressions can reflect a graded pattern of gaze-cue behaviour (see **Appendix A2: Pilot 1**, 164 for full details[9]). We found a clear and similarly sized difference between wholly valid and wholly invalid faces (original units £2.00; t(14)=2.69, p=.018, CI[0.41, 3.59], dz=.69), together with a clear effect across all six conditions (f(5,70)=5.37, p<.001, np^2=.28). This preparatory work provided confidence that the basic experimental design was working as anticipated. Importantly, it also provides evidence that people are sensitive to non-categorical social behaviour and reflect that trial-to-trial inconsistency in their final impressions.

Ethical Approval

All experiments were granted ethical approval by the School of Psychology at Bangor University's Ethics Committee, and all participants were Bangor University students who provided written consent prior to data collection taking place. In return for their time, participants received either course credit or a small cash payment (£4).

[9] A second pilot study was also performed, though for the sake of brevity, it is not reported in the submitted paper. The results can be seen in **Appendix A3: Pilot 2, p.165.**

Materials & Stimuli

Face stimuli

Six young, white adult male faces, each displaying a neutral expression, were used in all experiments to depict the various behavioural patterns. The faces, together with adapted versions to portray averted gaze to the left or right, were kindly provided by Strachan and Tipper (2016) who used them in a very similar gaze-cueing design and collected baseline ratings of attractiveness and trustworthiness for each face. These six (among others) were matched on both judgements. Faces (ids: AM 12, 16, 24, 25, 26, & 35) were originally taken from the KDEF face set (Lundqvist et al., 1998).

Gaze-cueing task

A gaze-cueing task was used to portray targets' behaviours (see **Figure 1**). Participants were asked to keep one finger from their left hand over the "c" key and one finger from their right hand over the "m" key and their task was to press the correct key when the target appeared; "c" for on the left, "m" for on the right. Participants were instructed to respond as quickly and as accurately as possible when the target appeared and instructions were provided on screen before the task began. There was no mention of the faces. Each trial began with a central fixation cross for 500ms followed by a forward-facing face for 1,500ms, which served to promote joint attention. The face then averted its gaze either to the left or right and after 450ms a letter "T", which served as the target, appeared to the left or right of the face. Once the target appeared, the screen remained the same until the participant made a response or 2,000ms lapsed, whichever came sooner. At this point the target disappeared and the face returned to looking forward for 1,000ms.

Valid trials were those where the gaze-cue direction matched the eventual location of the target. Invalid trials were those where the gaze-cue direction did not match the eventual location of the target. As such, we refer to valid cues as being helpful because they provide accurate information and invalid cues as being deceitful because they provide inaccurate information. Any change to a face's validity (helpfulness) over time always occurred halfway through the task; between the second and third blocks.

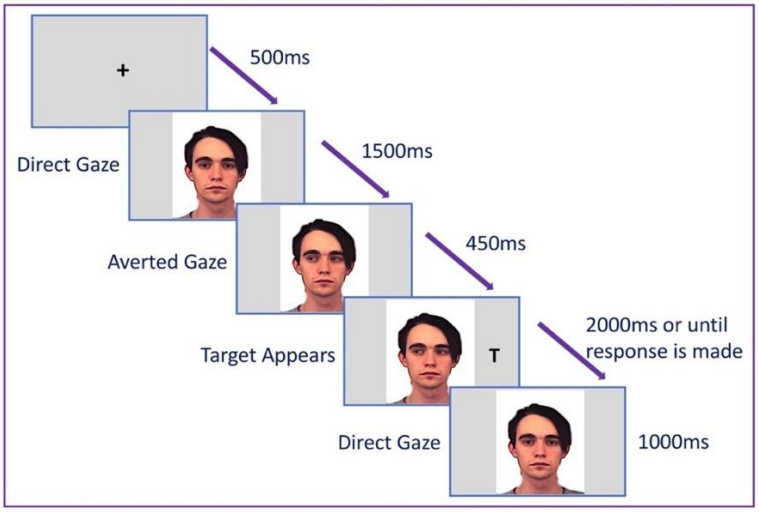


Figure 1*: An example of an invalid gaze-cueing trial, where the target appears on the side opposite to the location gazed at by the cue provider (ID: AM24)*

The task was structured into four blocks, with each block including 48 trials, for a total of 192 trials. This resulted in each of the six faces being presented eight times per block and 32 times overall. Each face looked to the left and right equally, and the total number of valid and invalid trials (across all faces) were equal across blocks and overall (with the exception of Experiments 4 and 5). To reduce anticipation and/or carry-over effects, no more than three trials of the same type (i.e., valid/invalid) could occur in succession. Blocks were pre-constructed so that trial order was the same for every participant within a particular block. Blocks 1 and 2 always came before blocks 3 and 4 but the order between blocks 1 and 2 and between blocks 3 and 4 was counter-balanced across participants. Most importantly, which face represented which condition was also counterbalanced across participants, ensuring that any effects of the characteristics (e.g., attractiveness, small differences in facial expression) of particular faces were minimised, as the same face that was always deceitful in the experience of one participant might provide the most help to another. The task took around 12 minutes to complete. To allow participants to rest their eyes, there was a break between each block where the task paused until the participant chose to continue.

Trust Decisions

Our primary dependent measure was taken from the trust (investment) game (Berg et al., 1995). In the 'one-shot' variant used here, participants had hypothetical £10s and were allowed to invest any amount from £0 to £10 with each trustee (face) only once. Participants were informed that any amount invested would be multiplied by three and the trustee would then have full discretion as to how much, if any, they returned to the participant. In this way, the task measures trusting behaviour, as any amount invested risks being lost and is, therefore, sent with an expectation that the other will reciprocate (Berg et al., 1995; Rousseau et al., 1998).

In order to check that participants understood the task and had the opportunity to ask questions, they completed two practice trials where they were shown two male faces (not included in the gaze-cueing task) and asked to decide how much they would invest with each one prior to the computerised task. Given the hypothetical nature of the task, participants were asked to act as though the decisions they were making had real consequences and would actually result in them winning or losing money. In this version of the game, often referred to as a 'one-shot' trust game, only the first 'investment' step was used and the trustees' responses were not revealed.

Face judgements

In addition to a social (trust based) decision, a simple social judgement measure was also collected. On a computer screen, participants were shown all the faces that they had seen during the gaze-cueing task and were asked to rate how nice they thought each face was using a 7-point Likert scale, where 1 represented "not nice at all" and 7 represented "very nice". As the primary measure of interest concerned trusting decisions, analyses for the niceness ratings are reported in supplementary materials. Effects were somewhat present for niceness ratings, although they were less marked than those observed for trust decisions (see **Appendix B1: Niceness Ratings**, p.167).

Procedure

Upon arrival, participants were informed they were going to complete three tasks; a reaction time-based computer task, a face rating computer task and some questionnaires. Included questionnaires varied across experiments and, as they did not form part of any primary hypotheses, their data is

not presented here. First, the trust game was explained, where participants were given an instruction sheet and the premise was re-iterated by the experimenter before the two practice trials were completed. Second, the computerised tasks were completed, starting with the gaze-cueing task, which was described to participants as a reaction time task. Immediately after, each face presented during the gaze-cueing task was rated on niceness ("How nice do you think this person is?") Followed by the one-shot trust games ("how much would you like to invest with this person?"). To reduce anchoring effects, the order in which faces associated with each condition were presented was randomised across participants.

Before being debriefed, participants were asked if they had guessed the purpose of the experiment. While many said that the experiment was "looking at how we judge people's faces", none made the direct link between the behaviour of the faces during the gaze-cueing task and its effect on judgements. Indeed, none reported noticing that the faces differed in their behaviour or that some faces were 'helping' while others were not. The whole experiment, including ratings, took approximately 30 minutes to complete.

Data Analysis

Though the results of the gaze-cueing tasks themselves were not part of the main hypotheses, data was analysed to ensure that the task generated the expected gaze-cueing effect; namely that invalid gaze-cues generated longer reaction times than valid trials, and that overall error rates were low. Incorrect trials and those with reaction times of less than 200ms or greater than 1,500ms were removed prior to all analyses. Average accuracy and reaction time data was then calculated based on trial validity and block number. Participants who scored less than 80% overall accuracy were designated as 'gaze-cueing' outliers and were removed from all analyses. Though this was not pre-registered, accuracy rates are generally high for gaze-cueing tasks, so a low accuracy score is suggestive of poor attention and, for our purposes, means it is less likely that the faces' behaviours were implicitly encoded. The number of gaze-cueing outliers was very low (range: 0-4) across all experiments. Analyses of all Experiments' gaze-cue data can be found in **Appendix B2: Gaze-cueing task results**, p.170-172.

Prior to analysis, responders designated as 'response outliers' were removed from the social data. A participant was pre-registered as a response outlier if they responded with the same niceness rating or investment amount for all six faces presented. This means that a participant could be removed from niceness analyses but still be included in trust analyses. While we cannot be sure that repeated responses were not genuine, these data were not included in final analyses as they are more likely due to a participant not engaging with the task. The number of response outliers was very low (range: 0-3) across all experiments, affecting only Experiments 1 (three participants) and 3 (one participant). We also analysed the data without the removal of outliers and it made no meaningful impact on the primary results.

Following the removal of outliers, our general approach was to first use a RM ANOVA and then follow with paired contrasts using mean differences (effect sizes in original units) and 95% cis. We also report standardised effect sizes using Cohen's dz calculated as: $\mathbf{t / \sqrt{(n)}}$. More specific details on analyses are provided in each experimental section. Where violations of sphericity occurred, Greenhouse-Geisser adjusted values are reported.

Following proposals by Gigerenzer (2018), we avoid interpreting results based solely on p-values and binary distinctions between "significant" and "non-significant". Instead, we base the direction and strength of our inferences on a range of metrics, which include a mixture of descriptive and inferential statistics. Hypothesis tests and p-values are considered in combination with an estimation approach, which emphasises the importance of estimating effect sizes (in original and standardised units) along with a measure of precision using 95% CIs (Cumming, 2014). Further, we have embedded conceptual replication into the design of the Experiment by using the same basic experimental design over a series of experiments (Zwaan et al., 2018).

Experiment 1

Overview

This first Experiment explored how impressions are influenced when faces' gaze-cue behaviour changes over time. Behavioural profiles were manipulated such that individual faces provided either wholly (in)valid or inconsistent behaviour during the beginning of the encounter and then

either increased, decreased or maintained their proportion of valid (helpful) cues during the second phase of the encounter (see **Table 1**). This design created two distinct faces groups, within which overall (average) helping behaviour (i.e., proportion of valid cues) was the same (i.e., three 'helpful' faces and three 'deceitful' faces) but individual faces differed in when they provided their valid cues (i.e., helpful-neutral or neutral-helpful).

Based on both previous work and our pilot data, we expected participants would form impressions of the faces during the first half of the task (blocks 1-2)[10]. Here, our primary research question probes how changes in helping behaviour during the second half of the task (blocks 3-4) would affect overall impressions. Impressions may not be affected and may continue to reflect early behaviour (primacy); they may be updated to reflect later behaviour (recency); or they may represent an assimilation of information across the whole task (averaging). In the current experiment, where participants are experiencing social behaviour that (incidentally) helps or hinders them in completing a task quickly, we expect the most recent behaviour to have a greater impact on final impressions than initial behaviour (Campellone & Kring, 2013; Chang et al., 2010; Maurer et al., 2018). However, one consequence of providing all behavioural information in one uninterrupted interaction may mean that participants calculate a running average of social information as it is experienced and adopt an 'anchor and adjust' (Hogarth & Einhorn, 1992) or average-based approach (Anderson, 1965). A simple average would seem unlikely, however, as negative behaviours (i.e., deceitful gaze-cues) are usually more salient and heavily weighted than positive (Baumeister et al., 2001; Fiske, 1980; Skowronski & Carlston, 1989). For example, when people learn about others' past (im)moral behaviours, updating is greater for moral-to-immoral behavioural changes than immoral-to-moral changes (Mende-Siedlecki, 2018; Mende-Siedlecki, et al., 2013; Reeder & Coovert, 1986). The direction of change (i.e., helping-to-hindering vs hindering-to-helping) could, therefore, impact the size of differences in perception between changing and stable pairs of faces. Our design helps us probe the influence of initial vs. recent

[10] Impressions were not, however, measured at the half way mark. This would have highlighted the faces (and their behaviours, together with the possible meaning *of* their behaviours) and, therefore, changed how people experienced them during the second half of the encounter. This would interfere with the incidental nature of the impression and likely induce anchor and adjustment effects (Hogarth & Einhorn, 1992) rather than the more implicit processing that occurred in the first half.

information, the direction of change (helping to hindering or vice-versa), and the extent to which participants are simply averaging their experience over time.

Method

Power analysis

In acknowledgment of the fact that effect sizes are likely to be smaller when reproduced (Camerer et al., 2018; Open, 2015) and given that we are not only seeking to assess a range of validities, but also to see if there is an effect of the order in which they are witnessed (i.e. Changing validities over time), the experiments were powered to detect effects smaller than those returned from pilot data (see **Appendix Appendix A: Chapter 2: Preparatory work**, p.161). Using G*Power 3 (Faul et al., 2007), a main effect of order (i.e., comparing faces who decreased vs. Maintained vs. Increased helping behaviour over time) was targeted and an a priori sample size calculation for a repeated measures design with three conditions was performed. A Cohen's f of .33 (np^2= .10) was set with 80% power and an alpha of .05. This returned a requisite sample size of 46, which was rounded to 50. A sample size of 50 provides 80% power to detect (unadjusted) differences between pairs of conditions of dz= 0.40 and higher. To take account of potential outliers, 55 datasets were targeted. Pre-registration: https://aspredicted.org/blind.php?x=fs3mb2.

Participants

Fifty five (female 44, missing 1; mean age 21.8 years, SD 3.9) participants were recruited, each of whom was compensated with course credits.

Design

The proportion of total valid and invalid gaze-cues provided by each face was manipulated such that there were two groups comprised of three faces (see **Table 1**). The individuals in one group each provided 75% valid cues (mostly helpful) when averaged across the entire task, while the individuals in the other group each averaged only 25% valid cues (mostly deceitful). While all faces within a group had the same overall number of valid cues, the order in which they displayed

Validity %	Overall Helpful Group (75%)			Overall Deceitful Group (25%)		
	Decrease	Stable	Increase	Decrease	Stable	Increase
Initial (Blocks 1&2)	100	75	50	50	25	0
Final (Blocks 3&4)	50	75	100	0	25	50
Average Validity	***75***	***75***	***75***	***25***	***25***	***25***

Table 1*: Gaze-cueing task conditions for Experiment 1****.***

their helping behaviour varied. One decreased its valid number of cues over time, one increased its valid cues and one maintained a consistent (albeit not necessarily wholly (in)valid) level of helping behaviour over time. This resulted in six unique gaze-cue profiles with some shared characteristics.

Data analysis

Levels of trust were analysed using a 2 (helping group: 75% average help vs 25% average help) by 3 (order: decrease, stable, increase) repeated measures Anova and paired contrasts. Of the 55 participants tested, three were removed as gaze-cueing outliers, resulting in 52 participants being included in the trust, niceness (see **Appendix B1: Niceness Ratings**, p.167**)** and gaze-cueing analyses (see **Appendix B2: Gaze-cueing task results**, p. 170).

Results

The first thing to note is that results do not support a primacy effect; the highest investments are not associated with the decrease (i.e., started helping) conditions (and vice versa) (see **Figure 2**). The second is that the mostly helpful face who, while not always helping (75% valid), did not change how much they helped over time (consistent), appears to be the most trusted, even among its group who averaged the same amount overall, and compared to the group member who increases to help the most in the second half. This suggests that neither initial, average nor recent behaviour clearly determines overall impressions formed.

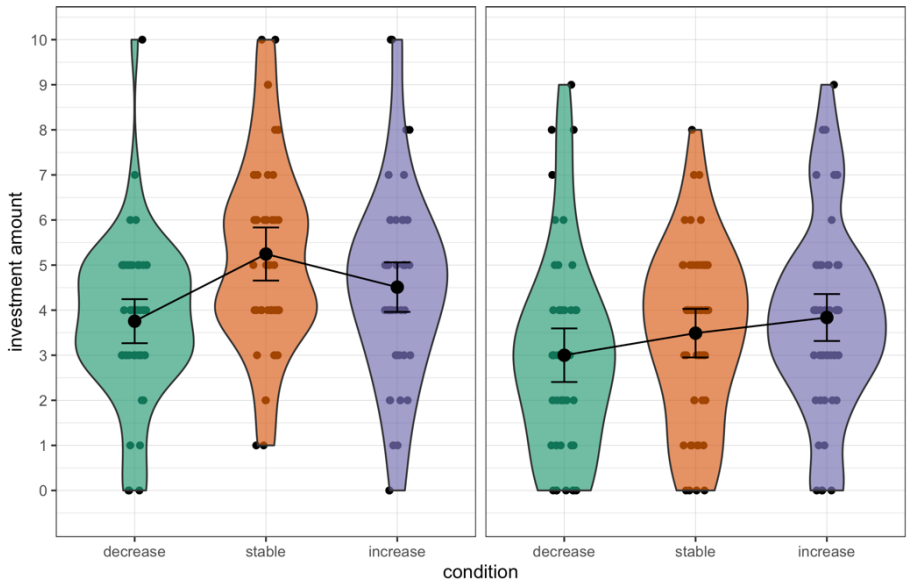


***Figure 2**: Violin plots showing average investments for Experiment 1. On the left is the overall helpful group and, on the right, the overall deceitful group. Bars represent 95% CI.*

There was a main effect of group $f(1,48)=19.11$, $p<.001$, $np^2=.29$ on investment decisions, with the mostly deceitful group (averaging 25% help) receiving lower (M £3.44, SD 1.44, CI[3.0,3.9]) investments than the mostly helpful group (averaging 75% help) (M £4.50, SD 1.36, CI[4.1,4.9]). There was also a main effect of order $f(2,96)=7.11$, $p=.001$, $np^2=.13$. Main effects were partly qualified by a marginal group*order interaction effect $f(2,85)=2.83$, $p=.071$, $np^2=.06$. To unpack the direction of these latter effects, we visualise the distribution of responses per condition and compute differences between conditions within each group (helpful and unhelpful).

As can be seen in **Figure 2**, across both groups, investments were lower in the decreasing conditions than the other conditions. Sidak adjusted comparisons showed the decrease conditions received less than both the stable conditions (mean difference £0.99, CI[-1.7,-0.3], $p=.005$) and the increase conditions (mean difference £0.79, CI[-1.4,-0.2], $p=.011$). On average, faces that maintained their helping behaviour throughout the task were trusted no more or less than faces that increased their levels of help (mean difference £0.20, CI[-0.5,0.9], $p=.862$), though this was not the same for both groups. As can be seen in **Figure 2**, there was a clear difference between how the conditions in the helpful group were trusted $f(2,96)=8.00$, $p=.001$, $np^2=.14$, but the difference was less apparent for the unhelpful group $f(2,96)=2.45$, $p=.092$, $np^2=.05$. Adjusted comparisons (see **Table 2**) revealed no difference between investments made with the decrease and stable faces

in the mostly unhelpful group but a clear difference in the helpful group. There was also a slight difference between the stable and increase faces in the helpful group. In other words, stability was valued most in the helpful group but not in the unhelpful group.

Our original design aimed to explore averaging and order effects. However, some of the results could, at least partly, be attributable to recency effects. That is because in this design, the most helpful faces also ended with higher levels of helping behaviour than the deceitful faces. Likewise, the decrease faces also ended with lower levels of helping behaviour than the other faces. In order to assess whether recency might also partly explain our effects, we directly compared the two conditions that were in different groups and changed their behaviour in opposite directions (one decreasing, the other increasing) but who both ended by helping the same amount (50% validity; conditions 1 and 6, **Figure 2**). Despite their behavioural differences, there was no difference in trust based decisions (**Table 2**). Conversely, there was no evidence for a primacy effect, as the two conditions that started helping the same (50%) amount were not trusted the same (**Table 2**). This suggests that recency may have played a role in determining trust levels in addition to other effects. Recency alone, however, cannot explain all the results. In the helpful group it is the consistent face, and not the face that ends by always helping, who receives the highest investment.

	Mean difference	95% CI Lower	95% CI Upper	t	Dz	Sig.
Helpful decrease > Helpful stable	-1.40	-2.60	-0.21	3.61	.50	.011
Helpful decrease > Helpful increase	-0.71	-1.75	0.32	2.11	.29	.454[a]
Helpful stable > Helpful increase	0.69	-0.32	1.70	2.11	.29	.455[a]
Deceitful decrease > Deceitful stable	-0.46	-1.70	0.77	1.15	.16	.988
Deceitful decrease > Deceitful increase	-0.79	-1.76	0.18	2.51	.35	.209[a]
Deceitful stable > Deceitful increase	-0.33	-1.41	0.76	0.93	.13	>.99
Helpful decrease > Deceitful increase *	-0.08	-1.09	0.94	0.23	.03	>.99
Helpful increase > Deceitful decrease ^	1.42	0.12	2.73	3.33	.46	.022

***Table 2**: Sidak adjusted pairwise differences in trust decisions for Experiment 1. Note: * these conditions both ended with 50% validity (i.e., recency driven) and ^ these conditions both started with 50% validity (i.e., not primacy driven). [a] when unadjusted these contrasts are p<.05.*

Discussion

Experiment 1 demonstrates that overall trust levels appear to be the product of a cognitive system that can track the order and recency of behaviour, as well as integrate behaviours into an overall average. Rather strikingly, in our view, even in a social situation where the trait-diagnostic information is subtle and impressions are formed while performing another task, we show that character impressions are formed in a manner that integrates a range of diverse signals over time. We show no evidence that the initial behaviour displayed by an individual influences trust. In the context of eye-gaze cues that change over time, therefore, we show no evidence of a primacy effect on impression formation.

To further probe the relationship between the average, most recent, and order of behaviour on impression formation, we ran two subsequent experiments. A primary concern with Experiment 1 was that the design might have been too complicated for people to sufficiently distinguish between the six individual faces over time. For instance, there was little difference in validity between some of the faces at any one time (e.g., some only differed by 25% during any one block) and faces only changed their behaviour by 50%, meaning a once unhelpful face would still be relatively unhelpful. Failure to adequately distinguish between behaviours may have reduced the ability of participants to track changes. Therefore, in subsequent experiments, we simplified the design in two different ways. First, we reduced the number of profiles to three and increased the size of initial differences in validity between each profile to 50% (0, 50 & 100%) (Experiment 2). Second, we again used only three levels of helpfulness at any one time, but we also increased how much the inconsistent faces changed from 50% (e.g., 0-50) to 100% (e.g., 0-100), representing a full reversal in behaviour (Experiment 3).

Experiment 2

Overview

Experiment 2 was designed to probe the effects of recency in a more direct manner than was possible in Experiment 1, where recency effects were confounded with order effects. This Experiment used a simplified design with only three conditions, each with a different average validity and sequence of helping behaviour. Crucially, all faces ended by offering the same level

of helping behaviour (see **Table 3**). If, despite all ending the same, faces are viewed differently, it would suggest that faces' behaviour is processed and incorporated over time and not that only the most recent behaviour determines overall impressions. In this instance, the face that decreases how much it helps over time may be perceived the most (as it averages the most helpful behaviour) or the least (as it decreases how often it helps over time) favourably.

Method

Power Analysis

As with Experiment 1, a power analysis was performed for a 1*3 repeated measures analysis of variance. Again, the design required a minimum of 46 participants and in consideration of outliers, 55 were targeted (Pre-registration: https://aspredicted.org/blind.php?x=q5wq2w).

Participants

Fifty-two new volunteers took part in this Experiment (41 female; mean age 21.59, SD 4.84), each of whom was compensated with either course credit or a small cash payment (£4).

Design

The validity of the gaze-cues provided by the faces was manipulated such that there were three conditions and only ever three variations in cue validity in any one block (see **Table 3**). Each condition was portrayed by two faces. As with Experiment 1, faces in this experiment either decreased, increased or maintained their level of helping behaviour over time. In order to disambiguate recency effects from the order and averaging effects seen in Experiment 1, the difference here was that all faces ended by displaying the same pattern of gaze-cue behaviour.

Validity %	Decrease	Stable	Increase
Initial (Blocks 1&2)	100	50	0
Final (Blocks 3&4)	50	50	50
Average Validity %	***75***	***50***	***25***

Table 3*: Gaze-cueing conditions for Experiment 2.*

Data Analysis

Prior to the main analyses, data from the pairs of faces producing the same behaviours were compared using within subject t-tests. There were no meaningful differences between any of the pairs (all p's >.05, unadjusted), and they were collapsed to create the three conditions; decrease, stable and increase. The main test of interest was a 1*3 repeated measures Anova.

Of the 52 people tested, one person was removed as a response outlier from both the niceness and trust datasets. There were no gaze-cueing task outliers. This resulted in 52 participants being included in the gaze-cueing analyses (see **Appendix B2: Gaze-cueing task results**, p.170**)** and 51 participants being included in both the trust and niceness (see **Appendix B1: Niceness Ratings**, p.167**)** analyses.

Results

Visual inspection of the data indicates that the condition that decreased helping over time, but helped the most overall (M £4.54, SD 2.10), had slightly higher investments than the condition that increased how much it helped, but overall provided the least amount of help (M £3.94, SD 1.98) (see **Figure 3**). A one-way repeated measures Anova, though, revealed no differences in trusting decisions between conditions f(2,100) =1.59, p=.209, np^2=.03 and a Sidak adjusted comparison showed that the difference between the two conditions was too small to provide clear or convincing evidence for a difference (mean difference £0.60, CI[-0.30,1.50], p=.271, dz=0.23).

Discussion

When the task was simplified through the use of only three conditions and all faces ended with the same amount of helping behaviour (50%), there were no clear differences between conditions. The results, therefore, are not consistent with an averaging approach over time being dominant. Indeed, there was only a relatively small difference in trust between the two conditions that changed their helping behaviour over time and offered substantially different levels of helping behaviour overall (75% vs 25%). In this design, where all faces end by offering the same amount of (neutral) helping

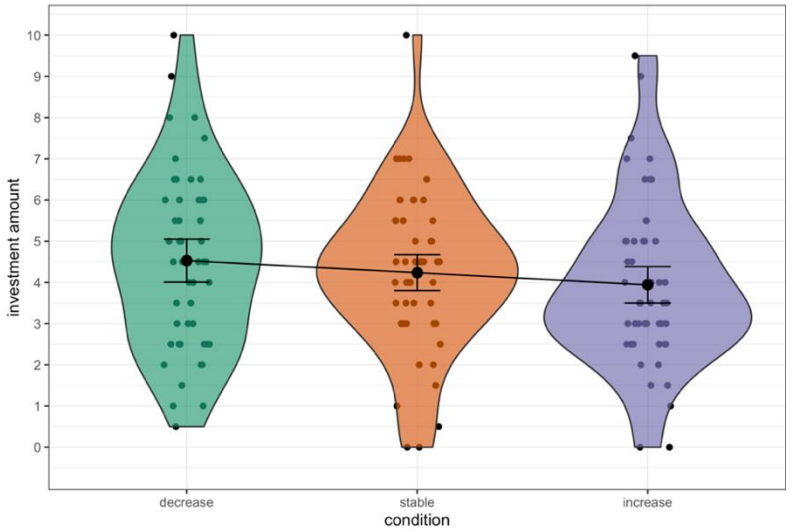


Figure 3: *Violin plots for trust decisions (investments) for Experiment 2. Bars represent 95% CI.*

behaviour, the results provide support for a recency effect with the potential for only a small influence from overall valence. In more explicit paradigms where individuals are initially presented as good, bad or neutral, before all are experienced as displaying neutral (50% positive behaviour) behaviour (e.g., Delgado et al., 2005; Fareri et al., 2012), impressions of the initially good person are downgraded to reflect their recent behaviour, however, they are still judged more positively than the original bad and neutral people. So, it is possible that there is a similar, albeit weaker, effect here, whereby the presentation of neutral information does not fully refute or invalidate a previous favourable impression.

We cannot conclude from these results that recency effects are always dominant during impression formation. By having a condition decrease to neutral and another increase to neutral, our design places averaging and order effects in direct opposition to each other, potentially reducing the influence of both. Thus, it is possible that participants' trust levels were the same across faces because of updated impressions rather than reflecting only recency. Additionally, although we simplified the design compared to Experiment 1, by presenting fewer gaze profiles overall, the change in behaviour between blocks 2 and 3 may not have been salient enough to strongly impact levels of trust. As was also true in Experiment 1, each changing condition only increased or decreased its helping level by 50%. Given the incidental nature of how impressions are formed in the gaze-cueing task, a more distinctive shift in behaviour may be needed to substantially impact

how the faces' behaviours are processed over time. Therefore, the dominance of recency that we observed here may, in part, reflect the way the experiment was constructed. To disentangle these issues, a third experiment was performed, which made both the shift in helping behaviour over time, and the difference between average helping behaviour overall, larger.

Experiment 3

Introduction

Experiment 3 used a simpler design than Experiment 1, while also making the change in behavioural valence between blocks 2 and 3 more distinct than Experiment 2. We aimed to identify if there were circumstances in which primacy, averaging or order effects might better explain impressions than recency effects. If, despite very different initial and average helping profiles, faces that end the same are trusted to the same degree, we would provide much stronger evidence that the most recent pattern of helping behaviour is the most dominant predictor of how impressions are formed. If, however, we find that there are other clear effects present, it would suggest that other effects can also contribute to the construction of impressions formed from inconsistent cue behaviour over time.

Further, for the first time in this series of experiments, wholly valid and wholly invalid faces that behaved consistently were included (see **Table 4**). This allowed for the conditions that changed (helpful-to-deceitful and deceitful-to-helpful) to be compared to those that remained helpful or deceitful. Thus, the magnitude of impression updating following a change in behaviour, and any differences in magnitude between updating positive and negative impressions, could be explored. This was of interest, because in studies using moral statements to provide indirect social information, updating for moral-to-immoral changes in behaviour is of greater magnitude than for immoral-to-moral changes, demonstrating that positive and negative impressions can be updated asymmetrically (e.g., Mende-Siedlecki et al., 2013). Here, the combinations of conditions that best approximate to the moral conditions are: 100-100 and 100-0 (helpful-to-deceitful) for moral-to-immoral, and 0-0 and 0-100 (deceitful-to-helpful) for immoral-to-moral. To explore this, the contrasts of conditions that start the same ('primacy' conditions, **Table 5**) can be compared to one another. If the difference in helpful-to-deceitful contrasts is larger than the difference in deceitful-

to-helpful contrasts, it would suggest that new negative information has a larger effect than new positive information.

Method

Power Analysis

The a priori power calculation for this experiment was based on a 1*5 repeated measures Anova. The previous two experiments were powered to detect medium to large effects and above. As effects detected in Experiment 2 were below that level, this experiment was powered to detect medium effects and above. At 80% power, with an alpha of .05 and Cohen's f of 0.25 (np^2=.06), this returned a requisite sample size of 50. In consideration of potential outliers, 55 datasets were targeted. Differences between specific combinations of conditions were also pre-registered (Pre-registration: https://aspredicted.org/blind.php?x=mn54eq). For paired sample t-tests (2 tailed), 50 participants provide 80% power to detect Cohen's dz of .40 and above.

Participants

Fifty-eight volunteers took part in Experiment 3 (44 female; mean age 22.4, SD 3.1), each of whom was compensated with either course credit or a small cash payment (£4).

Design

The previous two experiments involved changes of 50% in helping behaviour. Here, the changes that occurred were a full 100% reversal in helping behaviour. We also included three stable conditions, whereby individuals remained constant throughout at 100%, 50% or 0% validity (see **Table 4**). As a consequence, there were two pairs of conditions that ended by offering the same pattern of helping behaviour, but started by offering opposite levels of helping behaviour, allowing us to compare recency to averaging effects. Two faces offered the same, non-predictive behaviour throughout, and were collapsed into one condition to form the 'control'. By only having two faces change their behaviour over time (rather than four as in both Experiment 1 and 2), we hoped that the difference between the faces' behaviour would be more striking.

Validity	Wholly Helpful **'100-100'**	Decrease **'100-0'**	Increase **'0-100'**	Wholly Deceitful **'0-0'**	Control* **'50-50'**
Initial validity (Blocks 1&2) %	100	100	0	0	50
Final validity (Blocks 3&4) %	100	0	100	0	50
Average validity %	***100***	***50***	***50***	***0***	***50***

Table 4*: Gaze cueing conditions for Experiment 3. * This condition was displayed by two faces.*

Data Analysis

Data from the two control faces were compared prior to the main analyses. There was no significant difference between the two (p>.90) and so they were collapsed, creating five overall conditions. The test used to compare conditions overall was a 1*5 repeated measures Anova. The planned, main tests of interest were paired samples t-tests targeted to compare specific pairs of conditions; the conditions both ending at 100% validity and the two conditions both ending at 0% validity. As paired comparisons were planned, we used the least squares difference adjustment for multiple comparisons.

All 58 participants provided some variation in their responses, meaning none were removed as response outliers. Following a review of the gaze cueing data, four participants were removed from further analysis; three as their overall accuracy was less than 80% and one because of a technical error. This resulted in 54 participants being included in investment decisions, niceness ratings (see **Appendix B1: Niceness Ratings)** and gaze-cue task analyses (see **Appendix B2: Gaze-cueing task results**).

Results

A repeated measures Anova showed a main effect of condition f(3,163)=6.59, p<.001, np^2=.11 (see **Figure 4**). To directly address our research questions, we estimated the size of the differences between conditions by performing pairwise comparisons (**Table 5**). To assess the relative influence of primacy vs. recency, we used paired-sample t-tests to contrast the two pairs that started the same but ended differently (**Table 5** 'primacy' contrasts) and the two pairs that started differently but ended as the same (**Table 5** 'recency' contrasts). In line with the results from

experiments 1 and 2, there was no evidence for primacy. If the first information was driving impression formation, we would expect to find no differences between those that started the same, but ended differently. Instead, both the contrast between the 100-100 and the 100-0 conditions and the contrast between the 0-100 and 0-0 conditions revealed reliable differences (**Table 5**, 'primacy' contrasts).

On the other hand, if the most recent information was driving impression formation, we would expect no difference between those conditions that ended identically, regardless of how they started. Both the contrast between the 100-100 and 0-100, and the 100-0 and 0-0 conditions revealed only small differences, with effect sizes smaller than we can reliably detect (see **Table 5**, 'recency' contrasts). However, these differences do suggest that recency is not the only influence on impression formation. This is particularly evident when considered in "real" terms. Participants on average invested £0.76 more with the wholly helpful face than the face that only helped at the end, a difference that is nearly 8% of the possible investment and nearly 17% of the average investment.

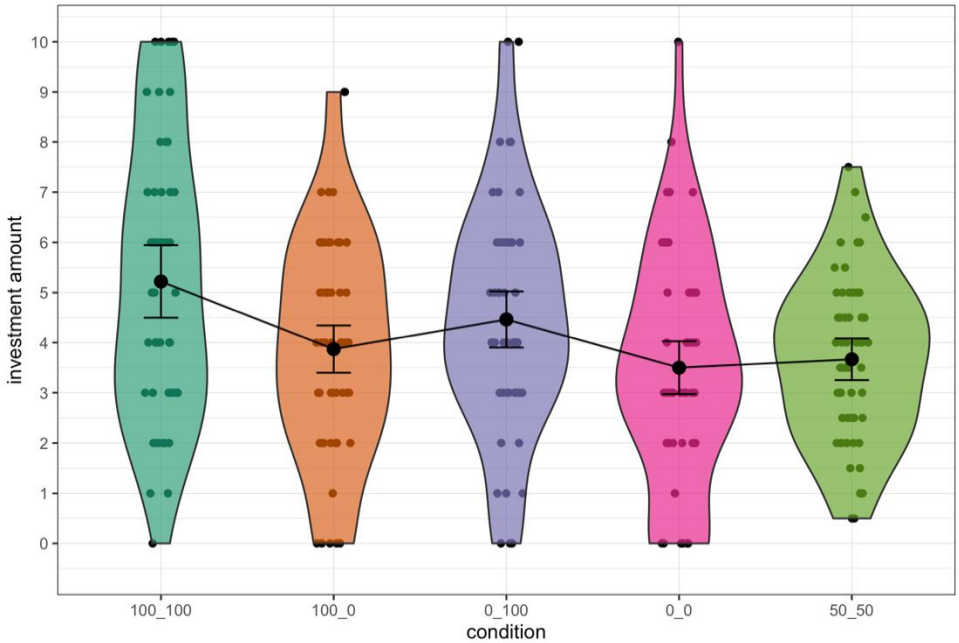


Figure 4: *Violin plot for Investment amounts for Experiment 3 where conditions show initial and final gaze-cue validity. Bars represent 95% CI.*

Similarly, we assessed the influence of averaging most simply by comparing the three condition combinations that averaged at 50% validity (100-0, 0-100, 50-50) (, 'averaging' contrasts). For the two changing conditions, the face that ended by helping was numerically trusted more than the face that changed to offer no help, but this effect was too small for us to reliably detect. The control face looked to be trusted less than the increasing (0-100) face (-80p), but very similarly to the decreasing (100-0) face (-20p). Indeed, the control and decrease conditions are trusted no more than the wholly deceitful (0-0) condition (**Table 5**, 'other' contrasts). This finding does not support averaging, recency, or order processes, but suggests that individuals who are neutral, or non-predictive, are viewed as negatively as those who are wholly invalid. Nonetheless, it is suggestive that a simple averaging mechanism alone cannot explain the current results.

Contrast	**Mean difference**	**95% CI Lower**	**95% CI Upper**	**t**	**dz**	**p**
Primacy						
100-100 > 100-0 ^	1.35	0.48	2.23	3.10	.41	.003 *
0-100 > 0-0 ^	0.96	0.19	1.73	2.51	.34	.015 *
Recency						
100-100 > 0-100 ^	0.76	-0.13	1.64	1.72	.23	.091
100-0 > 0-0 ^	0.37	-0.25	1.00	1.18	.16	.242
Averaging						
100-0 > 50-50	0.20	-0.66	1.07	0.69	.09	.999
0-100 > 50-50	0.80	-0.23	1.83	2.26	.31	.025
100-0 > 0-100	-0.59	-1.71	0.53	1.55	.21	.747
Other						
100-100 > 50-50	1.56	0.31	2.81	3.63	.49	.006 *
100-100 > 0-0	1.72	0.25	3.19	3.42	.47	.012 *
0-0 > 50-50	-0.17	-0.98	0.64	0.60	.08	.999

***Table 5**: Adjusted pairwise comparisons for Experiment 3. ^ these contrasts were planned and were not adjusted.*

Finally, the magnitude (degree of change) in impression updating was considered. As can be seen from **Table 5**, the difference between the 100-100 and 100-0 ('helpful-deceitful') conditions is larger than the difference between the 0-0 and 0-100 ('deceitful-helpful') conditions (i.e., comparing the sets of 'primacy' conditions). This suggests that positive impressions were decreased more when new behaviours were bad, than negative impressions were increased when new behaviours were good, and is in line with findings using indirect information about people's past (im)moral behaviours (Mende-Siedlecki, et al., 2013; Mende-Siedlecki & Todorov, 2016).

Discussion

In Experiment 3, when using a design that was simpler than Experiment 1 and provided more distinct changes in behaviour than Experiment 2, the results cannot be explained according to simple recency or averaging explanations alone. Rather, while recent behaviour is the best predictor of final impressions, the results suggest that encoding social information over time is nuanced and impression formation processes also depend on the overall valence of behaviour *and* the order in which it is provided. Indeed, unhelpful recent behaviour appears to over-write all previously helpful behaviour, but previous unhelpful behaviour may not be fully over-written by subsequent helpfulness. In other words, to be perceived positively, one must be consistently helpful, while negative behaviour impacts impression whether it is recent or not.

Across Experiments 1-3, we focused on an individual's social behaviour (gaze cueing behaviour). However, it is likely that impressions of one face have been influenced by impressions of other faces appearing in the same group (Carlston et al., 2015; Lammers et al., 2018). It would seem reasonable, therefore, that impressions formed from gaze-cue behaviour are relative and a product of both the behaviour of individuals and that of their associated group members. Experiments 4 and 5 test this assertion by presenting the same two behavioural profiles in two different social environments (link to pre-registration: https://aspredicted.org/blind.php?x=j8y5i4).

Experiment 4

Introduction

Experiment 4 investigated how consistent individuals are perceived when encountered in two directly opposed changing social environments. The valence of two target individuals' behaviour remained constant over time and across environments, but the behaviour of other group members changed both over time and across environments. In one environment, the overall levels of helping behaviour across group members improved over time, while in the other, group members' behaviour became less helpful over time. We predicted a contrast effect (e.g., Cline, 1956), whereby the targeted individuals would be trusted more in the environment where others become less helpful, rather than when they become more helpful.

Method

Power analysis

In keeping with previous experiments, a sample size of 55 was targeted. These designs were, however, 2*2 repeated measures designs, meaning we are only powered to detect larger effect sizes than those previously targeted. A sensitivity analysis for a two-level repeated measures Anova with an alpha of .05 calculated that partial eta squares of 0.14 (dz=0.38) or higher can be detected with 80% power.

Participants

A total of 64 (49 female, 10 male, five unknown; mean age 21.14, SD 5.21) volunteers began this two part experiment, although not all completed both sessions (see data analysis section). Demographic information was collected in the second session of the experiment, meaning that the information is not available for those participants who did not attend the second session. All participants were compensated for their time with course credit.

Design

Two faces maintained a consistent proportion of valid cues across two different environments. One individual was consistently helpful (75% valid) and the other was consistently unhelpful (25% valid). They were encountered in two different social environments, where faces either increased (improving environment) or decreased (worsening environment) how much they helped over time. The changeable faces that made up each environment were mirror images of each other, providing the same amount of help overall, but reversing the order in which they provided that help according to the environment (see **Table 6**). This created a situation where the ratio of valid to invalid cues differed both within and across blocks.

To be consistent with previous experiments, the same six faces were included in the gaze-cueing task and the same number of trials were included. This design thus required four faces to change their behaviour (improve/worsen) while the two target faces remained consistent over time. Consequently, there were mathematical constraints surrounding how the changing conditions could be represented in terms of their validity, and the selected design can be seen in **Table 6**. It was important that both the worsening and improving groups were the same in terms of overall helpfulness and differed only in terms of the order of their helpfulness. To achieve this, one condition (the middle condition that offered 50% help overall) of each environment was selected for repetition (see Table 6). This was a within-subjects design with the same participants experiencing both environments.

Validity	**Unhelpful Target**	**Helpful Target**	**Improve**	**Improve***	**Improve**
Initial (blocks 1&2)	25	75	0	25	50
End (blocks 3&4)	25	75	50	75	100
Average	25	75	25	50	75
			Worsen	**Worsen***	**Worsen**
Initial (blocks 1&2)	25	75	50	75	100
End (blocks 3&4)	25	75	0	25	50
Average	25	75	25	50	75

Table 6*: The table show the conditions and validities for Experiment 4 for the improving (top) and the worsening (bottom) environments. *This same condition was displayed by two faces.*

Procedure

The experiment was conducted across two sessions, with a period of at least three days between session one and session two. In the first session, the trust game was explained and practiced. Participants then completed either the improving or the worsening condition before rating each face for niceness and engaging in one-shot trust games. In the second session, participants completed the remaining condition and some questionnaires. The order in which the two tasks were completed was counterbalanced across participants.

Data Analysis

The two stable faces were compared across the two environments using a 2 (Environment: Improving, Worsening) *2 (Person: Helpful, Unhelpful) repeated measures Anova. To explore face and environmental effects separately, a series of planned paired samples t-tests were performed.

Due to the two-session nature of the experiment, there was some attrition, with 61 participants completing the improving environment, 59 the worsening environment and a total of 56 completing both. There were no response outliers. There was one gaze-cueing outlier and there were technical issues that resulted in data loss for four participants. These five participants' data was removed from all analyses, resulting in 51 subjects being carried through for final social judgement analyses (see **Appendix B1: Niceness Ratings**, p.169). Gaze-cue data were analysed separately for each task. In the improving environment, there was one outlier and missing data for two people, resulting in 58 datasets being analysed. In the worsening environment, the same outlier was removed and data was missing for four people, resulting in 54 datasets being analysed (see **Appendix B2: Gaze-cueing task results**, p.172).

Results

There was no effect of environment, with investments made in the improving environment (M 4.3, SD 1.7, CI[3.8,4.8]) being no different to those made in the worsening environment (M 4.3, SD

1.8, CI[3.8,4.8]), f(1,50) =0.05, p =.830, np^2 <.01. As expected, however, the helpful face that provided 75% valid gaze-cues (M £4.7, SD £1.9, CI[4.2,5.3]) received overall higher investment amounts than the unhelpful face that only offered 25% valid gaze-cues (mean £3.9, SD £1.8, CI[3.4,4.3]), f(1,50) =7.53, p =.008, np^2 =.13. Although numerically it does look as though the differences between the faces was more marked in the improving condition (**Figure 5**), perhaps due to a lack of power, the interaction was smaller than we could reliably detect f(1,50) = 2.68, p =.108, np^2 =.05.

In order to assess whether these potential differences are compelling enough to test in future work, we performed exploratory comparisons of simple differences between how the helpful and unhelpful faces were perceived within each environment (**Table 7**). For the improving environment, the helpful face was trusted more than the unhelpful face, whereas there was no meaningful difference in investments made to the two faces in the worsening environment. Planned paired t-tests were used to see whether the individual faces were judged differently in the two environments, but there were no clear effects present. Both the unhelpful and the helpful faces received similar amounts, irrespective of environment.

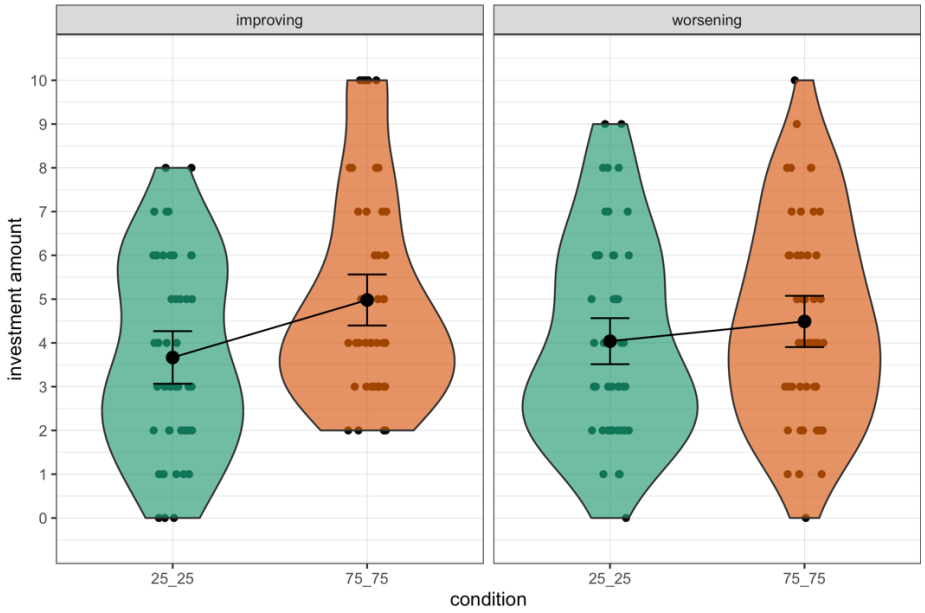


Figure 5: *Investment decisions for Experiment 4 for the consistent helpful (75% valid) and stable unhelpful (25% valid) faces across the two dynamic environments; improving (left) and worsening (right). Bars represent 95% CI.*

Contrast	Mean difference	95% CI Lower	95% CI Upper	t	dz	p
Within Environments						
25-25 Improving > 75-75 Improving	-1.31	-2.15	-0.48	3.15	.44	.003
25-25 Worsening > 75-75 Worsening	-0.45	-1.28	0.38	1.09	.15	.282
Between Environments						
25-25 Improving > 25-25 Worsening	-0.37	-1.16	0.41	0.96	.13	.344
75-75 Improving > 75-75 Worsening	0.49	-0.25	1.23	1.33	.19	.189

Table 7: *Pairwise comparisons between conditions both within and between environments for Experiment 4.*

Discussion

Our results show small differences in how people are judged based on changing group dynamics, but these differences were not entirely in line with our predictions. We predicted contrast effects, whereby the changing valence of the wider group environment would exaggerate the effect of oppositely valenced target individuals' behaviour on impression formation. Our results provide some evidence in support of this for the unhelpful person (25% valid) but show the opposite result for the helpful person (75% valid). Indeed, trust towards the helpful person appears to have been boosted by consistently being helpful in an improving environment. Our results give some, though certainly not unequivocal, support that group dynamics may operate on impression formation differently for individuals who are perceived positively vs. negatively. Perhaps unsurprisingly, group effects on impression formation, if truly present, are smaller than the effects due to differences in an individual's own behaviour. As such, our study was not properly powered to support (or deny) these smaller effects. Thus, at present, it is unclear what might be driving these possible group dynamic effects, and we encourage future research to probe the question further.

Experiment 5

Introduction

Experiment 5 used the converse approach to that of Experiment 4. Here, it was the groups' behaviour that remained consistent and two target individuals who changed their behaviour to help

more or less over time. As we designed and completed Experiment 5 in parallel with Experiment 4, we had similar general predictions. We expected contrast effects to manifest in two distinct ways. First, we expected both individuals to be perceived as less trustworthy in a helpful than an unhelpful environment, due to the fact that the overall unhelpful face, despite increasing how much it helps over time, is still helping less than other groups members and a face that helps less over time, despite helping a lot overall, is still deviating from the group. Further, we expected increased helpfulness over time to have a larger impact in the context of the unhelpful group than the helpful group (i.e., the face that increases in helpfulness will be judged more positively in the unhelpful environment). Likewise, we expected decreased helpfulness to have a larger impact in a generally helpful environment compared to an unhelpful environment.

Method

Participants

In return for course credit, a total of 57 people (female 40, missing 7, mean age 20.8, SD 3.4) took part in this study, although not all completed both sessions. Again, demographic information was collected in the second session of the study, meaning that the information is not available for those participants who did not attend both parts.

Design

Two faces varied in the proportion of valid cues they displayed overall and the order in which they provided their helping behaviour. The first face started by giving only valid gaze cues but then decreased to only provide 50% valid cues, resulting in 75% average helping behaviour. Conversely, the second face started by providing no valid gaze cues before increasing to offer 50% valid cues, resulting in 25% average helping (**Table 8**). These same behavioural profiles were used in Experiments 1 and 2, where we found no difference between how they were perceived. Any differences in how they are perceived in Experiment 5, then, are most likely to be due to the different social environments in which they are seen.

Validity	Target Decreasing	Target Increasing	Helpful	Helpful*	Helpful
Initial (blocks 1&2)	100	0	50	75	100
End (blocks 3&4)	50	50	50	75	100
Average	75	25	50	75	100
			Unhelpful	**Unhelpful***	**Unhelpful**
Initial (blocks 1&2)	100	0	50	25	0
End (blocks 3&4)	50	50	50	25	0
Average	75	25	50	25	0

Table 8: *Conditions for the consistently helpful environment (top) and the consistently unhelpful environment (bottom) in Experiment 5. * this condition was portrayed by two faces.*

Each face was encountered among two contrasting social environments, one where faces consistently displayed mostly valid cues (helpful environment) and another where faces consistently provided mostly invalid cues (unhelpful environment) (**Table 8**). Each block, therefore, had an unequal number of valid and invalid trials. Given the 'stable' nature of each Environment, this difference was constant across blocks, with more valid trials being present than invalid in the mostly helpful environment (2:1) and fewer being displayed in the unhelpful environment (1:2).

Data analysis

A 2 (person: increase, decrease) x 2 (environment: helpful, unhelpful) repeated measures Anova was used together with planned paired comparisons to estimate key differences between individual faces across environments.

Although 57 people attended session one, only 52 completed the unhelpful environment and 55 the helpful environment, with 51 completing both after data for one participant was removed as they accidentally completed the same version twice. There were no response outliers. One participant was removed as a gaze-cue outlier due to poor performance in the helpful environment, resulting in 50 being carried forward into the social judgement analyses (see **Appendix B1: Niceness Ratings**, p.169). After the removal of one gaze-cue outlier from the helpful environment,

gaze-cue analyses included 54 datasets for the helpful environment and 52 datasets for the unhelpful environment (see **Appendix B2: Gaze-cueing task results**, p.175**)**.

Results

In line with predictions, faces were generally trusted more in the unhelpful (M £5.5, SD 1.9, CI[4.0,5.1]) than the helpful (M £3.79, SD £1.56, CI[3.3,4.2]) environment f(1,49) =8.77, p=.005, np^2=.15. There was also an effect of person, with the increasing (0-50), yet least helpful, face (M £3.89, SD 1.6, CI[3.4,4.4]) receiving overall lower investments than the decreasing (100-50), yet most helpful, face (M £4.44, SD 1.8, CI[3.9,5.0]), f(4.52, p=.039, np^2=.08. This appeared to be driven by the least helpful face being trusted less in the helpful environment (see **Figure 6**). There was no interaction present f(1,49) =0.50, p=.482, np^2=.01, though this may be due to a lack of power to detect small differences.

Exploratory comparisons (**Table 9**) found that in the unhelpful environment, there was no substantial difference between the two faces. In the helpful environment, however, the increasing (overall unhelpful) face was trusted slightly less than the decreasing (overall helpful) face. Planned

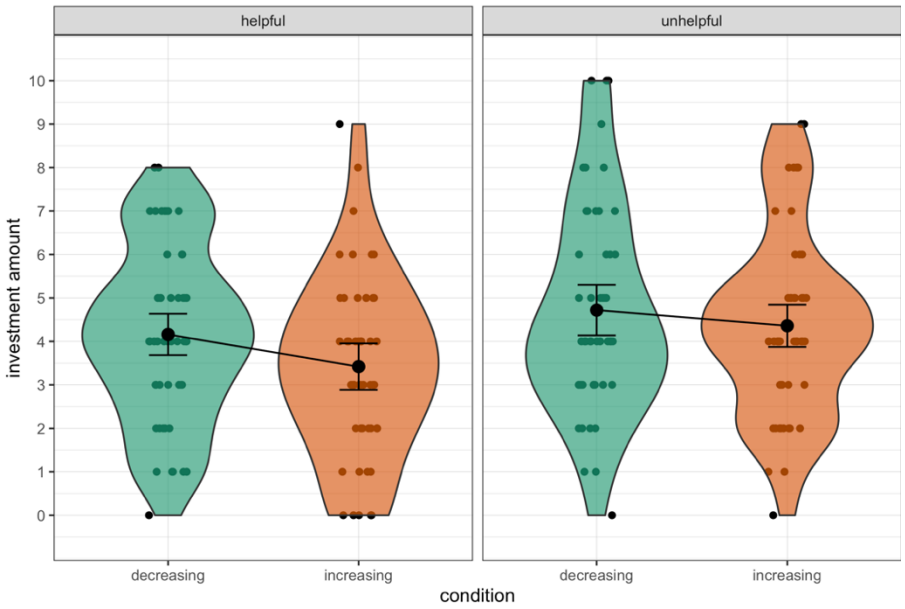


Figure 6: *Violin plots showing trust decisions for each target condition in Experiment 5 for the helpful (left) and unhelpful (right) social environments. Bars represent 95% CI.*

Contrast	Mean difference	95% CI Lower	95% CI Upper	t	dz	p
Within Environments						
100-50 > 0-50 (Helpful group)	0.74	-0.02	1.50	1.96	0.28	.055
100-50 > 0-50 (Unhelpful group)	0.36	-0.38	1.10	0.98	0.14	.334
Between Environments						
100-50 Helpful > 100-50 Unhelpful	-0.56	-1.30	0.18	-1.53	0.22	.133
0-50 Helpful > 0-50 Unhelpful	-0.94	-1.69	-0.19	-2.53	0.36	.015

Table 9: *Pairwise comparisons, both within and between environments, for the inconsistent, target conditions in Experiment 5.*

paired sample t-tests explored environmental differences for each face individually (**Table 9**). The decreasing face was trusted similarly in both the unhelpful environment, where it helped the most, and the helpful environment, where it did not, though the distribution of investments does seem to differ between environments (**Figure 6**). In contrast, the face that increased was viewed differently depending on the environment, as it was trusted more in the unhelpful environment.

Discussion

The two key profiles here (100% dropping to 50% vs 0% increasing to 50%) were also included in Experiments 1 and 2, where they were perceived similarly. Here, they are perceived differently, depending on the group environment. Taken together, they were trusted more in the unhelpful than helpful environment, and when group members' behaviour was mostly negative, both faces were again perceived similarly. However, when placed in a helpful environment of mostly valid cues, the face that (though increasing over time) provided the least help overall was trusted less than the decreasing (yet overall helpful) face.

This makes intuitive sense, given that in a helpful environment, helping is commonplace and so a face that deviates from the "common behaviour" may stand out more. It is unclear, however, why a similar contrast effect does not occur for the overall more helpful face in the unhelpful environment. Taken together, the results suggest a recency effect in the unhelpful environment and

an averaging effect in the helpful environment. However, much like in Experiment 4, larger sample sizes are required to adequately probe the effects of social context.

General Discussion

We demonstrate several ways in which incidental impression formation is sensitive to eye-gaze cues that change over time. Our results show that despite being task-irrelevant, people track and integrate inconsistent and varied eye-gaze behaviour over time, which they use to form nuanced judgements of people's trustworthiness. In addition, we demonstrate that in the context of non-verbal social cues, impression formation and updating is not dominated by a single cognitive process but is a complex product of many cognitive mechanisms that integrate average valence over time, the direction of behaviour changes, the recency of observed behaviour, and the group context in which the behaviour is observed. Our results, therefore, not only demonstrate just how tuned humans are to changes in non-verbal social information, but that our social assessments change flexibly in response to many different factors, including group context.

The results extend prior studies that used a similar gaze-cueing paradigm to manipulate trust (Bayliss & Tipper, 2006; Manssuer et al., 2015; Rogers et al., 2014; Strachan et al., 2017; Strachan & Tipper, 2017). Compared to such prior work, where people experienced only wholly valid and wholly invalid gaze-cues, we show that people are also capable of both tracking inconsistent and less categorical depictions of social behaviour, and of tracking *changes* in that behaviour over time. The most recent behaviour was generally the best predictor of subsequent impressions. However, we also show that the relative importance of any particular factor may be more or less important depending on the context and circumstances. Thus, among the factors that we manipulated, no one factor was dominant across all five experiments. Instead, a more complex picture emerged compared to prior research using the gaze-cueing paradigm.

Such added complexity was particularly apparent in two experimental contexts. First, in Experiments 4 and 5, we found some evidence that group context influences impression formation differently for individuals whose behaviour, in isolation, would be perceived as positive as

opposed to individuals who would be perceived negatively. This promising initial finding demonstrates that how we perceive an individual can be affected by behaviour of the wider group. Similarly, Lammers and colleagues (2018) found that the 'valence homogeneity' of others in the learning environment influences how an individual's behaviour is perceived. Second, in Experiments 1 and 3, faces that started out as helpful but became more deceitful over time were trusted as little as the individual that was consistently unhelpful. Thus, while new positive information does not over-write past bad behaviour, new negative information can almost entirely expunge previous positive behaviour. The observation that negative behaviour changes may have a larger effect than positive changes is consistent with previous research, which showed that faces providing only invalid cues were trusted less, but were remembered better, than valid cue providers (e.g., Bayliss et al., 2009; Bayliss & Tipper, 2006).

Our results can also be contextualised by considering the wider literature on impression formation that extends beyond gaze-cueing studies. Both primacy and recency effects have been documented in prior impression formation research exploring order effects. Such conflicting findings may be at least partially explained by task instructions (Forgas, 2011; Hogarth & Einhorn, 1992) and differences in the weight of negative and positive information (Baumeister et al., 2001; Fiske, 1980), where negative information received after positive information updates impressions more than positive information received after negative. Indeed, this asymmetry is the case in studies using moral statements, where impressions are asymmetrically updated depending on the direction of the valence change (Mende-Siedlecki, et al., 2013; Reeder & Coovert, 1986). Such studies have typically, though not always, used written descriptions of behaviour that are clearly trait-diagnostic. Here, when non-verbal eye-gaze cues are used as social signals, we find that the most recent behaviour generally has more influence than the initial behaviour on final impressions, though positive-to-negative changes are still more powerful than negative-to-positive changes. One possible practical benefit of using non-verbal and less trait-diagnostic cues is that they may be more easily integrated over time and therefore enable both more nuanced and more gradual changes in social behaviour to be studied.

A complementary interpretation for the dominance of recency over primacy in the context of eye-gaze cues is that recent information may be particularly powerful when the social cues are neither

the focus of attention nor strongly morally valenced. In the "real world", changes in eye-gaze cues are likely to be more indicative of a current and transient mood, which is more predictive of current behaviour towards us than enduring pro- or anti-social traits (Frith & Frith, 2006). Consequently, the most recent eye-gaze behaviour may dominate because of its superior functional relevance in this context. An alternative possibility for stronger effects of recency is that transient eye-gaze cues may be less readily encoded and/or tracked explicit and highly-valenced cues, such as written descriptions of behaviour. However, there are reasons to believe that rapid decay of incidental impressions from eye-gaze cues is not the sole driver of these results. Recent studies have found that the eye-gaze cueing effects on impression formation survived for up to a week, though the strength of the effect faded (Strachan et al., 2020; Strachan & Tipper, 2017).

Limitations and Future Directions

Some of the smaller effects that we observed need to be replicated in a future study that is properly powered to their size. However, we also expect that effects that are small in the context of a lab-based experiment could have potentially larger cumulative effects over time in real social interactions (Funder & Ozer, 2019). Of course, this needs to be tested empirically in future work, but we consider this to be particularly possible with incidental social cues as they are encountered many times a day, and our data strongly suggests that people are not only exquisitely sensitive to such cues, but also capable of tracking changes in such cues across time.

It is also important to note that participants did not perceive the "neutral", non-predictive (50-50 valid/invalid), behaviour as neutral. Instead, they generally judged such behaviour almost as negatively as they did individuals who were providing invalid cues a majority of the time. While we calculated the gazer's behaviour to be mathematically neutral, in reality, non-predictive behaviour is wholly unhelpful, as it can never be relied upon to make predictions. In this way, it makes sense that people would form negative opinions about such individuals. We acknowledge, therefore, that there may be no "neat" way to produce truly neutral behaviour when using this sort of task.

A further consideration is that we did not use an incentive compatible design for our trust measurement (Mailath, 1987). In other words, participants did not actually receive real monetary rewards. As a consequence, investment decisions might have been different if they actually stood to lose or gain money. While we accept that absolute investment amounts may have been different had we included a monetary incentive, we believe that the relationships found between conditions are still valid. Indeed, in our pilot work (see supplemental material), we found effects of a similar size to those found in work that did use an incentive compatible design (Rogers et al., 2014) and we consistently found large differences between valid and invalid faces in the expected direction and in spite of no risk or reward.

Finally, while the design used here is good for providing social behaviour covertly, it is not well-suited for teasing apart what kind of cognitive mechanisms underpin trait impressions from incidental eye-gaze cues, nor what element of the behaviour is driving incidental impression processes. One possibility is that behaviour directly drives impression formation. For example, invalid cues are perceived as deceitful, leading to the inference that the person making such cues is untrustworthy. Alternatively, the consequences of cue behaviour for participants could associatively drive impressions (Balliet et al., 2011; Behrens et al., 2008). For example, invalid cues slow participants' responses, so faces that provide invalid cues could be associated with a 'bad' outcome, or the task being more difficult. In this scenario, it is the cues relationship with our outcome that drives impressions rather than a judgement about the meaning of the behaviour itself. In our paradigm, these two possibilities are confounded. This distinction should be considered in future work.

Conclusion

Our results demonstrate that impression formation and updating is both complex and flexible, sensitive to how eye-gaze cues change over time in a variety of ways, as well as to the group context in which such cues are observed. Our results strongly suggest that we need a more comprehensive theory for how incidentally encountered social cues are integrated with contextual cues and more explicit social knowledge to impact on impression formation across time. As such, we feel this work reflects a compelling starting point for future research.

Chapter 3: Incidental impressions & Children

Overview

While reviewing the gaze-cueing and impression formation literature, it was noted that research concerning children had yet to be conducted. The ability to detect, extract social meaning from, and infer a person's intentions from their gaze-cue behaviour represents a complex form of social cognition and it is possible that it is a skill we evolve in the later stages of our social development. This Chapter was motivated, therefore, by the question of when the ability to form incidental impressions from gaze-cue behaviour matures. To explore this issue, an adolescent sample was targeted, as this period is both an important stage for developing higher social skills, which underlie social decision making (Blakemore ; Blakemore & Choudhury, 2006; van den Bos et al., 2010) and is marked by extensive changes in social behaviour (Kilford et al., 2016).

Background to data collection

Data was collected at two schools in the North Wales area. Although the parents of over 80 pupils at the first school were contacted, unfortunately less than half responded. Data was collected across a two-week period from 30 pupils: 17 in the 'younger' (11-13 years) age category and 13 in the 'older' (14-15 years) category. To increase the sample size, another school was approached where the first author had previously forged a relationship with the GCSE psychology teacher. Given the age-range of students in the class, we were only able to recruit students in the older age-range. In addition, data could only be collected during class-time, meaning that data collection was done in noisier and potentially more distracting conditions than the first school. Although this presented less than ideal testing conditions, allowing for only two participants to be tested at a time, I was able to collect a further 12 datasets. Unfortunately, at this point, pandemic restrictions made it impossible to continue. The original intention was to recruit at least twice the sample reported here. With the current sample size, only larger effects can be sensibly interpreted (see 'statistical power' section below).

Research question

Do children form incidental impressions from others' gaze-cue behaviour?

Contribution to Research

This is the first developmental study to investigate whether children extract social meaning from others' helpful and deceitful, and inconsistent, gaze-cue behaviour and form incidental impressions.

Contribution of Author

I conceptualised, designed, collected data for, analysed and wrote up the experiment detailed in this Chapter. An edited version of the gaze-cueing task used in Chapter 2 was used here.

Abstract

Gaze behaviour provides information about others' mental states and intentions that people readily interpret. While even infants routinely attend to another's gaze, the ability to extract the social relevance and meaning from others' looking patterns, as with other social skills, takes years to develop. Adolescence marks an important stage of development, as social relationships take centre stage in teenagers' lives, interactions become more complex, and social skills are refined. It is important, then, that adolescents are able to extract and use social information proficiently, even when attending to something else. This study explored trust learning from valid (helpful) and invalid (deceitful) gaze-cues in two adolescent age groups; 11-13 and 14-15 years. Preliminary results show that although both age groups demonstrate a reaction time based gaze-cueing affect (attentional effects), only the older age group extrapolated faces' behaviour in the task when making social judgements and decisions (social effects). Despite some faces helping more during the task than others and some faces changing how much they helped over time, the younger group rated all faces similarly for niceness and invested similar amounts with each in one-shot trust games. The older group, behaving more like adults, differentiated between the faces, both liking and investing less with the deceitful faces. Evidence tentatively suggests that the ability to detect faces' individual behaviours and make social inferences about their character is still in development during early adolescence.

Introduction

Adolescence marks the period between puberty and adult independence (Blakemore & Mills, 2014; Patton et al., 2016), and is a time period marked by rapid and intense social development, both behaviourally and neurologically (Blakemore & Choudhury, 2006; Choudhury et al., 2006; Kilford et al., 2016; Steinberg, 2005). During childhood, the social environments in which we find ourselves are often at least partially directed by our caregivers. As we progress into adolescence, however, we begin to take control over our social lives, make our own social decisions and judgements, and choose with whom we want to interact and socialise. Importantly, especially when encountering new people, adolescents need to independently discern people's characters and intentions, and decide if they can be trusted (Ames et al., 2011; Malle, 2011). One crucial step to becoming socially independent, therefore, is to develop the socio-cognitive skills necessary to use social cues to assess the trustworthiness of the people we meet so that we can safely and successfully navigate social situations (Frith & Singer, 2008; Renfrew et al., 2008). At the same time, the development of such interactional social-cognitive skills is particularly protracted in humans, continuing into early adulthood (Beaudoin & Beauchamp, 2020). Understanding how and under what circumstances adolescents are able to form impressions of others from social cues is, then, particularly important, as we know that adolescents are especially tuned to, and influenced by, social information (Crone & Dahl, 2012), and yet the skills needed to safely navigate their newly complex social landscape may not yet be fully developed. Further, whereas children can be guided or directed to useful (or directed away from potentially dangerous) social cues in their environment, adolescents need to detect and interpret this information for themselves.

The bulk of the literature looking at impression formation in adolescents uses tasks where participants directly attend to the target person. Further, the social behaviour they experience is directly relevant to the task they're engaged in and is highly diagnostic of the target's intentions (e.g., Lee et al., 2016; van den Bos et al., 2010). For example, in the trust game (Berg et al., 1995), there are clear rules whereby the target has the choice to share or steal from the participant. If a target repeatedly fails to reciprocate a participant's cooperative behaviour, the target has chosen to make the participant's financial situation worse, and their behaviour is simply and transparently negative. In real life, however, others' cues are often not so neatly presented or highlighted. Firstly,

we rarely set out to form impressions in case they are of use later and secondly, our attention may not be focused on the individual in question. Additionally, the meaning behind someone's social behaviour may not be so obvious and/or may require additional inferences to connect observed social behaviour to intentions or traits. As adolescents progress in their social worlds, they need to learn to deduce social meaning for themselves.

While it is difficult to capture "real-world" scenarios in the lab, a commonly used task, the gaze-cueing paradigm (Driver et al., 1999; Friesen & Kingstone, 1998), connects a powerful social cue (i.e., gaze) to trust-based impressions (Bayliss & Tipper, 2006; Rogers et al., 2014). In a gaze-cuing paradigm, faces provide either helpful (valid) or deceitful (invalid) gaze-cues by looking either towards or away from, respectively, where the perceiver's target appears shortly afterwards. Explained as a visual attention task, faces are neither referenced nor intrinsic to the task or the participant's goals, thus offering a unique way for exploring how we learn about other people by presenting social information outside the focus of attention. Targets' (gaze-based) behaviour can be provided relatively covertly (i.e., gaze is visually salient but is not directly task relevant) and any impressions formed are incidental (i.e., there is no explicit social goal in the gaze-cuing task itself). For both adults and children (Frith et al., 2003; Goldberg et al., 2008; van Rooijen et al., 2018), response times are facilitated by helpful (valid) and hindered by deceitful (invalid) cues. Adults both judge faces that give deceitful cues as less trustworthy (Bayliss & Tipper, 2006; J. Strachan & Tipper, 2015) and trust them with less money in trust-based decisions in economic games (Rogers et al., 2014). While clearly an artificial scenario, the gaze-cueing paradigm allows researchers in the lab to create three important elements of many real-life social experiences. First, gaze provides fleeting social cues that are not immediately and singularly diagnostic of trustworthiness, requiring them to be tracked/encoded over time to form an overall impression of the person. Secondly, unlike, say, words, which have a declarative meaning, visual social cues need to be decoded in reference to the context to be understood. Third, gaze-cues are readily perceived, despite not being vital to task performance or at the centre of participant's attention. The task also allows researchers to flexibly control "social" behaviour by changing the proportion of helpful vs. deceitful gaze cues. The result is a task capable of measuring social cognition under more complex conditions.

To make inferences about a person's character and make predictions about their future behaviour, all from only gaze behaviour, requires a number of increasingly complex cognitive judgements. Simply, a person's gaze conveys where, and at what, their attention is directed. This attentional information, combined with contextual information, can then be used to infer the person's goals; *why* are they looking that way? In other words, using another's gaze behaviour to identify what they are thinking and/or their intentions. These cues can then be tracked over repeated experience with that person and this behaviour over time can speak to their motives and character. In the gaze-cuing task, if they always look in the wrong direction, do they *intend* to hinder performance? If they always look in the correct direction, are they trying to help? These intentions can then allude to underlying character; if they have negative intentions, they are untrustworthy, if positive, then we are likely to judge them as more trustworthy. Finally, we then use these impressions of trustworthiness to make decisions about how to interact with that person in the future. In this way, we require a mentalistic understanding of gaze (Doherty, 2006) that we combine with contextual information (Hamilton, 2016) to interpret someone's gaze behaviour.

The recognition of gaze as an important social cue develops early. At just eight months old, infants have been shown to track the reliability of individual faces' cue behaviour (head movements and accompanying eye shifts), by visually attending to locations based on the cue providers' previous reliability (Tummeltshammer et al., 2014). Infants first learned about the reliability of a reliable (always cued the right location) and an unreliable (looked at the right location 25% of the time) cue provider over a series of familiarisation trials where an animal animation would appear (target) following the cue. Following familiarisation, test and generalisation trials were introduced. In test trials, both faces looked to a location (one of four corners) it had previously cued, however no animal appeared and instead all locations briefly flashed white so as to encourage saccades. On generalisation trials, the same occurred, only faces looked to the location that had never been cued during familiarisation trials. Impressively, the infants had encoded the faces' cue reliability, as they spent most of their gaze time during test trials looking at the location cued by the reliable face, but looked at all locations equally when the unreliable face produced the cue. The effect for the reliable face was even more pronounced for generalisation trials. These results demonstrate that from a very young age, we are able to encode others' cues and use the information to guide their future behaviour. Like adults, infants respond to others' gaze-cue behaviour (Farroni et al.,

2000; Hood et al., 1998) and by around two to three years of age, children begin to use eye gaze alone to make mental inferences about others' mental states (Lee et al., 1998). Children begin to understand others' intentions and goals around four or five (Wellman et al., 2001), the same age that they learn to use gaze cues to understand basic intentions and desires (e.g., she is reaching for the candy, therefore she wants it) as they develop a mentalistic understanding of gaze (Doherty, 2006). Indeed, by four to five years of age, 'mind-reading' on the basis of eye-gaze emerges; deemed a critical component for theory of mind development (Baron-Cohen et al., 1997). Children also begin to use an actor's gaze to identify the location of hidden objects at that age, including in scenarios where the actor claims not to know or says a location is different to where their gaze is directed (Freire et al., 2004).

From seven to 15 years of age, knowledge about eye-gaze and deception develops further, as this skill seems to develop in parallel to a child's own ability to control their own gaze when deceiving (McCarthy & Lee, 2009). Given all this, adolescents are almost certainly able to detect gaze-cues and infer intended cooperation or deception delivered via individual gaze-cues. What is less clear is whether adolescents will track these cues over repeated experience and use such 'accumulated' cues to make inferences about the person behind the cues' trustworthiness, especially if the gaze-cues are peripheral to an attentionally demanding task.

Trust is important during interactions with new people as we want to know how they will act towards us (De Bruin & Van Lange, 1999; Zarolia et al., 2017). Trust is an especially interesting subject from the perspective of development, because it touches on so many other skill-sets, both social and otherwise, that are likely developing in parallel. As adolescents start to become socially independent, it becomes imperative that they can detect and understand social cues that reveal others' intensions and use those cues to understand the personalities and characters of others in order to learn *who* they can actually trust (see Mills, 2013, for an article on developing trust and learning to doubt others). Learning to trust others is an essential feature of adolescence (Derks et al., 2014), and while general trusting behaviour (as measured by the trust game) has been shown to increase and develop with age (Bos et al., 2010; Lee et al., 2016; van den Bos et al., 2010, 2011), and stabilises between 16-22 years of age (Sutter & Kocher, 2007), less is known about the

development of the very necessary ability to differentiate between who we can and cannot trust (Evans et al., 2013).

Knowing who can be trusted is particularly important for adolescents as they move away from parental and adult influences and begin to forge their own social groups. Social information becomes highly salient when making most decisions during adolescence (Crone & Dahl, 2012), even before trust behaviour and impression formation abilities are fully mature and by early adolescence, people are able to use social information to form trust-based impressions. Who you are interacting with becomes more important with age, as investments made in the trust game become about the other person, rather than just about allocating money/tokens (Güroğlu et al., 2014). For example, like adults (Chang et al., 2010; van 't Wout & Sanfey, 2008), ten-year-old children send more "money" to trustworthy looking faces than untrustworthy looking faces (Ewing et al., 2015) when engaging in trust-based economic games (a 'token' version of the trust game, designed for children). A similar study, conducted around the same time, looked at the same effect in a large group (n=540) of adolescents (13-18 years old) (De Neys et al., 2015). Here, the faces were of people who had used 'cooperative' or 'abusive' strategies when playing the game in a previous study, and whose faces were readily recognised by adult participants as either trustworthy or untrustworthy, respectively. While all age groups in the study sent less money to people who had been identified as untrustworthy in the prior study, the size of the effect increased substantially with age. This result strongly suggests that the ability to detect trustworthiness from faces and use that information to make decisions about *how* to interact with people develops across adolescence. Lee and colleagues (2016) gave statements detailing people as 'good' or 'bad' before 12-18 (n>800) year olds played iterative trust games where the people either conformed to or deviated from their descriptions. Older adolescent (14-15 and 16-18) were able to adapt and invest less with people described as good who were not behaving cooperatively, younger (12-13) children, however, continued to invest, suggesting they were unable to update their belief about the person. Together, these results suggest that the ability to detect trustworthiness from social information is present by adolescence, but continues to mature across adolescence. Developmental changes in impression formation skills may be particularly evident, therefore, when considering the ability to track and use repeated peripheral social cues to form impressions all while focused on another (non-social) task.

The current study

Taking the experimental design used in **Chapter 2**, **Experiment 3**, the current study explores trust learning from gaze-cues in an adolescent sample. In this design, we use both faces that are consistent (give either wholly valid or invalid cues) and inconsistent (change their behaviour over time). Using this design, the basis of which is the standard gaze-cueing task, impression formation from peripheral, non-diagnostic social cues can be explored. This task, which does not rely on accuracy scores, also side-steps one potential difficulty with assessing social-cognitive skills across development, where ceiling effects are common when trying to assess older vs. younger children (Thomas et al., 2007).

We use the gaze-cueing effect to check that gaze-cues have an attentional effect in adolescents (i.e., that invalid cues interfere with response times) and social measures to test whether repeated exposure to faces displaying various behaviours has an additional social effect. Perceived "niceness" is used to assess whether faces' gaze-cue behaviour influences a simple one-sided social judgement, and one-shot trust games are used to assess the influence of behaviour on risky, trust-based decisions during an interaction. As a 'half-way' between attentional and social measures, participants are also asked who they remember seeing the most often during the task. The purpose being that if impressions of faces are not formed, is anything about their individual presence encoded? In line with other studies that explored the development of trust during adolescence (Lee et al., 2016; van den Bos et al., 2010), we look at developmental change *within* the adolescent age-range by collecting data in two age groups, 11–13 and 14–15. Owing to the novel nature of the study in this age group, no specific predictions were made. Instead, the study aimed to answer the question; does the validity of faces' gaze-cue behaviour influence adolescents' social judgements and decisions?

Method

Participants

Pupils were divided into two age groups; younger and older. A total of 17 younger (age range; 11–13 years, mean 12.3, SD 0.69) and 25 older (age range; 14–15 years, mean 14.7, SD 0.5)

adolescents took part. Participants were rewarded based on the outcome of one of the one-shot trust games they played, where they could win a maximum of £3.00. The younger participants were also given small gift bags comprised of stationery items. Ethical permission was granted by the Ethics Committee at the School of Psychology, Bangor University.

Design

This experiment used the same design as Experiment 3 in Chapter 3 (see **Table 10**). The design includes wholly valid and invalid (consistent) conditions, allowing us to assess whether the basic trust learning effect is present in children, and conditions that change their validity over time (inconsistent), allowing us to see whether (should they be formed) impressions are updated following a change in behaviour. Finally, the design allows us to compare the patterns of responses in adolescents to that found in adults.

Materials & Stimuli

Face stimuli

Six white young-adult male faces, each displaying a mildly warm expression were used in the experiment to depict the various behavioural patterns (see **Figure 7**). Faces with slight, closed-mouth smiles were chosen to depict neutral emotional content as expressionless faces, which are often used as "neutral" faces, can be perceived as being hostile or angry, particularly by children (Mattavelli et al., 2014). The faces, together with adapted versions with averted gaze to the left or right, were kindly provided by Manssuer and colleagues (2015), who used them in a very similar

	Valid	**Decrease**	**Control ***	**Increase**	**Invalid**
	100-100	**100-0**	**50-50**	**0-100**	**0-0**
Initial (Blocks 1&2) % Valid	100	100	50	0	0
Final (Blocks 3&4) % Valid	100	0	50	100	0
Average Validity %	**100**	**50**	**50**	**50**	**0**

Table 10**:** *The five conditions in the study were based on gaze-cue validity both within each block and over time. *this condition was portrayed by two faces.*

gaze-cueing design and also collected baseline ratings of attractiveness and trustworthiness for each face. The six faces used here (among others) were matched on both attractiveness and trustworthiness. Faces were originally taken from the Nimstim face set (Tottenham et al., 2009). Which face was used to portray which behavioural pattern (e.g., giving 100% valid cues vs 100% negative cues) was counter-balanced across participants.

Gaze-cueing task

With the exception of the change to the faces used in the task (see above), the gaze-cueing task was identical to that used in the gaze-cueing experiments previously described in **Chapter 2, Experiment 3**. See **Figure 7** for an illustration of a trial with a face included in the task.

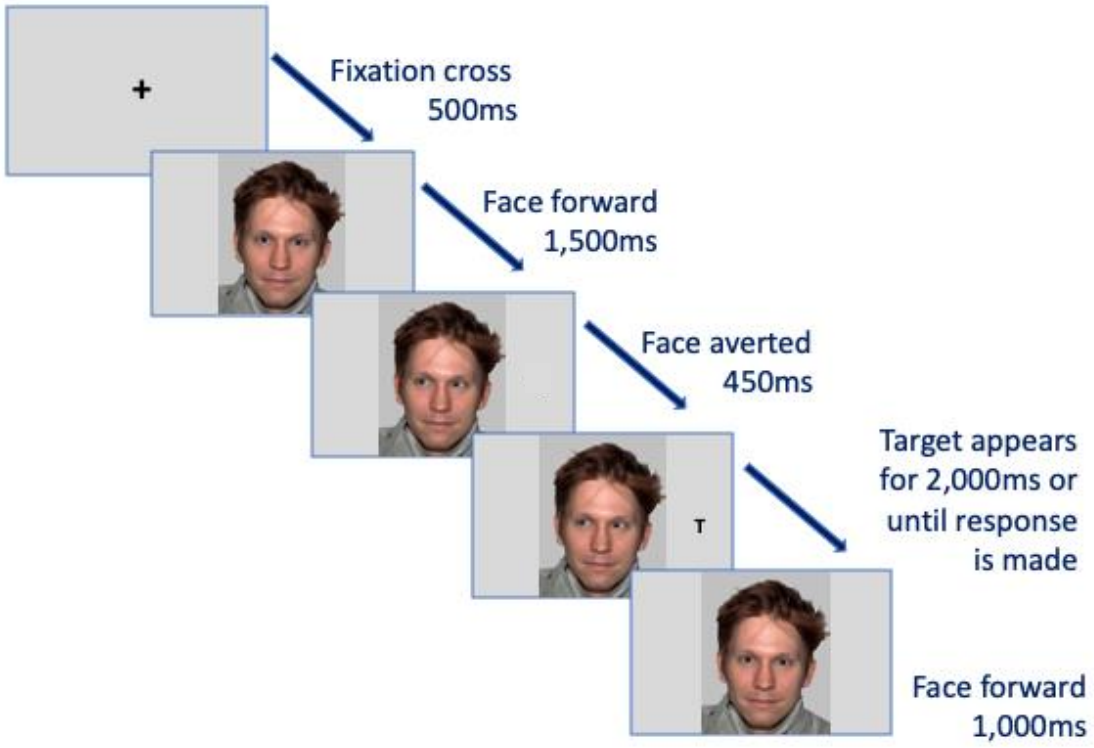


Figure 7: *An example of an invalid gaze-cueing trial, where the target appears in the non-gazed at location. The face depicted was included in the array of faces used in the experiment and has a 'calm' expression.*

Social measures

Face ratings

As a basic measure of social judgement, participants were asked to rate how nice they thought each face was after the end of the gaze-cuing task. Ratings used a 7-point scale, where 1 represented "not nice at all" and 7 represented "very nice".

Trust

As a measure of how participants would trust each face during an actual interaction, we used the trust (investment) game (Berg et al., 1995). In this case, the 'one-shot' variant was used (see Chapter 2, general Methods). Participants played a one-shot trust game with each face, after experiences their behaviour during the gaze-cueing task. Investments made were taken as an index of trusting behaviour.

Memory task

Participants were shown four faces from the gaze-cueing task (excluding the 'control' conditions) and asked to select the face which they thought had appeared most often. All faces had, in fact, been seen in equal numbers.

Questionnaires

Participants also completed questionnaires exploring self-esteem, mood, social value orientation and social anxiety. These primarily served the purpose of a filler task to occupy participants while others completed the gaze-cueing task. Although we were interested in exploring whether trust effects were mediated by other factors, the sample-size we were able to recruit was too small to meaningfully explore such individual differences. Thus, no analyses using data from these questionnaires were undertaken.

Procedure

After reading the information sheet and signing the assent form, participants were informed they were going to complete three tasks; a response-time-based computer task, a face rating task done

on the computer, and some questionnaires. These were described as *separate* individual tasks and participants were not informed that the face rating and trust-game tasks were related to the RT task. First, the trust game (referred to as the 'investment game') was explained, where participants were given an instruction sheet and the premise was re-iterated by the experimenter. Two pen and paper practice trials were then completed and pupils were encouraged to ask questions if they did not understand the rules and what they were required to do. At this point, they were informed that they would complete several of these one-shot games at the end of the testing session. They were also told that one of those games would be picked at random (based on the roll of a die) and that the trustee's decision about how much to return to them would determine how much, if anything, they had won. This was done to encourage participants to think carefully about their decisions in the trust games. They were told that any money they did not invest would be given to them and any money invested (and then multiplied by three) would either be kept by the trustee or shared with them, depending on what the randomly chosen trustee decided. In truth, this was determined based on the cue validity of the chosen face; the 100% valid face split the endowment 50/50, the 100% invalid face gave nothing back, and the other four faces – all on average across the game 50% valid – necessitated a further roll of the die to see if they were caught in a good (sharing) or bad (keeping) mood. Finally, participants were told that the amounts would be scaled down from the £10 capital they played with in the game to a real life amount of £2 capital, which could earn them a maximum of £3 (i.e., If the whole amount was invested – generating a £6 endowment – with the valid face, who would split the money equally).

Pupils were tested in smalls groups and the order in which the tasks were completed was counterbalanced within each group, with some pupils completing the questionnaires first while others completed the computer tasks before they swapped. The gaze-cueing task was described as a response time task, where responding both quickly and accurately to the target was emphasised and the faces were not mentioned. Prior to completing the actual task, participants practiced the gaze-cue task. Sixteen trials (using different faces to those used in the actual task) were included and participants were required to achieve 80% accuracy before they could progress to the main task. Only two participants, both from the older group, had to repeat the practice round. Immediately after the main gaze-cueing task was finished, each face presented during the task was shown individually and was rated on niceness ("How nice do you think this person is?").

Participants then engaged in one-shot trust games with each face ("How much would you like to invest with this person?"). To reduce anchoring effects, the order in which faces associated with each condition were presented was randomised. Finally, participants were shown an array of four gaze-cueing task faces (excluding 'control' faces) and asked "Who do you remember seeing most often during the computer task?". Again, the position that each condition appeared in was randomised.

Once all of the tasks were completed, the small group was gathered as a whole to see how much they had each won. Before being debriefed, participants were asked if they had guessed the purpose of the experiment. No one directly linked the gaze-cueing task with the social tasks in their responses, but a few of the younger pupils mentioned that some of the faces had been trying to "make them go wrong" and had distracted or annoyed them.

Data Analysis

There were no differences in mean responses for the two control faces for either the younger (niceness: t(15)=0.20, p=.844, trust: t(15)= 0.82, p=.424) or older (niceness: t(23)=0.892, p=.382, trust: t(23)=0.19, p=.853) groups. Data from the two control conditions for both niceness ratings and trust decisions were collapsed prior to analysis to form one 'neutral' condition, resulting in five overall conditions, as described in **Table 10** above.

The results of the gaze-cueing tasks were then analysed to ensure that the task generated the expected gaze-cueing effect and that overall error rates were low. Incorrect trials and those with response times of less than 200ms or greater than 1,500ms were removed prior to all analyses. Average accuracy and response time data was then calculated based on condition and block number. Although we intended to remove participants who scored less than 80% on overall accuracy as 'gaze-cueing' outliers, we only removed one participant in the younger age-group, who scored less than 60%. Two older participants scored 79% on the task, which we deemed close enough to 80% to be included, particularly as we were eager to retain as higher sample as we could, under the circumstances.

Power and Open science

The study was pre-registered (aspredicted.org/Y19_3GW), though only a target sample size was stated, as it was not known how many participants we could test and we wanted as much data as possible, as it was not known if younger adolescents would show a social effect. In the older group, a post-hoc sensitivity analysis shows a sample size of 25 provided enough power to interpret effects of Cohen's dz > 0.51 for a paired t-test to test for simple differences between wholly valid and invalid conditions (the observed effect was larger than this). Given effect sizes might be expected to increase with age (Neys et al., 2015), we would have needed around 50 younger adolescents to detect smaller effects of Cohen's dz = 0.35 (medium) or larger. With a sample size of 16 for the younger group, we actually had 80% power to detect effects of Cohen's dz > 0.65 (the observed effect was much smaller than this). When considering the groups separately, and including all five conditions, there was 80% power to interpret partial eta squares of 0.11 (medium-large) and 0.17 (large) for the older and younger groups, respectively, for a 1*5 repeated measures Anova.

Results

Gaze-cueing task

Accuracy

The gaze-cue task data was checked to confirm that both age groups demonstrated the expected effect of spatial attention. Although overall the younger group looked to have performed better on the task than the older group (see **Figure 8**, left), this apparent difference was not significant when simply average accuracy scores for the two groups were compared, [t(39)=1.30, p =0.200 (mean difference 2.1%, CI [-1.1, 5.2])]. The obvious statistical test to perform would have been a three-way Anova, with factors validity, block and group. Given the small sample size, the lack of any detectable difference between the two groups in overall accuracy and it does not form part of the research question, this was not undertaken. Instead, each age group's performance in the game was analysed separately using two-way (factors: validity and block), repeated measures Anovas.

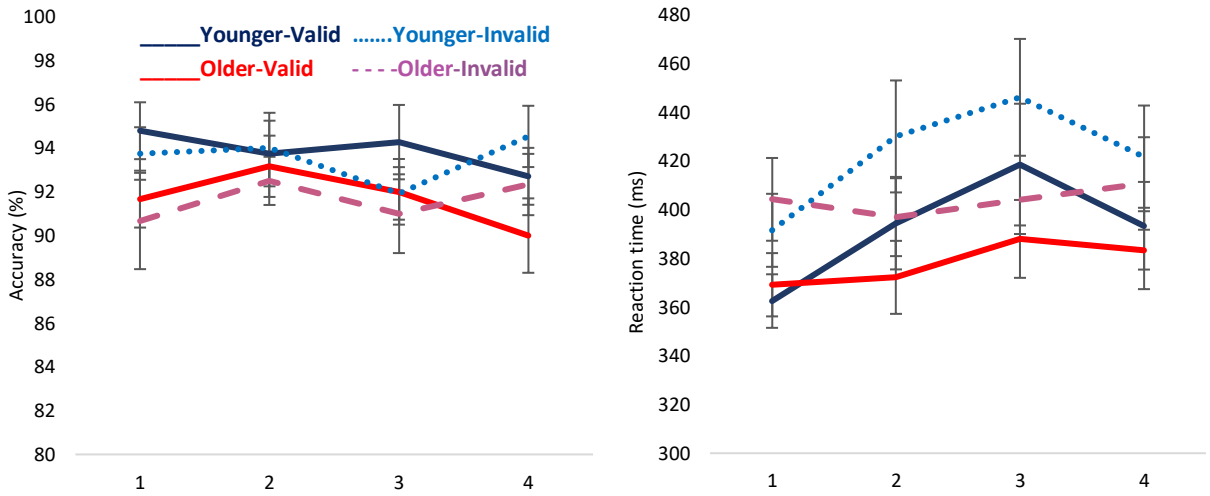


Figure 8*: Average accuracy scores (left) and response times (right) for each block split between valid and invalid trials for each age group. Older adolescents were a little less accurate but slightly faster. Bar represent SEM.*

For the younger group, there was no clear effect of validity, f(1,15)=0.21, p=.650, np^2=.01 (mean score valid: 93.9%, CI [91.5, 96.3], mean score invalid: 93.6%, CI [91.6, 95.5]). Though accuracy on invalid trials looks to have dipped in block 3 and valid in block 4 (Figure 2, left), there was no effect of block, f(3,45)=0.33, p=.804, np^2=.02, nor an interaction between block and validity, f(3,45)=1.43, p=.247, np^2=.09.

For the older group, accuracy seemed to increase between blocks 1 and 2 before decreasing in block 3, following the changes to faces' validity (**Figure 8**, left). However, there were no effects of validity, f(1,24)=0.01, p=.936, np^2<.01 (mean score valid: 91.7%, CI [89.6, 93.8], mean score invalid: 91.6%, CI [88.8, 94.4]), or block, f(2,51)=0.57, p=.0.635, np^2=.02, and no interaction, f(3,72)=1.35, p=.264, np^2=.05, present in the data. Small sample sizes likely reduced the power to detect any effects that may exist.

Response times

Gaze-cueing tasks produce a reliable cueing effect (difference between valid and invalid trials) which both groups were expected to demonstrate. The older group responded numerically faster than the younger group (see **Figure 8,** right) but a t-test comparing the groups' average reaction times, however, showed there was no real difference between the two, t(39)=0.72, p=.477 (mean difference -16ms, CI [-61, 29]).

As with accuracy, each age group's response times were then analysed using two-way (factors: validity and block), repeated measures Anovas to check for the presence of the gaze-cueing effect. For the younger group, there was a clear effect of validity, f(1,15)=21.09, p<.001, np^2=.58, with valid trials (mean 392ms, CI [356, 428]) being responded to more quickly than invalid trials (mean 422ms, CI [381, 464]). There was also an effect of block, f(2,27)=7.27, *p*=.0.004, np^2=.33. Pairwise comparisons showed that younger adolescents actually got slower over time before speeding up in the final block (see **Table 11**, left). There was no interaction between block and validity present, f(3,45)=0.17, *p*=.915, np^2=.01.

For the older group, the effects were a little different (see **Figure 8**, right). There was, again, a clear effect of validity, f(1,24)=25.89, p<.001, np^2=.52, with valid trials (mean 378ms, CI [350, 407]) being responded to more quickly than invalid trials (mean 404ms, CI [374, 434]). However, this was consistent over time, and there was no effect of block, f(2,38)=0.67, *p*=.0.574, np^2=.03 (see **Table 11**, right), and no interaction, f(3,72)=0.93, *p*=.432, np^2=.04.

Social Judgements & Decisions

Niceness ratings

Niceness ratings were initially analysed using a 5 (condition) by 2 (group) repeated measures Anova. This returned no main effect of condition, f(3,121)=1.35, *p*=.260, np^2=.03, no effect of group, f(1,39)=0.34, *p*=.564, np^2< .01, and no interaction between condition and group, f(4, 156)=1.12, *p*=.349, np^2=.03. Visual inspection of the data, however, suggests that there may be differences in how the two ages groups rated faces (**Figure 9**, left). While there does not appear to

	Younger				Older			
Blocks	**Mean diff. (ms)**	**95% CI Lower**	**95% CI Upper**	**Sig.**	**Mean diff. (ms)**	**95% CI Lower**	**95% CI Upper**	**Sig.**
1 and 2	-35	-61	-10	.011	2	-24	28	.866
2 and 3	-20	-37	-3	.024	-11	-22	-1	.036
3 and 4	25	-2	51	.064	-1	-17	15	.894

Table 11: *Paired contrasts for differences in reaction times across blocks for the younger (left) and older (right) age groups.*

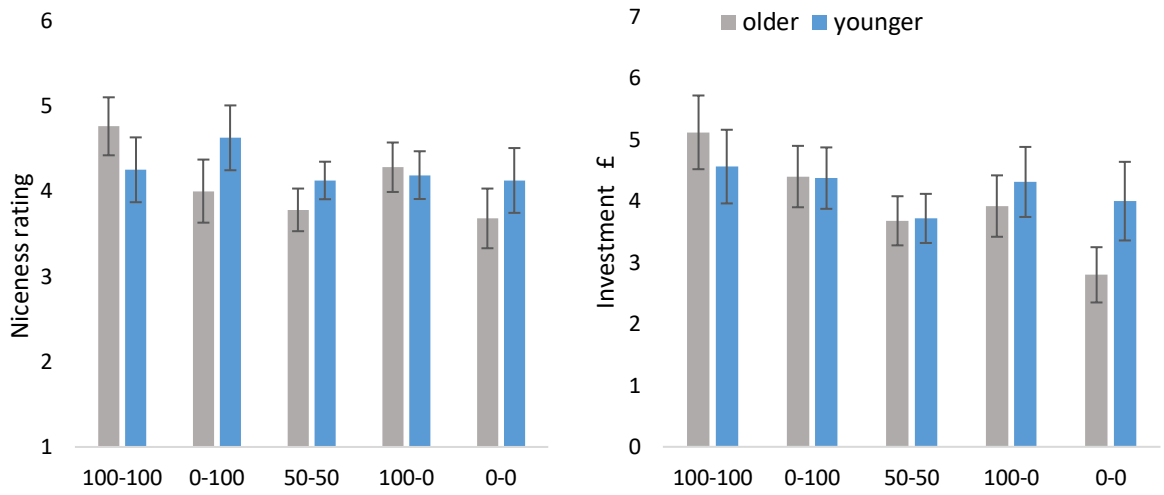


***Figure 9**: Average niceness ratings (left) and trust decisions (right) for each age group. Bars represent SEM.*

be much variation in ratings across condition for the younger group, the older group looks to have rated faces differently. To explore this further, the groups were analysed separately.

As expected, there was no effect present in the younger group, f(4,60)=0.51, p=.727, np^2=.03. While somewhat underpowered, a one-way, repeated measures Anova did find that the older group (marginally) distinguished between the conditions when it came to how nice each face was rated, f(3,68)=2.22, p=.097, np^2=.09. Unadjusted paired comparisons (see **Table 12**, left) showed that the wholly valid condition (100-100) was liked more than both wholly invalid (0-0) and increase to end the same (0-100) conditions. There was no difference, however, between how the invalid face and the face that decreased to end the same (100-0) were liked.

Trust

Again, a 5 (condition) by 2 (group) repeated measures Anova was used to explore differences between the groups. This time, there was a main effect of condition f(3,127)=3.23, p=.022, np^2=.08. As can be seen in **Figure 9** (right), differences were driven by the older age group, as the younger group invested similar amounts with each face. Similar average investments made meant

there was no main effect of group, f(1, 39)=0.18, p=.675, np^2<.01, and there was no interaction, f(4, 156)=1.06, p=.376, np^2=.03.

Similar to niceness ratings, the younger group did not appear to differentiate between the faces when making their investment decisions, f(4,60)=0.57, p=.695, np^2=.04. The older group, however, did, f(4,96)=4.32, p=.003, np^2=.15. Paired comparisons between key contrasts are reported in **Table 12**, right. Effects were mostly driven by the wholly invalid face (0-0), who received less than the wholly valid (100-100) face and the face that decreased how much it helped over time to end by not helping at all (100-0).

Comparison with adults

As this is the first study to explore trust learning from gaze-cues in children, it is helpful to compare its findings with an adult sample. This can be done by including the data from **Chapter 2**, **Experiment 3**, which used the same design. Given differences found between the two adolescent groups, it would not be useful to collapse them and compare overall results to the adult sample. Further, owing to small numbers in the adolescent sample, it is also not useful to perform inferential statistics that directly compare the groups. Instead, the results are compared visually and only the key (wholly valid vs. wholly invalid) contrast is compared.

	Niceness			**Investment**		
	Estimated difference	**95% CI**	**Sig.**	**Estimated difference**	**95% CI**	**Sig.**
100-100 and 0-0 (valid vs invalid)	1.08	[0.14,2.03]	.027	£2.32	[0.79,3.85]	.005
100-100 and 0-100 (both end valid)	0.76	[0.01,1.51]	.046	£0.72	[-0.77,2.21]	.327
0-100 and 100-0 (average same)	-0.28	[-1.16,0.60]	.518	0.48	[-0.83,1.80]	.457
100-0 and 0-0 (both end invalid)	0.60	[-0.25,1.45]	.159	1.12	[0.09,2.15]	.034

***Table 12**: Unadjusted pairwise comparisons of conditions for niceness ratings (left) and investment decisions (right) for the **older** group of adolescents.*

As can be seen in **Figure 10**, the older adolescents show a similar pattern of responses to the adults tested. Older adolescents actually appear to show a larger effect of validity between the wholly valid (left column) and wholly invalid (right column) conditions than the adults. When comparing only the wholly valid to wholly invalid faces, older adolescents (t(24)= 3.13, *p*=.005, dz= 0.62, mean difference £2.32) showed a slightly larger effect of gaze-cue validity on trust than adults (t(53)= 3.42, p< .001, dz= 0.47, mean difference £1.72) (see Ch.3, Exp.3). While both age groups invested similar amounts with the wholly valid face, the adolescents invested less with the invalid face. Indeed, unlike the adults, who invested similar amounts with non-predictive and invalid (either wholly or recently) faces, the older adolescents invested substantially less with the wholly invalid face as compared with any other condition. This is especially noteworthy given that wholly valid and wholly invalid faces were experienced among others who were performing more varied behaviours. This might suggest that adolescents are more sensitive to repeated negative social information than are adults, though this would certainly need to be replicated with a larger sample that included both adolescents and adults tested on the same stimuli.

Memory

Children were asked which of the (non-control) faces they thought they saw the most. This allowed us to explore whether, like adults (e.g., Bayliss & Tipper, 2006), adolescents also report having seen invalid faces the most often. Responses were analysed separately for each group using chi

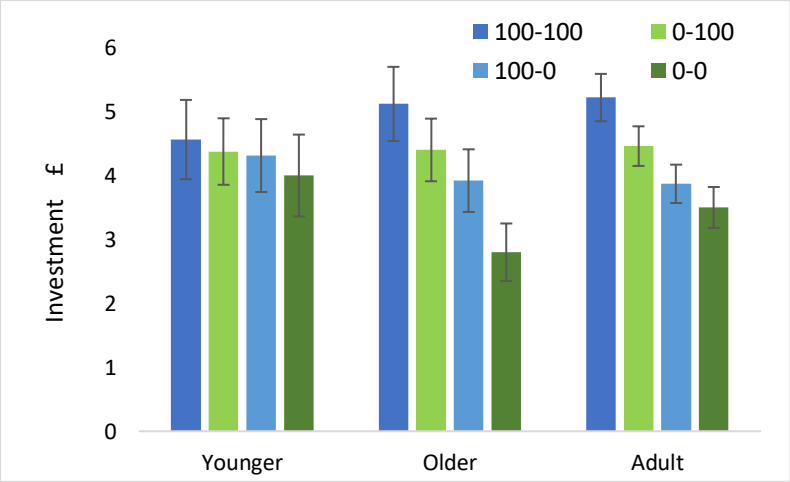


Figure 10*: A comparison of average adolescent and adult trust decisions for each condition (excluding control). Bars represent SEM.*

squared tests. Both the younger and older age groups seem *not* to have selected the wholly valid face as having appeared most frequently (see **Figure 11**). Rather, the older group appeared to choose the decrease condition the most, though, perhaps due to a lack of power, this did not reach statistical significance, $\chi^2(3)=4.28$, $p=.233$. The younger group looked to select either the decrease or the wholly invalid face (i.e., faces who ended by not helping), with only three participants selecting the wholly valid or increase condition. While this finding is visually clear, statistically, it was only tentatively supported, $\chi^2(3)=6.50$, $p=.090$.

Discussion

Implicitly extracting social information and meaning from the valence of faces' gaze-cue behaviour requires a complex level of inference on the part of the observer. Higher social skills are still in development during adolescence (Choudhury et al., 2006) and the current study tested whether trust learning from gaze-cues was present in children. Studies involving adults (18+) typically find that faces who always display deceitful (invalid) gaze-cues are trusted less than faces who always display helpful (valid) gaze-cues (Bayliss & Tipper, 2006; Rogers et al., 2014; Strachan & Tipper, 2017). The current study suggests that older adolescents (14-15 years) have developed the skill of linking the faces' gaze-cue behaviour with their intentions, such that those who always display invalid cues are not trusted. There was no such effect present for the younger group (11-13 years), who perceived all faces similarly, regardless of how the behaved. However, the analyses for this age group were underpowered, so it could be that effects do exist, they are just too small for use to have detected.

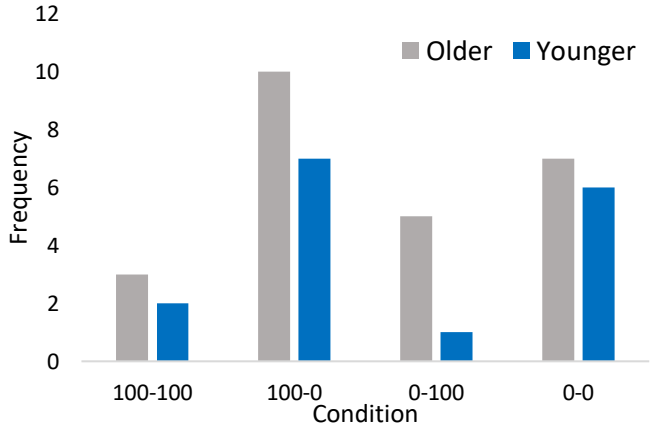


Figure 11: *Condition remembered as appearing the most during the gaze-cueing task, split by age group.*

The younger group did show the expected attentional effect in the gaze-cueing ask itself, confirming that the faces' behaviours were having an impact on their performance. Anecdotally, a few of them even mentioned the faces during the task saying things like "he's trying to make me go wrong". While, in the moment, these thoughts may have crossed their mind, they did not appear to be translated into beliefs about the actual person or any intention behind their behaviours, giving average niceness ratings for, and similar investments with (around £4), each face. This suggests that they did not perceive the faces negatively and that they were willing to a place reasonable amount of trust in each when investing with them. They did not, however, distinguish between the individual faces, indicating that while they may have developed general trust behaviours (i.e., as in van den Bos et al, 2010), they have not learned to relate the gaze-cues to the people producing them. Interestingly, though, there is tentative support for the finding that they did remember seeing the invalid (either wholly or recently) more often than the valid (either wholly or recently) faces. It is possible then, that the invalid faces' behaviours were encoded at some level, potentially due to their negative impact on performance; invalid faces became associated with worse performance, influencing memory *of* the faces but not *about* the faces. Age has been shown to positively correlate with performance on ToM tasks across adolescence (Kaland et al., 2007), and working memory and processing speed also increase during adolescence (Kail, 2003). Together this tentatively suggests that younger adolescents may not yet be able to process all of the information 'completely' to make any inferences about the person, but they are able to discriminate between faces. Indeed, participants were also attending to another (visually demanding) task, thus reducing attentional and working memory resources available to make the connection between behaviour and intentional meaning and, therefore, potentially contributing to a reduction in implicit impression formation.

It has been argued that it is difficult to study social cognition in older children due to ceiling effects; many social tests used with younger children are simply not hard enough for older adolescents and adults (Thomas et al., 2007). This is less than ideal, as we know that children undergo a great many socio-cognitive changes as they grow up. In this study, the social information was provided without reference and learned implicitly. This allowed for more covert, complex socio-cognitive abilities to be measured, as the child needed to connect the faces' behaviours to the context in order to infer their intention (over simply detecting a behaviour) and all without help or instruction. The

present results are, therefore, both interesting and important because we used a task that is capable of tapping into both social perception and social cognition without ceiling effects. Further, the design itself has a lot of flexibility, allowing for the influence of varied social contexts and social behaviours to be explored. Moreover, rather than adopting a more explicit approach such as 'form an impression of this person' or 'what are they thinking', this design allows for impressions to form naturally, which is not only more akin to real life, it also captures whether an impression *is* formed. By studying implicit social cognition (i.e., social processing happening outside of awareness) (Frith & Frith, 2008; Uleman et al., 2008) or incidental impression formation (i.e., impressions formed with no goal and/or reference to the person, we can begin to examine the subtle social changes that occur across development.

Limitations and Future directions

The obvious limitation here is sample size, and as such, this should be considered pilot data. That said, the results do suggest that this is an area worthy of further exploration. Many socio-cognitive abilities are 'online' by adolescence, but higher skills are still in development, and the present data supports the notion that trust learning from implicit social information is one such skill that has not been developed by early adolescence. Larger sample sizes would be able to confirm this. Further, it would be interesting to see if, at an individual differences level, the acquisition of this social ability correlates with other higher social skills (such as moral reasoning or understanding others' motives) or personal attributes (such as general or emotional intelligence).

As noted when discussing younger adolescents' findings, the main (non-social) task was itself attentionally demanding, and this may have contributed to the lack of social effects returned. It would be useful, therefore to a different task that is similar in regards to the fact the social targets are still neither referenced nor part of the main task, but different in terms of increased time and attention paid to the faces and their behaviours. This would allow for behaviours to be better encoded and not have them competing quite so much with another visual attention task, which may allow for any relationship between the faces and their behaviours to be inferred or perceived. A task such as that used in Chapter 4 would be highly suitable.

Conclusion

Adolescence represents an important part of social development, yet receives relatively little attention in the literature, perhaps owing to a difficulty in findings tasks that can capture developmental differences. This study investigated whether, like adults, adolescents of different ages socially process uncontextualised behavioural cues to form impressions of the people making the cues. Preliminary findings suggest that older adolescents process the behaviours similarly to adults, whereas younger adolescents may not yet make the connection between the behaviours and the people behind them. The flexible gaze-cueing paradigm may, therefore, offer an additional way to study the development of implicit social cognition.

Chapter 4: Incongruent Cues & Outcomes

Overview

In many impression formation studies involving a direct interaction, the target's behaviour largely determines the perceiver's outcome. For instance, in traditional gaze-cueing paradigms, a negative gaze-cue has a concomitant negative impact on performance. This congruity makes it difficult to distinguish between the relative importance of others' behaviours and our outcome when forming impressions. On the one hand, if the intention behind the other's gaze behaviour is perceived to be negative (i.e., they are deliberately trying to deceive us), it makes sense for us to not trust the person. Alternatively, our impression could be more about the association formed between a face and our outcome; when we see that person, we do not do very well. If the other's social behaviour (i.e., deceitful) is incongruent with the perceiver's outcome (i.e., positive), the straightforward nature of the relationship between cue and outcome can be separated, allowing for differences between the two to be studied. This chapter uses a new gaze based task to explore whether the social valence of a person's behaviour or its economic utility predicts impressions.

The second experiment in this Chapter was motivated by findings in Chapter 2, which demonstrated that people are capable of tracking and encoding others' inconsistent gaze-cue behaviour, however, the results were complex. The gaze-cueing method used did provide for a rather extreme impression formation experience, where the means of learning (i.e., implicit and from task irrelevant behaviour) and the behaviours experienced (i.e., fleeting) likely contributed to the results observed. This raises the question of whether impressions follow a similar pattern and magnitude when others' individual behaviours, while remaining unreferenced, are more salient and identifiable. Further, when learning during traditional gaze-cueing paradigms, the perceiver cannot benefit from detecting the other's cue validity, as invalid cues interfere with performance regardless of (implicit) knowledge, meaning a perceiver cannot use their knowledge to change their experience or outcome.

Here, I again used unreferenced gaze-cue behaviour to provide others' social behaviour while perceivers completed another task and did not expect to encounter the faces again. In this way, impressions formed remain incidental, however, the faces' behaviours were made more salient and explicit learning was made possible. By removing the visual attention (i.e., reaction time) component of the task and replacing it with a (forced choice alternative) decision-making task, a perceiver's experience with the other is altered in three important ways. First, perceivers 'see' the other's behaviour more clearly, as it does not simply flash before them but rather remains until the perceiver has made a self-paced decision, meaning they can more readily process and encode each individual's behaviour in relation to a trial's outcome. Second, and relatedly, because the perceiver is tasked with making a decision (as opposed to detecting a target) their outcome is not determined by the other's gaze behaviour. Another's gaze simply looks one of two options and the perceiver is free to make their own choice. This means while the faces' gaze behaviour may affect perceivers' visual attention, it does not concomitantly drive their outcome. Third, and significantly, perceivers can explicitly (i.e., with awareness) learn the faces' cue-outcome contingency and use the knowledge gained to their benefit. This is due to both veridical (faces who look at the right answer) and non-veridical (faces who look at the wrong answer) information having a predictive utility; for a non-veridical cue provider, the right answer is the simply the option they do *not* look at. This has the effect of breaking the congruence between the face's social behaviour (for instance negative) and the perceiver's outcome (which here would be positive), meaning the influence of each can be explored.

Research Questions

1. Are incidental impressions and social decisions driven by the inferred social meaning of the other's behaviour or the perceiver's own outcome?

2. Are incidental impressions formed when others' gaze-cue behaviour is more salient and capable of being learned explicitly, similar to those formed when behaviour is experienced using a more implicit learning (i.e., the traditional gaze-cueing) task?

Contribution to Research

This is the first impression formation (and updating) study to separate the concomitant relationship between the other's behaviour and a perceiver's outcome (i.e., if they are negative, we do badly), testing whether impressions and decisions are guided by the other person's behaviour or our financial outcome.

Results of the two experiments detailed in this chapter have been prepared for publication under the title:

"*Your behaviour may help me, but I don't like you; Inferred intent, not economic utility predicts social decisions*" *(Newey, Rauwolf, Ramsey & Koldewyn, in prep.)*

Raw data for each Experiment are available from the project's Open Science Framework page: https://osf.io/6erfq/

Contribution of Author

I conceptualised, designed, collected data for, analysed and wrote up the experiments detailed in this Chapter. The programming of the experiments and the majority of mixed effect models' analyses were done by Dr. Paul Rauwolf, to whom I am incredibly grateful. Experiments were written and run in Qualtrics.

Abstract

When we interact with new people, we use our observations of their social behaviours and cues to infer their intentions, guide our own behaviour, and form impressions of their character and personality. Often, the valence of someone's social behaviour is both congruent with the valence of our outcome and consistent over time – nice people are usually nice and almost always try to help us to reach our own goals. This congruence between 'social valence' and our own outcome facilitates impression formation and simplifies social decision-making processes. Impression formation and social decision making may, however, become more complex if the perceived social intent behind a person's behaviour is incongruent with its economic consequences, for example when another's negative social behaviours have positive outcomes. This study had two main aims; first, to explore how people are perceived when their negative social cues come to be associated with positive financial outcomes (Exp.1) and, second, to examine how social decisions are affected when the valence of others' behaviour is inconsistent over time, when valence and outcome are not necessarily aligned (Exp.2). Participants played an incentivised, forced choice, card guessing game during which unreferenced faces provided either veridical (looked at the right answer) or non-veridical (looked at the wrong answer) gaze-cues, both of which facilitated high accuracy if detected and learned. Importantly, all faces' cues predicted the correct answer and participants quickly learned to use both cue types equally. Social judgements and decisions about veridical and non-veridical faces diverged, however, following the inferred social meaning of a target's behaviour, not its overt utility. Despite facilitating and being associated with winning, non-veridical faces were disliked, distrusted, punished, and even avoided in future games where their predictive behaviour would be economically beneficial (Exp.1). This pattern of impressions persisted even when some of the other faces behaved inconsistently and disrupted performance (Exp.2). Together, results demonstrate that when it comes to social decision making, the perceived intention behind other's behaviour is more important to us than its impact on our outcome.

Introduction

We are quick to form impressions of people based on their behaviour, especially if we are affected by their actions. In many circumstances, the social valence of others' behaviour is correlated with our own outcome: if they are helpful, we benefit from their action but if they behave selfishly or deceitfully towards us, we usually experience a personal loss or poor outcome. When someone's deceitful behaviour negatively affects us, it makes sense for us to form a negative impression of them and to not trust them in future interactions. In such circumstances, the social implications are congruent with the perceiver's own outcome, making it difficult to identify whether an impression is driven by the social valence of the other's behaviour, by its consequence, or both. Further, the straightforward, congruent nature of the relationship between another's behaviour (e.g., deceives) and a perceiver's outcome (e.g., performs poorly) presents no conflict for perceivers when forming impressions or making decisions about possible future interactions. How we might perceive and interact with someone whose negative social behaviour has positive consequences, however, is a more complex and interesting question. How do people make decisions when another's behaviour has negative social connotations but is economically beneficial? In such situations, when social valence and outcome conflict, do we judge the person based on social behaviour or economic utility? Moreover, what happens if the social valence of someone's behaviour changes over time in such conflicting circumstances? Here, we probe these questions across two experiments by separating the valence of a person's behaviour from its economic utility and measuring both social judgements and social decisions informed by others' consistent (Experiment 1) and inconsistent (Experiment 2) social behaviour.

We process social information to form impressions rapidly and effortlessly and we do so from even minimal social information (for reviews see Ames et al., 2011; Olivola et al., 2014; Uleman et al., 2008; Uleman & Kressel, 2013). The trait that we are both fastest and most likely to evaluate is trustworthiness, which speaks to a person's trait warmth (Cuddy et al., 2008; Pakrashi et al., 2009; Todorov, 2008; Willis & Todorov, 2006). Perceptions of trustworthiness correlate with other social characteristics, such as likeability (Todorov, 2008), and whether we think someone is good or bad helps us to decide how they may treat us (Brambilla et al., 2021; Cuddy et al., 2008; Frith & Frith, 2006; Frith & Singer, 2008; Lammers et al., 2018). Simply seeing a person's

(un)trustworthy looking face (Chang et al., 2010; Rezlescu et al., 2012; van 't Wout & Sanfey, 2008) or reading about their past (im)moral or (anti)social deeds (De Bruin & Van Lange, 1999; Delgado et al., 2005; Maurer et al., 2018; Zarolia et al., 2017) is enough to influence how much we initially trust them in a risky economic exchange, such as the trust game (Berg et al., 1995). The ability to form quick impressions from even sparse social information is functional, as it not only helps us to predict the actions and intentions of others in real time, it also allows us to plan our own actions in the moment to increase our own chances of success (Ames et al., 2011; De Bruin & Van Lange, 1999; Fiske, 1992).

While we do form initial impressions quickly and from sparse information, we also use social behaviours to learn about people across time, updating initial impressions as we gain relevant new information (Mende-Siedlecki & Todorov, 2016). We often learn about people from experience (Behrens et al., 2008) and during a direct first person encounter, we can gain knowledge about a person's actual trustworthiness or goodness by consciously evaluating their behaviour and our outcome in context so as to infer their intentions (Frith & Frith, 2006; Malle, 2011; Malle & Holbrook, 2012). Trust can be learned over time through a series of direct positive exchanges (Balliet et al., 2011; Behrens et al., 2008; Chang et al., 2010), but we also monitor and aggregate more subtle social information over time to form incidental impressions (Rogers et al., 2014). Indeed, social information that doesn't provide compelling evidence about trustworthiness or social intention when observed only once, can provide powerful social information when that behaviour is consistently repeated. One example of this is gaze behaviour. A shift in another person's gaze is a compelling attentional cue which automatically orients *our* own visual attention to where *they* are looking (Emery, 2000; Frischen et al., 2007; Langton & Bruce, 1999). Researchers have used gaze-cueing paradigms, where the gaze shift either facilitates or interferes with performing an unrelated target detection task, as a way to provide flexible social behaviour. In such paradigms, veridical faces' cues facilitate response times by orienting us towards the correct target location, while non-veridical faces' cues interfere with response times by orienting us towards the incorrect target location (Driver et al., 1999; Friesen & Kingstone, 1998). Faces that repeatedly interfere with our responses are consistently rated as less trustworthy than those whose gaze behaviour improves our performance (Bayliss & Tipper, 2006; Strachan et al., 2016, 2020; Strachan & Tipper, 2017). When faces change how they behave over time (e.g., switch from

veridical information most of the time to giving non-veridical information most of the time), impressions and trust decisions tend to be driven by the most recent behaviour if that behaviour is negatively valenced, while consistently positive behaviour is valued over only recently positive behaviour (**Chapter 2**). Additionally, in many social situations, we use people's gaze behaviour to infer their mental states (Argyle & Cook, 1976; Doherty, 2006; Hamilton, 2016; Shepherd, 2010) and there is evidence to suggest that gaze-cueing based impressions are driven by the 'perception of minds, not eyes' (Colombatto et al., 2020). For example, the gaze-cueing studies here suggest that participants perceive faces that display non-veridical cues as intentionally trying to hinder us, which leads to distrust. This is noteworthy, because any intention on the part of the person providing the gaze cues must be entirely inferred, as their behaviours are neither explained nor referenced. Altogether, these results suggest we track people's social behaviour across time and use that (sometimes implicit) information to form impressions about them that we then use to infer their mental states and their likely future intentions.

Indeed, the human tendency to attribute intention to actions is ubiquitous. As we interact with and observe people, we do not simply perceive their actions, we tend to think about *why* they chose particular behaviours and what they *intended*, as we consider behaviours to be the product of goals (Frith & Singer, 2008; Malle, 1999, 2011; Malle & Holbrook, 2012; Reeder, 2009). When evaluating someone's behaviour or the outcome of an event, knowledge of the person's intentions influences moral judgements. For instance, if a negative action with terrible consequences is performed unintentionally (e.g., poisoning someone's coffee with arsenic that is thought to be sugar), then the actor is judged much less harshly than when the same action is performed intentionally (Young et al., 2007). Indeed, even if there is no action taken, so long as a failure to occur is for external reasons, the person is still regarded as immoral (Hirozawa et al., 2020), demonstrating that a person's motives can have a powerful effect on how we process social information.

A person's intentions towards us are often clearly conveyed when we engage in direct interactions where our actions can directly affect each other. Economic games are a good example of such interactions and are often used to study social learning and decision making (Balliet et al., 2011; Camerer, 2003; Sanfey, 2007). One such game, the trust game (Berg et al., 1995), is often used to

study how we learn about others' trustworthiness over time. In this version of an impression formation task, we have a financial goal and another person's economic utility to us should be of primary importance. However, though the rules of the game may be simple, engaging in the game requires complex reasoning about one's partner's (or opponent's) mind, as their behaviour determines our outcome (Rilling & Sanfey, 2010; Sanfey, 2007). In the trust game, one player, the 'trustor' is given a monetary windfall (e.g. £10) which they can choose to invest with the other player, the 'trustee'. Any investment made is multiplied by some number (typically three) and the trustee can decide whether to share (act fairly) or steal (act selfishly) the money. The amount a trustor invests is taken as a measure of how much they trust the other player, as making an investment involves taking a risk on the trustee (Rousseau et al., 1998). Deciding, then, how much (if anything) to invest requires the perceiver to infer the trustee's intention in order to determine whether they can be trusted to share, or whether they might keep all the money. When people engage in iterative trust games with novel people, both investment decisions and trustworthiness judgements are predicted by the other's reciprocity, with a person who shares being more positively perceived than one who is greedy (King-Casas et al., 2005; Phan et al., 2010). If a person continually steals from us, we can reason that their intention is not to cooperate with us, but to win the most for themselves, and this knowledge tells us we cannot trust the person.

Direct interactions comprise both the other person's behaviour and our own outcome. In many impression formation studies, the emphasis is placed on we how process and evaluate the other person's behaviour. For instance, in the trust game, we are concerned with the other person's overt sharing rate, while in a gaze-cueing paradigm, it is the inferred valence of their social cue. While sharing behaviour and gaze-cue behaviour do represent rather different ends of the spectrum of tasks that have been used to investigate impression formation, providing behaviours in very different contexts, they do have one thing in common as regards the perceiver's outcome. In both situations, the social valence of the other person's behaviour (i.e., steal or deceive) is congruent with the valence of our own outcome (i.e., lose money or perform badly). This conflates impressions based on the other's behaviour with impressions based on our outcome. From the prior literature, we know that together both outcome and social valence are powerful factors in driving impressions. Less is known about how impressions are formed when social valence and outcome are incongruent, and if in such situations our judgements of others would be driven by social

factors or by our own outcome. How do we perceive people when their behaviour is socially negative, but has positive utility for our own outcome? How are initial impressions in such situations formed, and how do they update when individuals change their behaviour? To investigate these questions, we used a card-game designed to partially dissociate the social valence and economic utility of others' social behaviour and measured both impressions and social decisions across two experiments.

Current Study

Interacting with and making sense of other people successfully requires that we sometimes reconcile complex and incompatible information (Freeman & Ambady, 2011; Hughes et al., 2017). The current study considered how we do this when the relationship between the negative valence of a person's social behaviour is incongruent with its economic utility, both when behaviour is consistent (Experiment 1) and inconsistent (Experiment 2) over time. To our knowledge, this approach has not been taken before. We chose to use gaze-cues to deliver social information because they can be used to communicate social information or intentions (Emery, 2000; Frischen et al., 2007; Frith & Frith, 2008; Gobel et al., 2015; Itier & Batty, 2009; Shepherd, 2010) and, while they have a powerful attentional effect in an observer, in our design they do not automatically impact performance as they are used to provide suggestions rather than determine outcomes.

Participants played a forced choice, incentivised game where they had to guess whether the value of an unseen card was higher or lower than a visible card. During the game, participants received gaze-cues from faces that all provided the correct answer, but varied in how they did so. Some faces looked at the correct answer (veridical cues), suggesting positively valanced or cooperative behaviour, while others looked at the incorrect answer (non-veridical cues), suggesting negatively valenced or deceptive behaviour. Importantly, once participants learned each face's cue-outcome contingency, *all* the faces provided information that revealed the correct answer and increased rewards. A player's performance was not, therefore, negatively affected by negative social information. Rather, all cues had equal utility regardless of their social intention, meaning all cue types were associated with winning. Subsequent decisions could, therefore, be driven by inferred

social valence (i.e., about the other's perceived social intention), or the overt economic utility of the cues (i.e., about our own outcome).

If someone always looks towards the incorrect answer (non-veridical cue) they must, by definition, know the correct answer. While it is possible that they have no social goal, players that notice and attend to the faces are likely to deduce that such faces intend to deceive them. It seems likely, then, that a social valence will be attributed to the gaze-cue behaviours and corresponding impressions will be influenced by how the person is perceived to have treated us. In other words, we expect that faces providing non-veridical information will be disliked and distrusted. On the other hand, reasoning about the faces' intentions is not necessary to succeed at the task, which can be approached using a 'cue-outcome' rule and it is therefore possible that people will process behaviours not based on mentalising about the person's intentions, but rather by associating cues with their own outcome which, once the rule is learned, is always positive. Associative or reward (Balliet et al., 2011; Behrens et al., 2008; Harris & Fiske, 2010; V. K. Lee & Harris, 2013) learning models might, therefore, predict that faces will be perceived equally, owing to the fact that each comes to be associated with winning and should have similar reward value (Cox et al., 2005; Sims et al., 2012).

It is also plausible that social valence and economic utility will be weighed differently depending on the *type* of social decision that is being made. For this reason, we chose to use several different measures to capture the influence of socially negative but economically positive behaviour on a range of judgements and decisions. We chose likability as a purely social measure with no implications for either party, expecting this measure to be most influenced by social valence. To explore how behaviour in the game translated to trust learning, one-shot trust games (Berg et al., 1995) were used. In the trust game there are two social elements, as low investments can signal both a lack of trust and a desire to prevent the 'trustee' from getting any money. Similarly, we used one-shot ultimatum games (Güth et al., 1982) to explore whether participants would actually pay an economic cost themselves in order to 'punish' faces that they perceived to be socially 'bad' (Henrich et al., 2006). While it was expected that people would dislike and distrust faces who provided non-veridical cues, we did not expect that this would extend to forsaking a fair monetary offer and hurting one's own outcome simply to 'punish' the other. Finally, forced choice measures

were used to see whether players would choose to interact with faces again for both more rounds of the trust game and more rounds of the card game. Trust decisions were expected to be similar to those found for the actual one-shot games (i.e., if participants did not trust the face, they would choose *not* to play more trust games with them). However, we expected that participants would choose to play more rounds of the card game with familiar faces, regardless of whether they had given veridical or non-veridical cues. In other words, we expected that only the utility of the cue would matter in making this decision, as refusing to interact with a non-veridical cue provider (no matter what you think of them socially) would only hurt oneself. It was here, then, that we most expected socially motivated decisions to deviate from economically motived decisions.

Experiment 1

In this first Experiment, two faces always provided veridical gaze-cues and two faces always provided non-veridical gaze-cues during the card game, each looking to the right and wrong answers, respectively. Upon completion of the game, participants were asked a series of socially oriented questions regarding the four card game faces, and four unfamiliar faces, designed to probe how people judged each of them independently and as compared to one another. In addition to investigating how negative social behaviour impacts impressions when such behaviour results in positive economic outcomes, this first experiment had two further aims. Firstly, we explored how learning which faces were veridical and which non-veridical during the card game task might relate to the strength of impressions formed. Secondly, this first experiment served both to ensure that (most) people notice the faces and learn their behaviours and what initial impression are formed prior to any change in behaviour taking place, as occurred in Experiment 2.

Method

Open Science & Power Analysis

This experiment was pre-registered (aspredicted.org?FT3_21Y) and an a priori power analysis using G*Power 3 (Faul et al., 2007) was performed. All data is freely available online on the project's OSF page: https://osf.io/6erfq/. All manipulations and measures are reported, and all exclusions are specified. The main effect of interest was the potential difference in trust ratings

between game faces that displayed different social behaviour and so a paired samples t-test was used for the power calculation. With a pre-determined sample of 50 participants, the experiment had over 90% power to detect Cohen's dz of 0.50 or higher. In consideration of potential outliers, a minimum of 55 participants were targeted.

Participants

Fifty-eight (female 41; mean age 20.1 years, SD 3.4) students at Bangor University, UK completed the experiment. Each participant was compensated for their time with course credits and was remunerated for their performance in the card game, where a maximum of £2.88 could be won. The experiment was granted ethical approval by the School of Psychology at Bangor University's Ethics Committee, and all participants provided written consent prior to data collection taking place.

Design & Stimuli

Face Stimuli

Eight, young, white adult male faces, each displaying a neutral expression, were used in the experiment. Faces were taken from the Oslo Face dataset (Chelnokova et al., 2014) which includes natural images for both eyes forward and eyes averted (left and right). The dataset also provides ratings (on a scale of 0-10) for individual faces based on responses from ~40 people (50% female) along three dimensions: attractiveness, trustworthiness and dominance. The faces chosen for this study were rated similarly to each other across all three dimensions, and to the average of the (male) group as a whole (see **Appendix C1: Selection of faces**, p.177).

Card Game Task

Participants played a simple forced choice game, for real money, based on a card game of 'higher or lower'. Details of the task's parameters and design can be found in **Appendix C2: Card Game parameters** (p.179).

During each trial, players had to guess whether a face down (unknown) playing card was higher or lower than a face up (visible) playing card. Prior to playing, participants were informed that the value of the cards were limited, such that the face-up card $\in \{4,5,6,7\}$, while the face down card $\in \{2,3,4 \ldots 9\}$. Further, they were told that the face down card would never be identical to the face up card. The task can be approached by using a probabilistic strategy. For example, if the visible face up card is a 7, then there is a 71% chance that the face down card is lower (2-6) rather than higher (8-9). If the participant uses this strategy, on average they would guess correctly 62.5% of the time in the current task – comfortably above chance, but not nearly as well as they *could* do if they attended to and learned the contingency of the gaze cues. Once players made their choice, the face down (unknown) card was revealed and they learned whether they were correct. To motivate participants to attend to the task and to do well, each correct trial resulted in a small monetary gain of £0.03, while incorrect trials resulted in a loss of £0.02.

In addition to the explicit card information, during each trial, an unreferenced human face looked towards either the higher or lower option (see **Figure 12**). A total of four faces were shown during the game, two providing veridical cues (by looking to the correct answer) and two providing non-veridical cues (by looking to the incorrect answer). Crucially, those players who detected and tracked the behaviour of the faces could quickly learn which faces were veridical and which nonveridical. By learning the behaviours of any given face, players could then increase their accuracy, and reward, for the remaining trials. This was true even for faces providing non-veridical cues, as the participant could simply select the option *opposite* to the one they cued. Importantly, not all trials allow for equal learning opportunities. Because some answers are more probable than others, a clear indication of when participants had learned the gaze-cue contingencies was when they started to make 'irrational' decisions by following the gaze-cues to pick the less probable, but correct, answer. In this way, we were able to capture both when people begin to trust others' cues in spite of conflicting probability information, and whether learning occurred at different rates for veridical and non-veridical cue providers.

Each trial began with a face positioned in the centre of the screen whose eyes were looking straight forward, a face-down (unknown) playing card positioned directly beneath the face, text boxes saying "Lower (F)" and "Higher (J)" positioned to the left and right of the face, respectively and

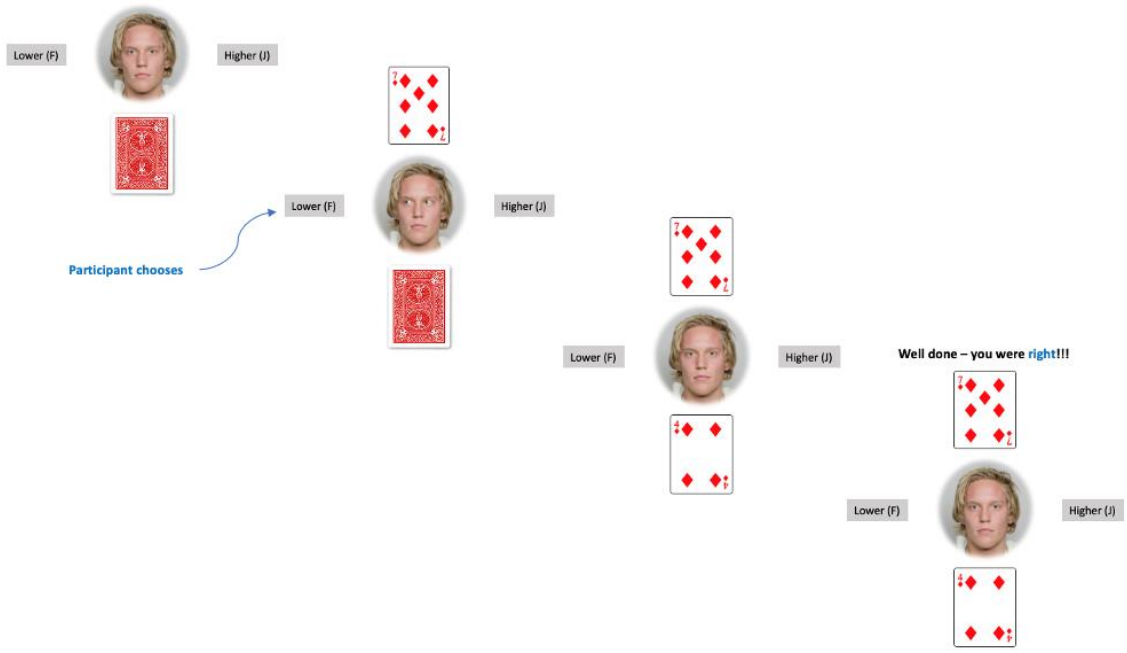


***Figure 12**: An example of a trial with a 'veridical' face where the person looks towards the correct answer, and in this instance, also the most probable answer. The participant selects the correct answer, winning some money as a result. The face depicted here was not used in the experiment.*

text reading "Do you think the face down card is LOWER or HIGHER than the card below?" displayed at the top of the screen. The image remained for 1,000ms before the face looked either to the left at 'Lower' or to the right at 'Higher'. At the same time, a face-up (visible) playing card appeared directly above the face. This image remained on screen until participants made their choice. Participants were instructed to press the "f" key if they thought the face down card was lower than the face up card and "j" if they thought it was higher. Once they made their selection, the face returned to the direct gaze position for 1,000ms before the face down card was revealed and the message "Well done – you were right!!!" or "Unlucky – you were wrong" was displayed at the top of the screen for 1,000ms (see **Figure 12**). The number of trials where the answer was 'lower' was equal to the number of trials where the outcome was 'higher'.

The task was structured into three blocks, with each block including 32 trials, for a total of 96 trials. This resulted in each of the four faces being presented eight times per block and 24 times overall. Blocks were pre-constructed so that trials were the same for every participant. This ensured that the position of veridical and non-veridical, as well as probable and improbable trials, did not

interact with individual participant's experience or learning. Importantly, which faces were veridical vs. non-veridical was counter-balanced across participants, ensuring that any effects of the characteristics (e.g., attractiveness, small differences in facial expression) of particular faces were minimised, as the same face that provided veridical cues for one participant would provide non-veridical information for another. Each individual face always looked at either the correct or incorrect answer. However, the answer itself was not always the most likely outcome, given the value of the face-up card (4-7). When a face-down card (2-9) is drawn at random, it has a 36% chance of being one of the least probable options available (e.g., an 8 or 9 when the face-up card is a 7). In order for the game to approximately reproduce outcomes conforming to this probability, while also ensuring that conditions were each associated with the same number of probable and improbable outcomes, 12 (37.5%) of the 32 trials per block had an improbable outcome (see **Appendix C2: Card Game parameters**, p.177). This resulted in each face being paired with five probable and three improbable trials per block.

Social Measures

Social Judgements

Participants were asked to rate eight faces - four from the card game and four unfamiliar - for niceness on a 1-7 point scale (with end points labelled 'not nice at all' and 'very nice', respectively). The order in which the faces were shown varied across the different versions of the task, however, all participants were first shown an unfamiliar face (see **Appendix C3: Social Measures**, p.179).

Social Decisions

Three different measures were used to explore how the valence of targets' gaze-cue behaviour influenced social decision making. First was the trust game (Berg, et al, 1995), referred to as an 'investment game', where participants, acting as the investor, could invest £0-10 with each trustee, in this instance each of the eight faces, only once. Though no real money was exchanged, participants were informed that any money invested would be tripled in the hands of the trustee, who would decide if and how much of the tripled amount they would share with the participant. Again, the order in which faces were shown varied across versions but always started with an

unfamiliar face. Second, participants were asked a series of forced choice questions designed to probe which faces participants would choose to interact with in the future. Across two different question sets, participants were presented with a choice of two faces, displayed side-by-side. In the first series of questions, participants were asked which person they would rather play with if they were asked to play more rounds of the card game. In the second series, they were asked which person they would like to play with if they were asked to play multiple rounds of the investment game. There were eight pair combinations for each series, with veridical and non-veridical faces being paired with both unfamiliar faces and each other. When choosing partners for more rounds of the card game, the rational choice is to never select the unfamiliar faces because both veridical and non-veridical faces are consistently predictive, whereas nothing is known about how the unfamiliar faces might play the game. In contrast, if participants perceive the veridical face as helpful and the non-veridical face as deceitful, the rational choice would be to never select the non-veridical faces and instead choose to play with the unfamiliar face, whose trustworthiness is unknown. For both games, the veridical face should always be selected.

The final task involved a number of rounds of the Ultimatum Game (Güth et al., 1982), where participants played the role of the responder. In this economic game, one party, the 'proposer' (here played by the faces) has a sum of money (here, £10) and must decide how much they will share with the other party, the 'responder'. The responder decides whether they will accept or reject the offer. Crucially, if they reject the offer, both parties get nothing, while if they accept, the split goes ahead as proposed. The rational choice is for the responder to accept any amount offered, however, in real life this is often not the case, as people often reject offers they deem to be 'unfair' (Camerer, 2003; Henrich et al., 2006). Participants were told that, unlike in the investment game where the 'other' person gets to make the final decision, in this task it is they who determine the outcome of the interaction. Participants engaged in 16 rounds of the game; three with each of the faces from the card game, who each offered £3, £4 and £5 and one with each of the unfamiliar faces who offered either £2[11], £3, £4 or £5. The order of the trials was randomised for each participant.

[11] This amount was only offered by an unfamiliar face, each of whom had to offer something different. This value (as opposed to, say, £6) was included because it may have been informative when considering offers accepted/rejected from non-veridical faces; if a low offer is rejected, it may be more related to the amount than the condition. However, if a very low offer is accepted from an unfamiliar face, while a higher offer from a non-

Procedure

All tasks were performed on a computer and participants were randomly assigned to one of the four versions of the task, where each face portrayed a different condition in each version (see Table C3, p.179). The rules of the card game were explained on screen and participants had to pass four 'sense-check' questions before they could start the game to ensure that they had read the instructions and understood the rules of the game. There was no mention of the faces in either the rules or checks. Once the card game was completed, participants were asked a series of socially oriented questions, the order of which was kept constant across participants. Unfamiliar faces, those who were not present during the card game, were introduced at this point, serving as both a baseline measure against which game faces could be compared and to break-up the game faces during their social appraisals. The experiment ended by revealing how well players had done during the card game and how much money they had won.

During debriefing, when participants we paid their winnings, participants were asked if they had noticed the behaviour of the faces during the card game. Many clearly indicated that they had, while a small number had failed to detect the contingency or notice that the faces were useful. Participants were also asked whether they had understood the rules for each of the tasks performed and all indicated that they had. The whole experiment took around 25 minutes to complete.

Exclusion criteria

The forced choice card game task is reasonably easy, so it was important to remove participants who were not paying attention or who were playing randomly. However, it is not obvious how this should be done. Those who are playing randomly can (by chance) score higher than 50%. Given that there is a 50% chance of guessing correctly on each of the 96 trials, we know from the binomial distribution that 46% of players adopting random strategies will score above 50% --- P(Bin(96,0.5) > 48) = 0.459. As such, it was insufficient to remove players who scored 50% or lower, as that would not remove 46% of those playing random strategies. However, it was also important not to

veridical face is rejected, it would suggest the condition is driving decisions. This point became moot, however, as even fair offers were rejected when made by non-veridical faces, meaning this condition was not analysed further.

remove players who were paying attention. A score of 62.5% can be achieved by playing the cards. Any approach for removing people playing randomly, therefore, needs to strike a balance with not removing those who were playing the cards, which was a valid strategy. Using a high level of confidence (such as 95%) to remove almost all of those playing randomly (true negatives) would greatly increase the chance of also removing those playing the cards (false negatives). With that in mind, we chose to use a 90% confidence level. Given a binomial distribution with 96 trials, only 10% of players with random strategies will achieve a score of 55/96 (57%) or more --- P(Bin(96,0.5) >= 55) = .092. By excluding those who score less than 55/96, then (on average) 90% of those playing randomly are removed, while we retain a high likelihood of not excluding those who are playing the cards, who are allowed a permissible error rate of five trials (60-55) before they would be removed. Nine participants scored less than 57% and were removed. The remaining 49 participants were carried through for analysis.

The card game and each of the social measures had their own set of analyses, as is laid out in the pre-registrations and specified in the result's section. Adjusted values are reported for any violations of sphericity, partial eta squared effect sizes are reported for Anovas and, unless otherwise stated, pairwise comparisons are Sidak adjusted, with Cohen's dz ($\mathbf{t / \sqrt{(n)}}$) used to report effect sizes.

Results

Data were analysed in a number of ways. Mixed effect models (performed using the lme4 and glm4 packages in R – abbreviated to 'mixed models' hereafter) were used to closely examine fixed effects of condition(s) when controlling for random effects of participants and stimuli (faces). For example, irrespective of condition, participants could be expected to vary both in terms of how they ascribe likeability ratings in general (random intercepts) and how much they discriminate between the faces (random slopes). Further, while the faces used were selected because they were rated as being similar by another group of people, this sample may perceive individual faces, and any differences between them, differently (random intercept and slope). Base model selection (random effects) together with the models used to explore fixed effects (and the bases for any adjustments made) and results are reported in **Appendix D: Chapter 4 – Mixed effects model**

selection and results (p.180), and referenced throughout the results presented here. More traditional tests, such as analyses of variance (Anovas), allow for a simple exploration of fixed effects, and the results for these tests can be found in **Appendix E: Chapter 4 – Traditional analyses** (p.196).

Card Game

Data from the card game allowed us to measure changes in accuracy and response times across the game, as well as any impact that condition (veridical vs non-veridical) might have. Specifically, we tested whether accuracy increased and reaction times decreased over time; both overall, and separately for veridical and non-veridical trials. Reaction time analyses can be found in **Appendix D1-1: Card Game** (p.182) (mixed models) and **Appendix E1-1: Card Game** (p.198) (traditional). Response times were highest for the first few trials, as participants familiarised themselves with the task, then decreased and stabilised across time. There was no difference in response times between veridical and non-veridical cues, suggesting the incongruent relationship between the cue and the outcome displayed by the non-veridical cues did not present a processing problem for participants.

Accuracy

Performance in the game varied, as some players clearly learned to use the face's gaze-cues very quickly and others did not. As can be seen, accuracy rates for probable trials were high throughout the game (see **Figure 13**, top) . Accuracy for improbable trials began low but increased over time (see **Figure 13**, bottom). It was these trials where participants could improve their accuracy by following the faces' gaze-cues. In this way, while accuracy analyses consider probable and improbable trials together, an increase in accuracy is mostly driven by participants getting more 'improbable' trials correct over time as they learn to use the faces' cues rather than the cards' values to guide their decisions.

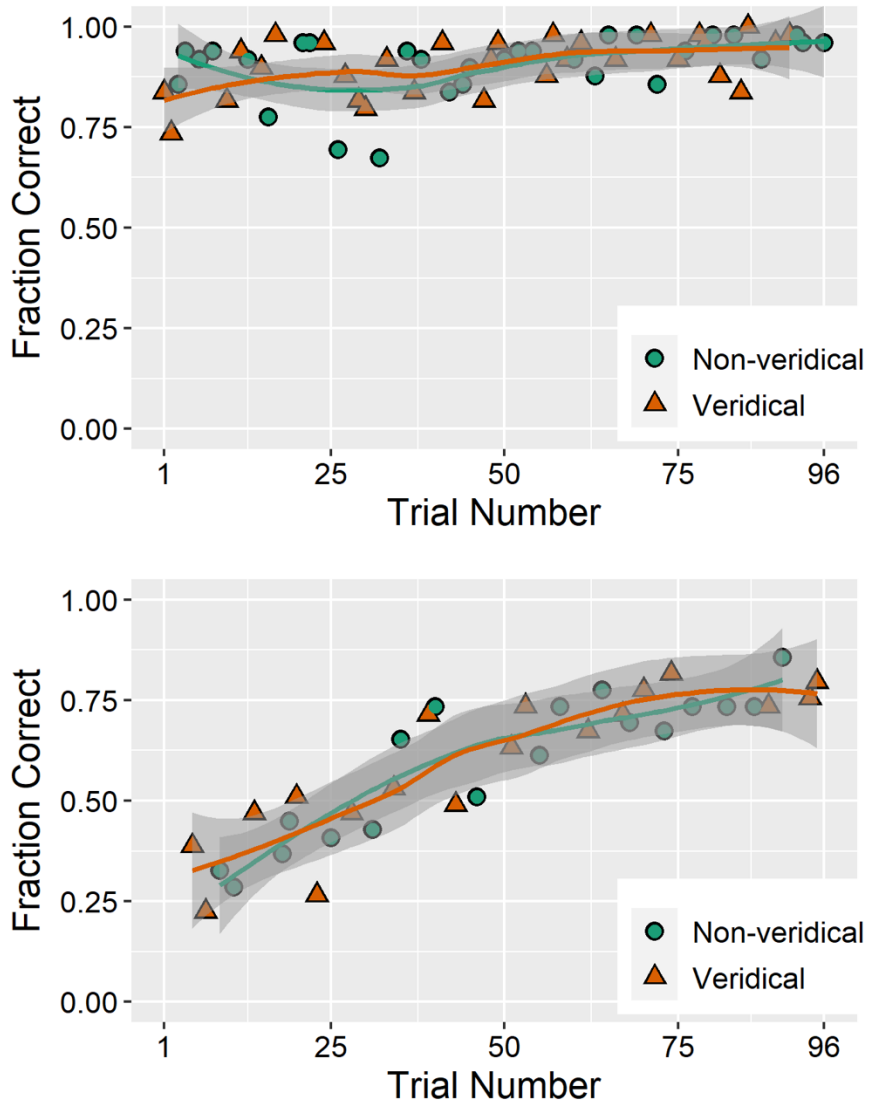


***Figure 13**: The proportion of participants getting each trial correct is displayed by the dots according to whether the trial was accompanied by a 'veridical' face (orange triangles) or a 'non-veridical' face (green circles). A smoothed function shows changes in group accuracy over time based on trial type. Shading represents confidence intervals. The difference in accuracy for probable (top) and improbable (bottom) trials can clearly be seen, with accuracy for probable trials remaining high and stable over time, while accuracy for improbable trials steadily increases over time as players begin to trust the faces' cues.*

For any given trial, participants can guess correctly or incorrectly. To test whether participants performed better over time, and were thus using the cues, general logistic mixed models were used. To test whether fixed effects of time and condition better predicted accuracy than the Base Model (BM), several models were used (p.181). Participants did increase their accuracy over time

($\chi^2(1)=40.07$, p< .0001). Adding a fixed effect of time to the BM significantly increased model fit, as is clearly illustrated in **Figure 13**. There was no main effect of condition ($\chi2(1)=0.05$, $p=.822$), meaning there was no difference in accuracy rates for veridical and non-veridical faces over time. Finally, there was no evidence of an interaction between Time and Condition ($\chi^2(1)=1.55$, $p=.214$). Again, this is illustrated by **Figure 13**, which shows that the learning rate for trials associated with veridical and non-veridical faces increases at the same rate.

Social Measures

Niceness

Linear mixed models were used to consider whether faces' behaviour during the card game predicted how nice participants thought they were. Analyses included the two veridical faces, two non-veridical faces, as well as four unfamiliar faces, which together produced three conditions; veridical, non-veridical, and unfamiliar (see **Appendix D1-2: Social Measures**, p184). for model details). There was a main effect of condition ($\chi^2(2)=29.11$, $p< .0001$). Pairwise comparisons showed that there was a difference between all combinations of conditions, where veridical faces were liked the most, followed by unfamiliar faces, while non-veridical faces were liked the least (see **Table 13**, left). Differences between the conditions were confirmed with repeated measures ANOVAs (see **Appendix E1-2: Social Measures**, p.200) and are clearly illustrated in **Figure 14**, left.

Trust

To explore the effects of faces' behaviour on investments in the trust game, the same approach as that used for niceness ratings was followed (see **Appendix D1-2: Social Measures**, p.185 for model details). There was a large effect of condition on investments made in the one-shot trust games ($\chi^2(2)=24.08$, $p< .0001$). Similar to niceness ratings, pairwise comparisons showed that each condition received a different amount than the others, with veridical faces receiving the most investment and non-veridical faces by far the least (**Table 13**, right). This is clearly visible in **Figure 14**, right. Differences between the conditions were confirmed with repeated measures ANOVAs (see **Appendix E1-2: Social Measures**, p.201).

	Niceness			Trust		
	Estimated difference	(t-ratio)	sig.	Estimated difference	(t-ratio)	sig.
Non-Veridical – Veridical	-2.78	(-8.1)	<.0001	-3.94	(-7.6)	<.0001
Unfamiliar - Veridical	-1.51	(-6.4)	.0001	-2.63	(-7.7)	<.0001
Non-Veridical - Unfamiliar	-1.27	(-4.3)	.0005	-1.31	(-4.4)	.0013

***Table 13**: Pairwise comparisons between conditions for niceness ratings (left) and investment decisions in the trust game (right)*

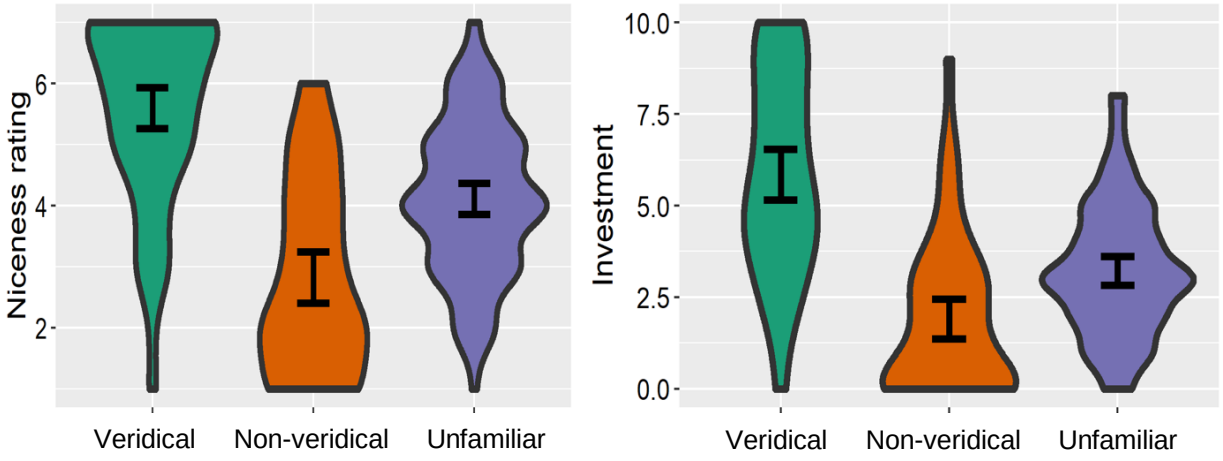


***Figure 14**: Violin plots for Niceness ratings (left) and Investment decisions in the trust game (right) for each condition. Bars represent 95% CI around the mean.*

Forced choice tasks

Pairs of faces were presented side-by-side over two rounds each of eight pairs of two forced choice alternatives. Responses were initially tested using binomial t-tests, with a test proportion of 50%. The results of these tests can be found in **Appendix E1-2: Social Measures**, p.201. For the first round, participants were asked with whom they would rather play more rounds of the card game and for the second, whom they would rather play with in more rounds of the investment game. For both rounds, each veridical face was paired with each non-veridical face and one unfamiliar face, producing 6 pairs. Each non-veridical face was also paired with one unfamiliar face, producing

two further pairs. The eight pairs in total thus produced three directional combinations for testing as follows:

- Veridical being selected over Non-veridical (4 pairings)
- Veridical being selected over Unfamiliar (2 pairings)
- Non-veridical being selected over Unfamiliar (2 pairings)

Within each combination, the condition of interest could be selected one or more times. For example, participants saw combinations of veridical and non-veridical faces four times, meaning they could choose the veridical face between zero and four times, making five possible occurrences. If there was no pattern in which face was selected, we would expect that each occurrence would occur with equal frequency. Chi squared tests were used to explore whether people tended to select one condition more often than another.

When asked to select a partner for more rounds of the card game, veridical faces were chosen substantially more than non-veridical faces (36/49 people chose veridical faces all four times), $\chi^2(4)=88.86$, $p<.0001$ and unfamiliar faces (36/49 people chose veridical faces both times), $\chi^2(2)=39.24$, $p<.0001$, while non-veridical faces were rarely chosen over unfamiliar faces (26/49 people never chose the non-veridical face), $\chi^2(2)=8.61$, $p<.001$. Despite the differing nature of the tasks in question, results were very similar when people were asked with whom they would rather play more rounds of the trust game. Veridical faces were again chosen over non-veridical (34/49 people choose veridical faces all four times), $\chi^2(4)=76.20$, $p<.001$ and unfamiliar (35/49 people chose veridical faces both times), $\chi^2(2)=35.06$, $p< .0001$ faces, while non-veridical faces were not selected over unfamiliar faces (28/49 people never chose the non-veridical face $\chi^2(2)=13.27$, $p=.001$.

For illustrative purposes, repeated pairings (e.g., a veridical face with a non-veridical face) were collapsed to produce the average proportion of times a condition was chosen over another across all participants. As can be seen from **Figure 15**, while the pair type being considered clearly influenced who was chosen, the type of game did not. To directly compare choices made in the card game with choices made for the trust game, the number of times a condition was selected in

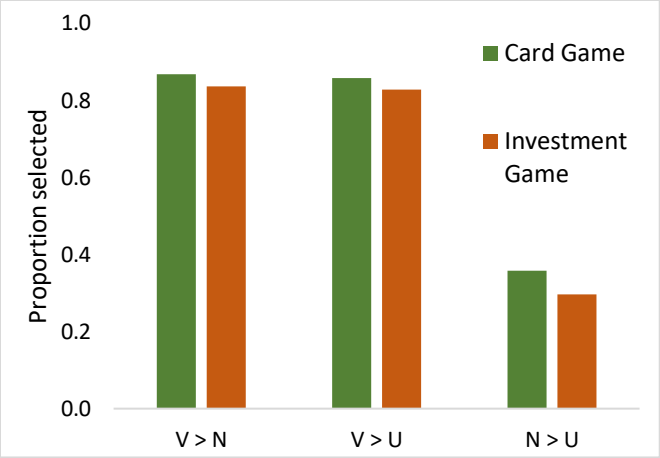


***Figure 15**: Repeat pair combinations collapsed for the card game and investment game forced choice rounds to show the proportion of occasions the condition stated on the left was selected over ('>') the condition stated on the right (V=veridical, N=non-Veridical, U=unfamiliar).*

any one pair combination type was treated as a continuous variable. Paired samples t-tests showed that there were no significant differences between the two tasks for veridical faces being chosen over non-veridical faces, t(48)=1.63, p=.110, dz= 0.23, veridical faces being chosen over unfamiliar faces , t(48)=0.70, p=.485, dz= 0.10, or non-veridical faces being chosen over unfamiliar ones, t(48)=0.90, p=.371, dz= 0.13.

Acceptance of offers in the Ultimatum Game

Each face from the card game made separate offers of £3, £4 and £5. Unfamiliar faces made one offer each of either £2, £3, £4 or £5. Participants could either accept or reject the offer. Responses from the £2 offer were omitted from analyses so that acceptance rates from unfamiliar faces could be compared to those for veridical and non-veridical faces. Acceptance rates for individual offers were tested using binomial t-tests, the results of which can be found in **Appendix E1-2: Social Measures**, p.202. Together, the offers conformed to a 3 (Condition: Correct, Incorrect and Unfamiliar) by 3 (Amount: £3, 4 or 5) design. We used a logistic mixed model to predict acceptance rates (see **Appendix D1-2: Social Measures**, 186).

Condition did affect acceptance rates ($\chi^2(2)=22.50$, p< .0001). Logit pairwise comparisons showed that the likelihood of an offer being accepted differs for each pairwise combination of conditions.

These effects are illustrated in **Figure 16**. The likelihood of acceptance is much more likely if the offer is made from a veridical face compared to a non-veridical ($p<.001$) or unfamiliar ($p=.005$) face, while an offer from an unfamiliar face is more likely to be accepted than one from a non-veridical face ($p=.046$). As expected, there was also a main effect of offer amount ($\chi^2(1)=28.84$, $p< .0001$). Unsurprisingly, when controlling for Condition, the higher the amount offered, the more likely an individual is to accept the offer. Finally, there was no interaction between Condition and Amount ($\chi^2(2)=0.39$, $p=.822$). While condition predicted acceptance rates, the influence of amount was similar across the three conditions. Interestingly, only ~60% of the £5 offers made by non-veridical faces were accepted. This is noteworthy as it means participants are rejecting a fair, even split of the £10 windfall, which, everything else being equal, is rarely rejected (Camerer & Thaler, 1995; Camerer, 2003; Camerer, 2010). Importantly, this means that participants are willing to pay a substantial price to 'punish' the non-veridical faces.

Effect of learning rate (accuracy) on person perception (niceness and trust)

Next, we tested whether success in the card game (i.e., overall accuracy) predicted niceness ratings and/or trusting decisions. Scoring highly in the card game required participants to learn the association between a face's cue and the correct answer. As such, faster learners would receive a higher overall score. Might the effect of condition on impressions be more pronounced for those who learned each face's contingency faster?

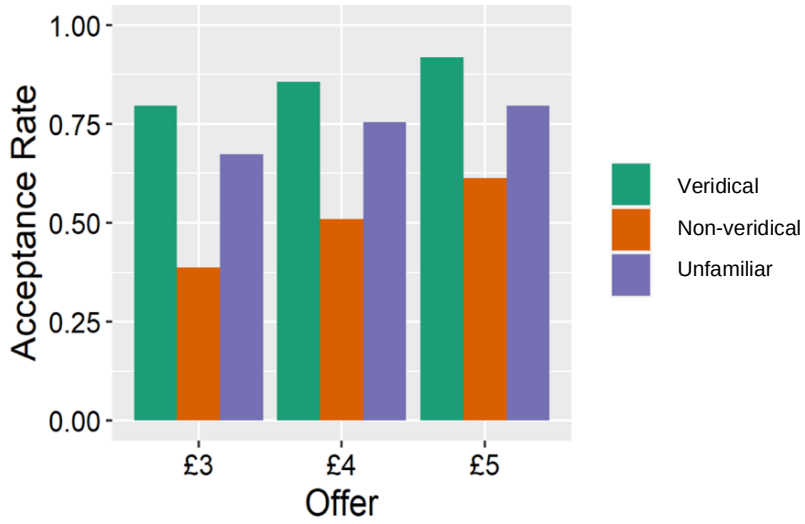


***Figure 16**: Acceptance rates for each amount offered by each condition in the Ultimatum Games.*

Linear mixed models were used to explore if performance in the card game predicted social judgements and decisions (see **Appendix D1-2: Social Measures** p.187, for full model details and results). There was no main effect of learning for either niceness ratings ($\chi^2(1)=0.45$, $p=.500$) or trust decisions ($\chi^2(1)=0.19$, $p=.660$). This was likely because differences between veridical and non-veridical conditions were averaged out when collapsed across different accuracy rates. To test whether there was an interaction between condition and learning rate, further models were compared. The interaction of condition and learning rate did predict both niceness ($\chi^2(2)=15.23$, $p=.001$) and trust ($\chi^2(2)=12.33$, $p=.002$). This is clearly illustrated in **Figure 17**, which shows that differences in judgements between veridical and non-veridical faces increases as accuracy increases. This suggests that those who learned the contingency early in the task went on to have more extreme impressions of the respective faces as compared to those who did not learn the predictive nature of the cues as quickly. Indeed, for those people who used a probabilistic strategy throughout the game, there was no discernible effect of validity at all (though see **Appendix E1-2: Social Measures**, p.204).

However, there is a tangible difference between not learning the facial contingencies at all and not learning them quickly and the interaction appears to be driven by those with the lowest scores. Those who scored above, say, 70% overall likely did eventually learn the association between gaze-cues and correct answers, albeit slowly. Those at 62.5% or below, however, likely simply used the cards; since using the cards and ignoring the faces leads to an accuracy rating of 62.5%. If we exclude the participants (n=6) who scored less than or equal to 62.5%, then the interaction remains significant for niceness ratings ($\chi^2(2)=7.54$, $p=.023$) but is at only trend levels for Trust ($\chi^2(2)=4.99$, $p=.083$).

In broad terms, once accuracy levels reach around 80% or more, ratings seem to stabilise. Importantly, unfamiliar faces were perceived similarly irrespective of performance during the game. This suggests that differences between how the card game faces were perceived across participants was the product of learning and the faces' behaviours.

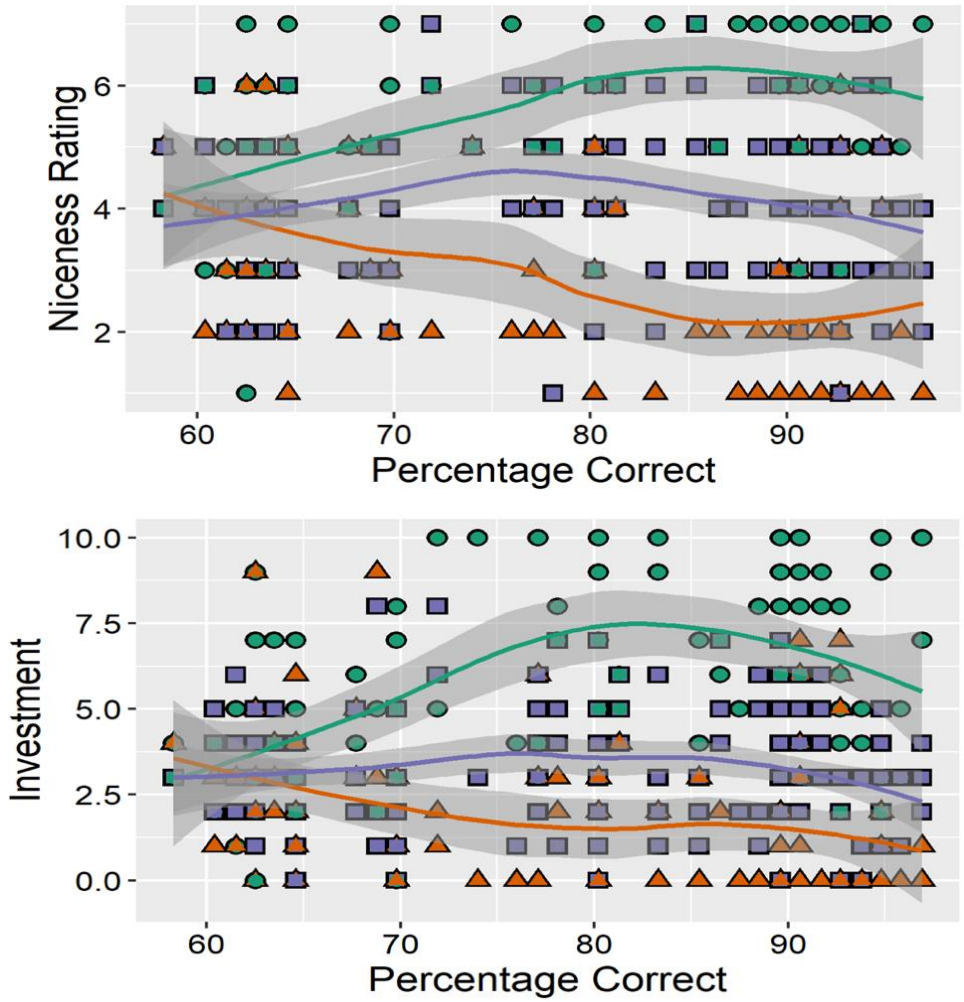


Figure 17: *Individual participants' niceness ratings (top) and investment decisions (bottom) for each condition plotted against their overall score in the card game. Green circles represent 'veridical' faces, orange triangles represent 'non-veridical' faces and purple squares represent 'unfamiliar' faces. The smoothed line represents average ratings and decisions by accuracy.*

Discussion

Although faces were not referenced before the task, most people quickly detected that their gaze-cue behaviour provided information about the outcome of a trial. By the end of the game, most players were using the gaze-cues (social information) over the cards (probabilistic information). Although non-veridical faces looked at the wrong answers, causing conflict players must overcome to choose the right answer, and veridical gazers simply looked at the correct answer, there was no difference in how quickly players learned to use cue contingency for the two types of faces. In this

way, the non-veridical cue providers' contrary, yet consistent, form of gaze-cues came to be trusted and acted upon in the same way as were the more straight-forward veridical cue providers. Participants earned as much money in interactions with veridical and non-veridical faces. Despite this, veridical cue providers were consistently perceived favourably, while non-veridical cue providers were consistently perceived negatively, suggesting players perceived the gaze cues in a social context and judged veridical faces as 'helpful' and non-veridical faces as 'deceptive'. Importantly, contrary to our expectations, participants were consistent across all judgements and interactions. Not only did they not like or trust the non-veridical faces, they were willing to lose money in order to punish them and even chose *not* to play the card game with them again, despite knowing they would win more money with non-veridical faces than they would with a new player they did not yet know.

This first Experiment demonstrated that (most) people both learn to detect and use cooperative and deceptive cues equally, and that the perceived valence of others' behaviour, not its economic utility, appears to predict overall impressions and a range of social decisions. Indeed, negatively perceived cue providers are not selected for future interactions even though in game they behaved consistently and always predict the correct answer. Experiment 2 explores these findings further by assessing how changes in gaze-cue behaviour across the game might affect impression formation, particularly for faces that change their cue-contingencies and, thus, are both unpredictable and cost players money.

Experiment 2

In Experiment 1, we established that exposure to (three blocks of) wholly veridical and non-veridical gaze-cue behaviour was sufficient to nearly all players to learn the face contingencies and generate highly positive (or negative) impressions. Here, we explore what happens when some faces change their behaviour after a period of time. In the real world, people themselves are often inconsistent and can change how they behave over time, making impression formation a dynamic process (Brambilla et al., 2019; Mende-Siedlecki, et al., 2013), where we must adjust to others' unexpected behaviours and incorporate conflicting information. In this Experiment, two of the four faces in the card game reversed their helping behaviour (e.g., changed from providing all

cooperative to all deceitful cues) after the third block and displayed only their new cue-answer contingency for an additional three blocks (i.e., the last half of the game).

In the card game, a change in the faces' behaviour will cost players money if not detected, meaning the association between a face's behaviour and a trial's outcome must be updated if they want to continue to do well. Importantly, if players have learned the initial cue contingencies, they can recover from a violation quickly, as their outcome is not concomitant with the target's behaviour. In the trust game and traditional gaze-cueing tasks, however, if a target negatively changes their behaviour, the perceiver is destined to be negatively affected, and a switch from sharing/helping to stealing/deceiving is met with a reduction in investments and a downgrading of impressions (Campellone & Kring, 2013; **Chapter 2, Experiment 3**). This distinction could mean that initial impressions formed during the card game are not updated following a change in targets' behaviour, as there is no (long term) corresponding impact on performance. Given just how divergent impressions of veridical and non-veridical faces were in Experiment 1, however, this seems doubtful. If impressions follow the perceived intentions of targets' behaviours, then a person who changes to always displaying non-veridical gaze-cues should no longer be thought of favourably. Conversely, someone who changes from all non-veridical to all veridical cues may be seen as having a change of heart and hence come to be seen more positively. Considering how negative initial impressions were, however, the degree of any change is unknown. In judgement based studies where inconsistent, indirect moral behaviours are presented sequentially, greater updating is found for moral-to-immoral changes in behaviour than for immoral-to-moral changes (Brambilla et al., 2019; Mende-Siedlecki, 2018; Mende-Siedlecki, et al., 2013; Mende-Siedlecki, et al., 2013; Mende-Siedlecki & Todorov, 2016; Reeder & Coovert, 1986). Here, then, we might expect to find that the difference between a consistently cooperative person and a person who was once cooperative but later deceives us, is greater than the difference between a consistently deceitful person and someone who was once deceitful and later changes to cooperate.

Method

Power and Open science

Sample sizes and hypotheses were pre-registered (aspredicted.org/KKB_SF6). All manipulations and measures are reported, and all exclusions are specified. A smaller effect size was targeted here, as perceived differences between faces may be less marked when the number of cue profiles is increased and overall differences between faces are decreased. An a priori power analysis using a two-tailed, paired t-test as the condition of interest with 80% power to detect Cohen's dz of 0.35 or higher returned a requisite sample size of 67.

Participants

Seventy-two (female 49, male 22, unspecified 1; mean age years 21.9, SD 3.6) students at Bangor University, UK completed the experiment. Each participant was compensated for their time with course credits or, where data collection did not take place in person, a nominal fee of £4.00. Due to lockdown restrictions, the final 15 datasets were collected online with no experimenter present. All participants were rewarded for their performance in the card game, where a maximum of £5.76 could be won. The experiment was granted ethical approval by the School of Psychology at Bangor University's Ethics Committee, and all participants provided explicit consent prior to data collection taking place.

Stimuli & Procedure

This experiment was identical to Experiment 1, with one exception. The card game was extended to include a further three blocks and incorporated a change in two faces' gaze-cue behaviour (see **Table 14**). There were now 192 trials in total. The first three blocks of the card game exactly the same as those experienced by participants in Experiment 1; only the final three blocks differed. In the last three blocks, the ratio of overall correct to incorrect answers associated with faces remained the same (50:50), only the faces providing some of them changed (50%). One face originally giving wholly veridical cues swapped to giving non-veridical cues, and one face originally giving wholly non-veridical cues swapped to giving veridical cues.

	Face 1 '100-100'	Face 2 '100-0'	Face 3 '0-100'	Face 4 '0-0'
Blocks 1-3	100	100	0	0
Blocks 4-6	100	0	100	0
Average	***100***	***50***	***50***	***0***

Table 14*: Each face's gaze-cue profile, expressed as a percentage of veridical cues per block, for the first (1-3) and second (4-6) half of the card game, together with their overall proportion of veridical cue-gaze behaviour.*

Data exclusion criteria

The same procedure as that used for Experiment 1 was used here to remove participants who were playing randomly. Those who scored less than 57%, on average, over the first three blocks of the card game were removed from all analyses. A total of six participants were removed, leaving 66 datasets for analyses. Fifteen datasets were collected online and without experimenter supervision. It was noted that one participant took an inordinately long time (> 1hr, with one trial response time of 32 minutes) to complete the card game and another spent a very long time on block 6 only (averaging 11 seconds per trial as compared with less than 2 seconds for other blocks). So as not to skew the results, these two participants were removed from all reaction time analyses, resulting in 64 datasets being included (see **Appendix D2-1: Card Game**, p.191). Both of these participants' overall accuracy levels, however, were high (average score across all blocks: 83 and 80%, respectively), so they were retained for analyses using accuracy and all social measure analyses.

Results

Where possible, results were analysed using mixed models. Full details of all model selection and outcomes can be found in **Appendix D2: Experiment 2**. Results of more traditional tests or additional analyses are reported in **Appendix E2: Experiment 2**.

Card Game

In Experiment 2, measuring accuracy and response times was slightly more difficult. Since half of the faces changed their behaviour after trial 96, accuracy rates were expected to drop, and reaction

times rise, immediately following the change in cue-answer contingency for two of the faces. This adds non-linearity to the data, making it difficult to assess without using complex models (which struggle to converge). Therefore, to better understand the effects of time and condition on accuracy and reaction times, the data was broken into subsets which were analysed separately as follows:

- First 96 trials (Blocks 1-3) – The three blocks prior to strategy change (same as Exp. 1)
- Last 96 trials (Blocks 4-6) – The three blocks after strategy change (unique to Exp. 2)
- Trials 97 – 128 (Block 4) – The block immediately following validity changes
- Last 64 trials (Blocks 5-6) – The last two blocks, after the participants had some time to learn the change in two faces' behaviour

Reaction time analyses can be found in **Appendix D2-1: Card Game** p. 191 (mixed models) and **Appendix E2-1: Card Game** p.211 (traditional). Later, the relationship between accuracy (learning) and the magnitude of social impressions was explored.

Accuracy

As anticipated, performance increased over the first three blocks of the game. As can be seen in **Figure 18** below, accuracy on improbable trials started low before approaching ceiling just prior to the two faces changing their behaviour. Again, accuracy rates appeared to be similar for each of the faces, regardless of their cue-outcome contingency. There was a temporary drop in accuracy in the fourth block when the two faces reversed their cue veracity. Interestingly, it took very few trials for people to update their belief about the faces' cue veracity and adjust their behaviour accordingly, and a change in some faces' behaviour even caused brief doubt in the consistently of the other face's trustworthiness (see **Figure 18**, top – green circle around trial 100).

As in Study 1, to test whether participants performed better over time, general logistic mixed effect models were used (see **Appendix D2**, p.189). In terms of time, it was clear that individuals increased their accuracy over time for the first three blocks ($\chi^2(1) = 73.74$, $p < .0001$), the last three blocks ($\chi^2(1) = 68.18$, $p < .0001$) and the fourth block ($\chi^2(1) = 82.90$, $p < .0001$). However, accuracy rates did not increase over the 5^{th} and 6^{th} blocks ($\chi^2(1) = 0.76$, $p = .384$), demonstrating that participants had learned the faces' cue-answer contingency by the end of the fourth block.

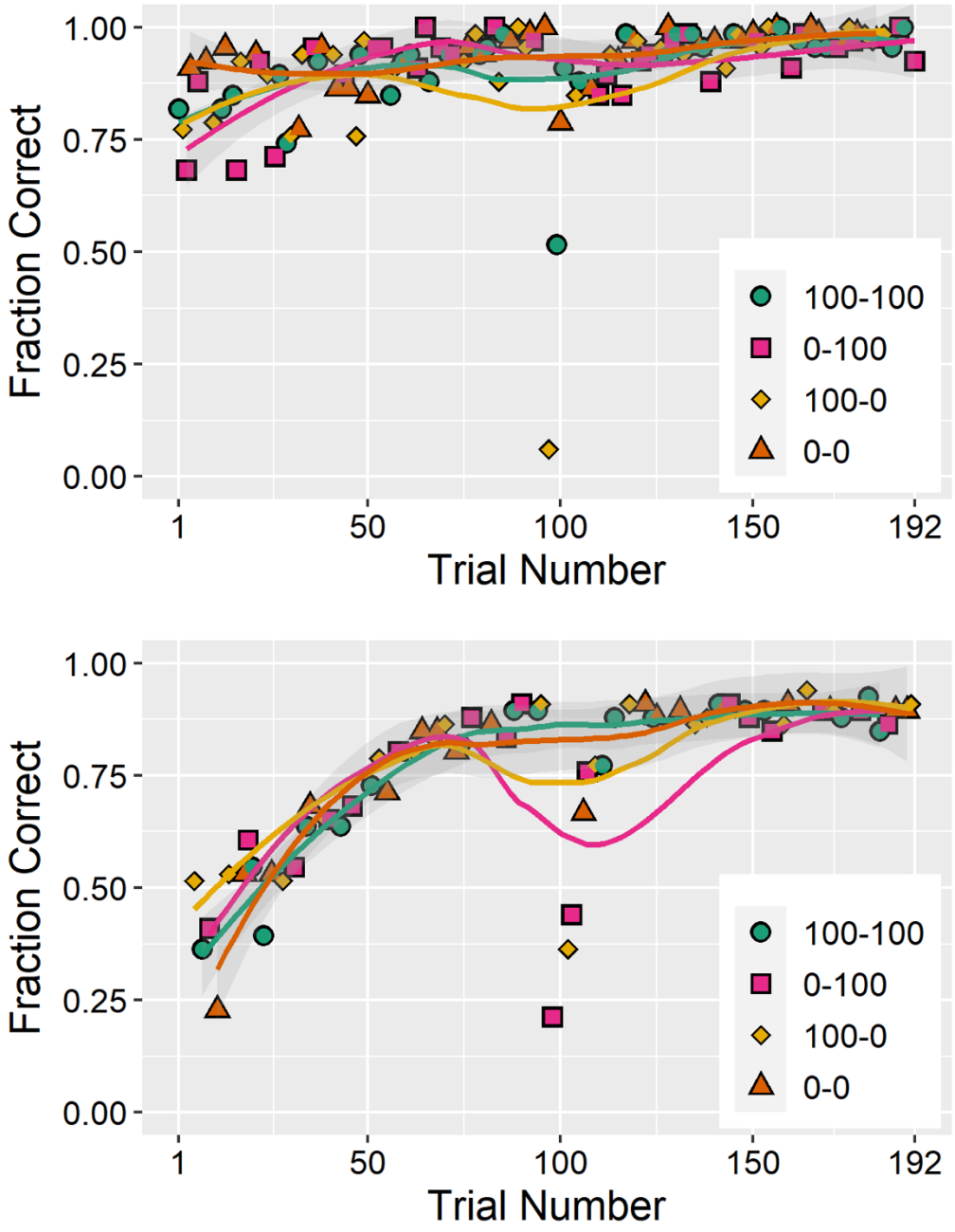


***Figure 18**: Scatter plots showing average accuracy scores for each trial separated by condition (initial – final proportion of veridical cues) and trial type: probable (top) and improbable (bottom). By around trial 50, errors have all but disappeared, as faces' gaze-cues are used. Errors occur briefly after trial 96, when two of the faces reversed the relationship between their gaze-cue and the answer. Shaded area represents 95% confidence intervals.*

As expected, and in line with Experiment 1, there was no main effect of condition on accuracy rates during the first three blocks ($\chi^2(3)=4.48$, $p=.214$), but there was a main effect of condition in the final three blocks ($\chi^2(3)=42.83$, $p< .0001$). However, this effect was driven by the drop in accuracy in the rounds immediately following the change in strategies (Block 4). When looking at Block 4 on its own, accuracy for the two faces who changed their behaviour dropped at the beginning of the block. However, by the time participants played the last two blocks, accuracy rates were no longer significantly predicted by condition ($\chi^2(3)=6.25$, $p=.100$), demonstrating that by the end of the task, the accuracy rates for all four faces were comparable. Indeed, by the end of the task, participants were close to ceiling level performance. The interaction of condition and time was not assessed, as the model would not converge with any random slopes. Visual inspection of the data (**Figure 18**) does not suggest that there is any meaningful interaction present.

To illustrate the degree of learning experienced within the sample, frequency histograms for the fourth block, in which updating took place (see **Figure 19**, left), and average accuracy for the final two blocks (see **Figure 19**, right) can be used. As can be seen, by the end of the task, 48/66 people got every single trial correct and a much smaller, distinct group of seven people had failed to detect (or use) the faces' cues when making their decisions.

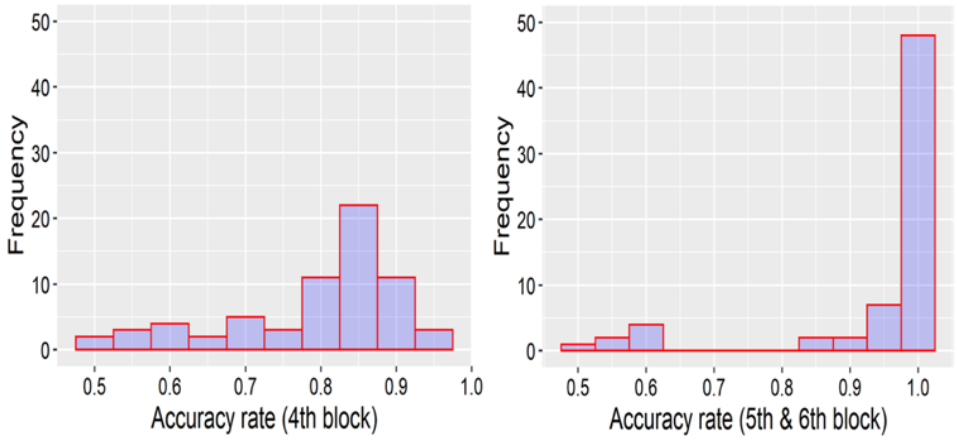


***Figure 19**: Frequency distribution of accuracy scores (percent correct) during the card game for the 4th block only (left) and final two blocks together (right). The 4th block shows variation in scores while the final blocks shows two, distinct and different sized groups.*

Social Measures

Niceness ratings

Participants again rated each of the eight faces for niceness on a scale (1-7) and differences in ratings between the faces were explored using linear mixed effect models (see **D2-2: Social Measures** p.192). There was a large main effect of condition for niceness ratings ($\chi^2(4)= 104.47$, $p< .0001$). Excluding comparisons with unfamiliar faces, paired contrasts found a significant effect for every combination of card game faces (**Table 15**, left). This effect is illustrated in **Figure 20** (left). The face that helped the most (100-100) was liked the most and, correspondingly, the face that helped the least was liked the least. The two changing faces, while helping the same amount overall, were not judged to be the same, nor were they perceived similarly to either the fully veridical face or the fully non-veridical face. Rather, the face that initially helped but changed to never helping was liked less than the face that did the opposite (i.e., the most recent behaviour had more influence on the final niceness ratings than the initial behaviour). Finally, once again, the unfamiliar faces were judged somewhere in the middle of the card game faces. A repeated-measures Anova validated these results (see **Appendix E2-2: Social Measures**, 212).

	Niceness			**Trust**		
	Mean difference	SE	Sig.	Mean difference	SE	Sig.
100-100 and 0-100	1.12	0.24	<.001	1.89	0.34	<.001
100-100 and 100-0	1.97	0.27	<.001	3.59	0.38	<.001
100-100 and 0-0	2.92	0.27	<.001	4.47	0.38	<.001
100-0 and 0-100	-0.85	0.27	.0150	-1.70	0.38	<.001
100-0 and 0-0	0.95	0.24	.0006	0.89	0.34	.0662
0-100 and 0-0	1.80	0.27	<.001	2.58	0.38	<.001

***Table 15**: Results of paired contrasts between conditions in the card game for niceness ratings (left) and investments in the trust game (right). Mean differences are shown in original units.*

Trust

Participants engaged in one-shot trust games with each of eight faces and differences in amounts sent were explored using linear mixed effect models (see **D2-2: Social Measures** p.192). For the models, five conditions were including, four card-game faces and one that averaged across the four unfamiliar faces. There was a large main effect of condition on trust ($\chi^2(4)=128.18$, $p<.0001$). Pairwise comparisons, again excluding those with unfamiliar faces, were checked to see how investments made differed between pairs of faces. Each of the card game faces was, on average, sent a different amount (see **Table 15**, right). The pattern of results was similar to that found for niceness ratings, with the exception of the condition where the face became unhelpful after initially helping (100-0; see **Figure 20**, right). Participants trust this face only slightly more than the fully non-veridical face (quote stats). To validate these results, we ran a repeated-measures ANOVA which also found differences between all card game conditions, except 100-0 and 0-0 ($p=.081$) (see **Appendix E2-2: Social Measures**, 212).

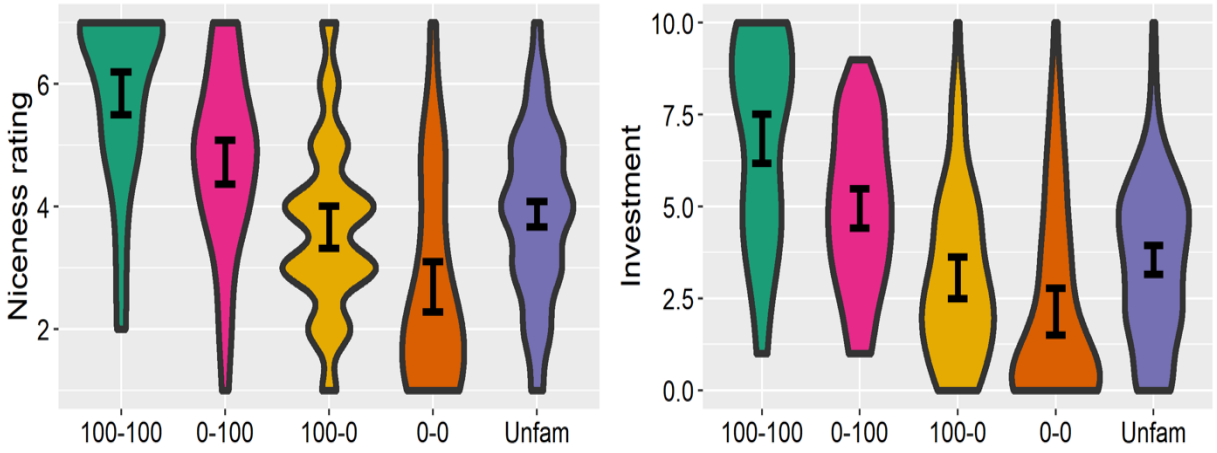


Figure 20: *Violin plots showing niceness ratings (left) and investment decisions (right) for each of the four card game conditions (initial – final proportion of veridical cues) and unfamiliar ('Unfam') faces. Responses for the four unfamiliar faces were collapsed and their average is presented here. Bars represent 95% CI around the mean.*

Forced choice

There were eight pair combinations for each round of the forced choice tasks. Each pair was tested to see whether one condition (left in table) was selected at an above chance (50%) level using 2-tailed binomial t-tests. The results of the tests are shown in **Table 16**. While the face that was chosen clearly depended on the pair being presented, the selection of the face within each pair was once again the same regardless of whether participants were choosing a future partner for more rounds of the card game (**Table 16**, left) or trust game (**Table 16**, right). These results clearly show that the wholly veridical face (100-100) was selected most of all. Of the faces that changed, the direction seemed to matter, with the face that changed to provide only veridical cues being chosen far more often than both the face that changed to offer only non-veridical cues and an unfamiliar face. Interestingly, even when a face changes to only provide non-veridical information – which could be construed as particularly negative – they are still selected for both future trust games *and* card games when pitted against a wholly non-veridical face.

Acceptance rates in the Ultimatum Game

As in Experiment 1, participants received offers of varying amounts (£3-5) from each of the five conditions. Offers could be accepted or rejected in a binary fashion. Logistic mixed models were

	More rounds of the Card Game		**More rounds of the Trust Game**	
	Proportion	**P value**	**Proportion**	**P value**
100-100 > Unfamiliar	.91	<.001	.94	<.001
100-100 > 0-100	.83	<.001	.83	<.001
100-100 > 0-0	.88	<.001	.86	<.001
100-0 > Unfamiliar	.59	.175	.52	.902
100-0 > 0-100	.21	<.001	.23	<.001
100-0 > 0-0	.64	.036	.68	.004
0-100 > Unfamiliar	.79	<.001	.80	<.001
0-0 > Unfamiliar	.47	.712	.41	.175

Table 16: *Outcomes of binomial t-tests for forced choice tasks concerning partner choice for more rounds of the card game (left) and more rounds of the trust game (right) for each pair choice. The proportion selected refers to the condition listed first.*

used to predict acceptance in the Ultimatum Game based on the amount being offered and the face (condition) making the offer (see **D2-2: Social Measures,** p.193). There was a clear main effect of amount ($\chi^2(1)=40.08$, $p< .0001$), with higher amount receiving higher acceptance rates, as might well be expected (see **Figure 21**). The face (condition) making the offer also predicted acceptance rates ($\chi^2(4)=49.76$, $p<.001$). Follow-up pairwise comparisons, excluding contrasts with the unfamiliar condition, can be seen in **Table 17**. Roughly speaking, the results echo the prior tests, as participants were numerically more likely to accept offers form the entirely veridical face, followed by the face that became helpful, then by the face that became unhelpful. However, differences between these three face-types were relatively small, and (mostly) non-significant. The 100% non-veridical face, however, was significantly different than all other faces, and offers from that face were rejected the most, despite being encountered among inconsistent others that actually cost participant's success and money in the game.

Models designed to test for any interaction between condition and validity were problematic and are not reported here (see **Appendix D2-2: Social Measures**, p.193). A factorial Anova on minimal acceptable offers (MOA) was performed and is reported in **Appendix E2-2: Social Measures**, p.214. This considered the amount at which participants would accept an offer from each condition. The main effects of amount and condition were present, but there was no interaction.

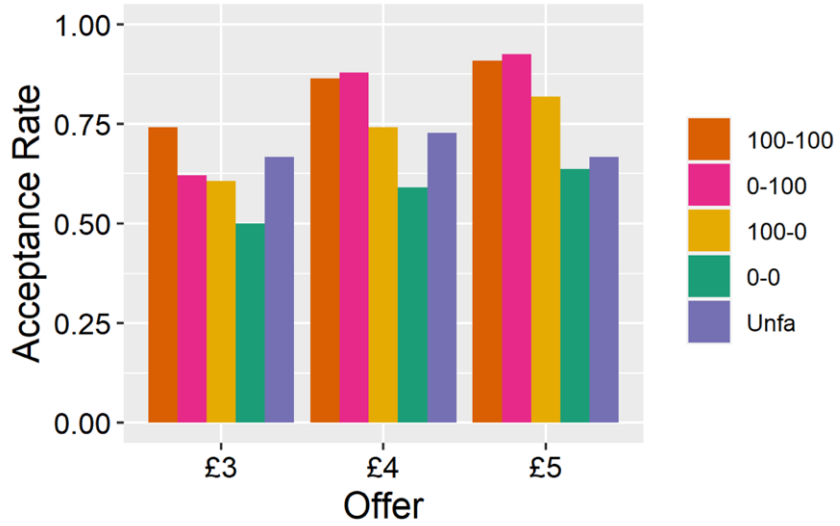


Figure 21: Average acceptance rates in the Ultimatum Game for each condition and amount it offered.

Conditions	p value
100-100 and 0-100	.9232
100-100 and 100-0	.0305
100-100 and 0-0	<.001
100-0 and 0-100	.2157
100-0 and 0-0	.0111
0-100 and 0-0	<.001

***Table 17**: Paired contrasts for differences in acceptance rates in the Ultimatum game between conditions (ignoring amount offered). Due to the nature of the analyses performed, data such as confidence intervals, is not available.*

Effect of learning rate on person perception

As with Experiment 1, we were interested whether participant performance in the card game influenced their final impressions about faces. As the faces' behaviour was more complex here, a measure of learning rate was non-trivial. Looking only at accuracy, most people learned the change in validity quickly, and accuracy rates for Block 5 and 6 approached ceiling (see **Figure 19**, right). Thus, we chose to use accuracy rates for Block 4 alone to look at learning, as it had a larger variance in accuracy across individuals than the final three blocks as a unit. For parity with Study 1, the influence of accuracy rates for just Blocks 1-3 was also explored. However, given the subsequent changes in faces' behaviours, it could be expected that accuracy here would not predict impressions collected after the changes had been experienced. Alternatively, high accuracy in the earlier part of the game might have a greater impact on updated impressions, as those participants would have perceived the importance of the faces sooner and followed more gaze-cues and earned more money across the whole game than those who learned more slowly.

Using linear mixed models, participants' accuracy scores, both averaged across Blocks 1-3 and for Block 4 alone, were used to predict niceness ratings and investment decisions in the trust game (see **D2-2: Social Measures,** p.194). As in Experiment 1, there was no main effect of accuracy when considering Blocks 1-3 ($\chi^2(1)=1.77$, $p=.183$) or Block 4 ($\chi^2(1)=0.48$, $p=.487$) (see **Figure 22**, left). The same held for investment decisions, where there was no main effect of learning on

trust for either Block 4 ($\chi^2(1)=0.13$, $p=.719$) or Blocks 1-3 ($\chi^2(1)=1.09$, $p=.297$) (see **Figure 22**, right).

Again, as in Experiment 1, there was a significant interaction of condition and learning when considering both Block 1-3 accuracy rates ($\chi^2(4)=18.74$, $p=.009$) and Block 4 accuracy rates ($\chi^2(4)=16.97$, $p=.002$) on niceness ratings. The results were similar for trust where there was a significant interaction when considering Block 1-3 accuracy rates ($\chi^2(4)=23.75$, $p<.0001$) and

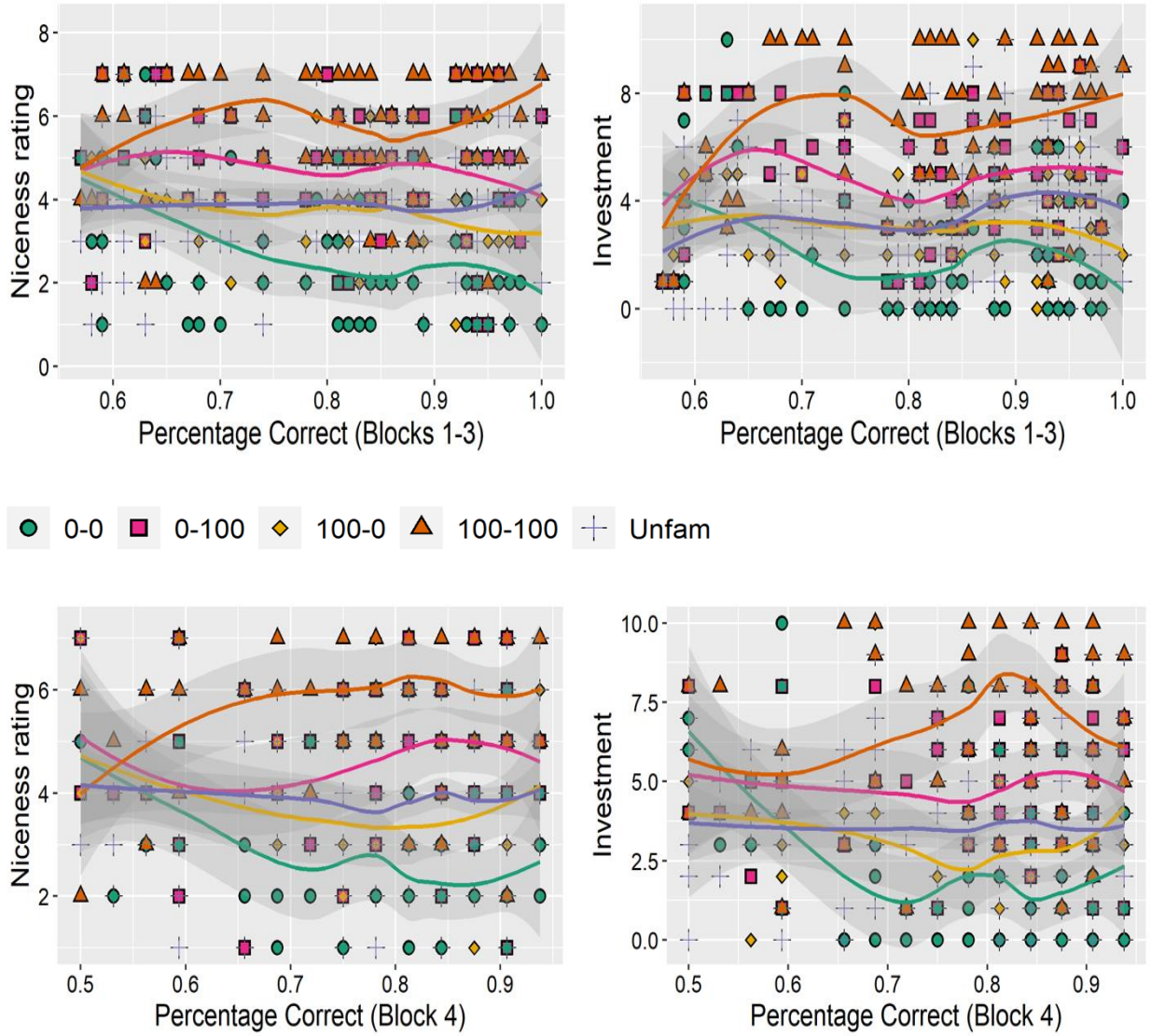


***Figure 22**: Plots showing niceness ratings (left) and investments made (right) with each condition given overall accuracy scores for all participants for Blocks 1-3 (top) and Block 4 only (bottom).*

significant interaction when considering Block 1-3 accuracy rates ($\chi^2(4)=23.75$, $p<.0001$) and Block 4 accuracy rates ($\chi^2(4)=18.61$, $p<.001$). Visual inspection of the graphs suggests, however, that these interactions are driven by a lack of differentiation between conditions among low-scoring participants. In other words, those who did not learn the faces' gaze-cue contingencies also did not form distinct social impressions about them. For those who did learn, be it quickly or slowly, judgements and decisions appeared to be similar across all scores. At the end of the game, just seven players were still using the cards to determine their choices. If these seven players (i.e., those who did not learn) are excluded, then the interaction effect goes away for both niceness ratings (Blocks 1-3 ($\chi^2(4)=7.56$, $p=.109$) and Block 4 ($\chi^2(4)=2.56$, $p=.635$) and trust decisions (Blocks 1-3 ($\chi^2(4)=7.96$, $p=.093$ and Block 4 ($\chi^2(4)=1.91$, $p=.753$).

Discussion

Replicating Experiment 1, players again quickly learned to use faces' gaze-cues to increase their performance in the game. Somewhat unexpectedly, they were also extremely quick to update their expectations of faces' behaviour in response to two of the faces changing their cue-answer contingency halfway through the game. Players' accuracy was only briefly affected by the change. Participants also updated their impressions of faces in the direction of their change. Indeed, although faces that changed on average gave the same number of veridical and non-veridical cues (50%), participants formed more positive impressions of the face that became more veridical than the face that became less veridical. Together with the results from Experiment 1, these findings demonstrate that initial impressions (as inferred from Experiment 1) are revised with new experience when faces change the valence of their behaviour and that this revision occurs is spite of there being no corresponding (long-term) change to performance.

Importantly, participants apparently again judged faces entirely on the social valence of their behaviour, with little reference to their economic utility during the game. The wholly veridical cue provider was, unsurprisingly, still perceived the most favourably. Somewhat surprisingly, the wholly non-veridical cue provider was still perceived the least favourably, even when encountered among others who acted inconsistently, causing disruption to performance and necessitating participants to change strategy. This suggests that even someone who turns against us after helping

us is preferable to someone who always deceives us yet consistently benefits us. This implies that consistency is only valued in those we perceive as *intending* to help us, rather than those that *actually* help. Indeed, only offers from wholly invalid faces were consistently rejected in the ultimatum games, suggesting that they were the only face-type that participants were consistently willing to pay a cost to 'punish'. That being said, the face who changed from veridical to non-veridical cues was perceived more negatively than his changeable counterpart, who replaced non-veridical with veridical cues and was viewed more favourably. This indicates that the direction of change – rather than overall valence of behaviour – does matter. While some help is better than none, becoming deceitful is perceived much more negatively than becoming helpful.

This Experiment used the same gaze-cue profiles[12] as that of **Chapter 2, Experiment 3**. In that Experiment, the two faces who both ended by providing helpful behaviour were perceived differently, with the face who was consistently helpful being trusted the most while faces who both ended by displaying deceitful behaviour were perceived similarly (see **Figure 4**, p.65). Here, all four faces were perceived differently, with the face who changed to provide only non-veridical cues being preferred to the face that always provided non-veridical cues. Though the parameters of the task are somewhat different, for the wholly 'invalid' face to be the most disliked person in the present context might reflect the fact that here, their uncooperative intention was more apparent, or easier to infer, making its consistently negative connotations that much more impactful. Conversely, helping behaviour when participants are trying to learn how to play the game may be particularly impactful and impossible to fully 'forget' so that initial positive impressions may not be fully written-off when a person later behaves deceitfully. Alternatively, owing to the implicit nature of learning in the gaze-cueing task as compared to the explicit acquisition of knowledge in the card game, effects are likely to be smaller in the gaze-cueing task and, therefore, differences between the conditions may simply be harder to detect. With a larger sample size, it is possible that the same differences found here may be present when using the traditional task, albeit with smaller effect sizes. Indeed, when considering the differences between the two learning environments, it is encouraging to see a not too dissimilar pattern of impressions being formed.

[12] Same as regards the predictive faces. In Chapter 2, six faces were included in the task and the two 'control' (non-predictive) faces from Experiment 3 were absent here.

General Discussion

Across two experiments we find evidence that it is the perceived valence of a person's behaviour, and not its economic utility or effect on the perceiver's outcome, which dominates impression formation and social decision-making processes. Rather strikingly, social motives appear to drive decisions even when the disliked person can neither negatively affect us nor be negatively affected by us, and we are willing to pay a cost to avoid them. When faces changed their behaviour during the encounter, participants readily updated their impressions upwards or downwards, depending on the direction of change. This is similar to other studies, where a change in the valence of another's behaviour has a congruent effect on a perceiver's outcome (Campellone & Kring, 2013; King-Casas et al., 2005), only here, the direction of change in behaviour was of no financial consequence. Further, in both consistent and inconsistent environments, the face that only ever provided non-veridical cues was unequivocally disapproved of, suggesting that perceived unwavering negative intention was particularly important to people when making decisions. Moreover, results demonstrate that people attribute meaning to behaviours (Macrae & Quadflieg, 2010; Malle, 2011; Reeder, 2009) even when the behaviour has not been referenced and is not explicitly explained or diagnostic of a character trait. Together, results highlight how sensitive we are to the meaning behind people's behaviour (Ames et al., 2011; Frith & Frith, 2006, 2011), and how much we value cooperation and dislike 'liars' (Balliet et al., 2011; Baumeister et al., 2001; Henrich et al., 2006). Indeed, our impression about the other is so pervasive that we use our social belief to make economic decisions.

In previous studies where a target's behaviour is learned over time, the valence of their behaviour is often both clearly positive or negative and is congruent with the perceiver's outcome. In these scenarios, the other's intentions can be readily inferred from their behaviour, and negative behaviour generates negative consequences (Campellone & Kring, 2013; Chang et al., 2010; Fouragnan et al., 2013; King-Casas et al., 2005; Maurer et al., 2018; Phan et al., 2010; Zarolia et al., 2017). Here, however, any meaning behind a person's behaviour required additional decoding and participants were actually helped by non-veridical cue providers, thus presenting participants with a potential conflict between social valence and economic utility. Our results, however, provide no evidence that participants suffered from any such dilemma; all their post-game

judgements and decisions indicate that they had categorised the non-veridical person as 'bad', even to the point of being willing to pay a cost themselves to punish them (in the ultimatum game).

A recent study by Hackel and colleagues (2020) explored how people learn about and perceive others based on their sharing behaviour when splitting a pool of money. Participants learned about people or slot machines (i.e., non-people) based on their individual reward value (allocated amount) and generosity level (allocated proportion), where a target person could be selfish but have a high reward value or generous with a low reward value. It was found that for human targets, participants both preferred generous people and relied on them more when learning, whereas for slot machines, both learning and impressions were driven only by their reward value. In line with our findings, this work demonstrates that the trait exhibited by a person's behaviour means more to us than the benefit we obtain from their behaviour (i.e., we value generosity more than reward value). Similarly, we find that an inferred negative intention behind a less overtly diagnostic behaviour (i.e., the 'rules' governing their behaviour are unknown), that is not the subject of our attention, also influences our impression more than the rewards we accrue from their behaviour, though learning rates here were equal for both types of behaviour. Our study further demonstrates that others' intentions also outweigh personal rewards when we make future decisions.

Players were free to make their decisions independently of the gaze cues, yet they chose to use the fully non-veridical faces' cues to guide their decisions during the game. Consequently, they learned that the cue provider was consistent, allowing them to always get the answer right and earn more money. This should mean that participants had no reason to think the non-veridical face would behave differently in the future and that they would be a good partner to play the card game with, even if the participant does not like them personally. However, players did not pick the non-veridical face for future rounds of the card game and instead selected unfamiliar individuals whose behaviour during the game is entirely unknown. This is noteworthy because impression formation is believed to guide decisions in order to help us to achieve our own goals (De Bruin & Van Lange, 1999; Fiske, 1992) and the best predictor of a person's future behaviour is their past behaviour (Axelrod & Hamilton, 1981; King-Casas et al., 2005). Together, this should mean that people – out of self-interest – will select the candidate who helps them to achieve their purely financial goal,

irrespective of the social meaning behind the candidate's behaviour. Unlike the ultimatum game, where players pay a cost but also gain the ability to punish the other, in this scenario, only the participant themselves suffers as a result of their decision. This means that others' (only implied) negative social intent is sufficient for us to deprive ourselves of a (highly likely) positive economic outcome. The consistency of participants' choices across post-game measures suggests they were relying on social valence alone and did not wish to encounter non-veridical faces again, under any circumstances, no matter the economic costs.

Across both experiments, the majority of players did learn the face-outcome contingency, with most participants reaching ceiling-level performance in the last block of the game. Intriguingly, a solid minority of players (~12%), in both experiments, failed to use/detect the relationship between the faces' cues and the answers, achieving scores consistent with using only the card values to make game decisions. This is particularly surprisingly in Experiment 2, where participants had nearly 200 trials on which to notice that the gaze cues contained relevant information! Those who did not learn the face-answer contingencies also showed no evidence of an effect of condition on liking or trust (though in such a small group we don't have the power to rule out small effects). This suggests that experience with the faces and learning/using the gaze contingencies is necessary for impression formation under these circumstances. This finding is somewhat in contrast to studies that have used more traditional gaze-cueing paradigms, where the faces' gaze-cues impact our goal related behaviour by automatically orienting our attention and pursuantly affecting our performance (Driver et al., 1999; Friesen & Kingstone, 1998). In such tasks, (e.g., Bayliss & Tipper, 2006; Rogers et al., 2014; Chapter 2: Inconsistent Behaviour), even though participants do not explicitly learn the gaze-cue contingencies, most do form incidental social impressions. In our design, participants' ultimate behaviour is only affected if they choose to use the gaze-cue when making their decision. As such, our outcome is automatically affected in a traditional paradigm and only voluntarily affected in the present task. It seems plausible, therefore, that forming an impression of a person from their gaze-cue behaviour requires more than just the visual processing of a cue's spatial relationship with a target, we need to also be behaviourally affected by the other person at the level of our outcome. Indeed, during debrief, around half of those who did not learn, and did not form strong impression of the faces, declared that they had noticed the faces' eyes

were moving, but had not spotted the contingency. Despite the inherent differences between the two tasks, that both 'traditional' gaze-cuing tasks where impressions are formed implicitly and our card-game task where impressions are formed more explicitly result in similar impressions suggests that the same underlying social processing mechanisms are involved.

Limitations and future directions

In both Experiments, there were distinct groups of people who did not use the faces' cues to guide their decisions in the card game. While this seems quite remarkable, the design may have contributed to a lack of learning for some people. The faces were prominent (i.e., in the centre of the screen), but they were not central to the task and the face up card was the immediate focus of player's visual attention; it provided the information from which a decision could be made. Consequently, participants may have experienced no or a minimal gaze-cueing effect (i.e., no effect on visual attention) (Langton & Bruce, 1999), reducing the salience of cue providers' behaviours. It is possible, then, that a small minority of people simply did not engage with the faces, as they had other information available on which to base their decision. This possibility could be confirmed with eye-tracking. The high number of incorrect 'improbable' trials that such participants would experience should have encouraged them to look for other information to help them do better, so ignoring the faces cues under these circumstances is particularly surprising. Indeed, during data collection (where the researcher could deduce whether the participant had 'clocked' the contingency when paying their winnings, which were linked to accuracy) some participants were surprised to learn that the faces had been behaving with a strategy. Future studies should explore who this group of people is and what, if anything, they have in common.

As already noted, following the change in faces' behaviour, most people accurately and rapidly updated their own behaviour. This was likely facilitated by the faces completely reversing the valence of their behaviour, which made their new contingency easy to learn. If new behaviour was non-predictive (i.e., 50% veridical, 50% non-veridical), there would be no new rule to learn (as the person will cooperate as often as they will deceive), which prior work suggests can interfere with learning (Delgado et al., 2005; Fareri et al., 2012). It would be worthwhile, therefore, to

explore how behaviour during the game and pursuant social decisions are affected when faces change to provide non-predictive behaviour. Firstly, impressions of the wholly non-veridical face might be boosted, as others' non-predictive behaviour would be highly unreliable (a negative social trait) and greatly interfere with performance. Additionally, it may allow for any differences in learning rates between veridical and non-veridical cues providers to emerge, as players would likely focus initially on impression-consistent behaviours (i.e., trials where the original cue-answer contingency is present) (Fareri et al., 2012), which could mean that behaviour is updated differently in relation to formerly veridical and non-veridical cue providers.

When making their decision, participants did forsake potential monetary gains and/or paid costs to either punish or avoid non-veridical faces. However, participants neither risked nor declined real money in the trust (Berg et al., 1995) and ultimatum (Güth et al., 1982) games. Though studies have shown that investment levels are comparable whether hypothesised or incentivised versions are used (e.g., Thielmann et al., 2016), it is possible that the kind of decisions that participants made in these tasks may have been more driven by monetary gain if real money was in play. Results, however, show that people did distinguish between how they would interact with each face, supporting the notion that while the magnitude of effects found here may be different when real money is risked or offered, there is no reason to believe that differences would not, nonetheless, still be present.

Finally, only the non-veridical face's 'negative' cues were incongruent with the perceiver's positive outcome. Put another way, a veridical face's 'positive' cues were congruent with the perceiver's outcome, meaning that the study did not test how positive behaviour that generates negative consequences is perceived. One way that this could be done is by using a traditional gaze-cueing design, only increasing the temporal gap between the face's cue and the target's appearance. In that situation, an inhibition of return is commonly experienced, whereby response times are actually longer for valid (helpful) gaze-cues, as the perceiver has already visually disengaged from the location by the time the target appears (Klein, 2000). By associating worse performance with a valid cue provider, it would be possible to see whether the behaviour is perceived positively or negatively. People are sensitive to the intention behind another's behaviour, such that good

'surface level' behaviours are not interpreted favourably if the motives behind them are negative (Malle & Knobe, 1997). In this scenario, where the other's behaviour is not referenced or explained, it is possible that by hindering performance, participants will infer a negative intention. Alternatively, because the target does eventually appear in the cued location, the behaviour may still be processed as having a positive intent even if performance is negatively impacted.

Conclusion

To date, impressions based on the social meaning behind other people's behaviour have often been conflated with impressions based on the perceiver's outcome. Our results clearly show that incidental impressions are driven by our perception of the other person's behaviour, not our own outcome. Indeed, our decisions appear to be so socially motivated that we are prepared to forsake personal benefits and take unnecessary risks on strangers to avoid someone we don't like. While we will update our impression if someone changes how they behave, a change (for the better or worse) is not enough to fully reverse a previous impression, at least under these circumstances. Taken together, these experiments demonstrate that the social nature of a person's behaviour, not the actual monetary reward, is prioritised during social reward learning.

Chapter 5: General Discussion

As social animals, we surround ourselves with social information, even when engaged in other activities. As a result, forming impressions incidentally is likely ubiquitous in our everyday life. However, the process and mechanisms of how we do so are relatively understudied. Additionally, observing and reconciling inconsistent social information is also something we experience on a regular basis, yet research has only relatively recently begun to address how impressions are updated empirically (Mende-Siedlecki, et al., 2013). The current work contributes to both of these areas of research by addressing open questions around how experiencing others' inconsistent behaviour over time influences incidental impression formation and social decision-making. Together, the studies presented here provide for a clearer and more precise picture of how we process and incorporate others' task-irrelevant, yet socially informative, gaze behaviour.

Incidental Impressions

Chapter 2 demonstrated that people can still form incidental impressions of others when their behaviour is both less than fully categorical and inconsistent over time. Given that the context in which the other is experienced includes multiple others, minimal social information and the need to attend to an unrelated task, this is quite impressive, not least because it demonstrates how sensitive and tuned to others' social signals we are (Frith & Frith, 2007). In Chapter 3, despite the learning environment and total lack of a social goal, older adolescents show a similar trust learning effect as adults. This demonstrates that by around 14 years of age, children are able to track, process, and associate others' gaze-based behaviour in a social manner and form impressions incidentally. Albeit underpowered results suggest younger adolescents may not yet make the connection between another's gaze behaviour and the social meaning behind it, at least when the behaviour is neither explained nor goal related. There was tentative evidence, however, that something social about the faces was being processed by the younger group; those who always or ultimately displayed unhelpful cues were better remembered. This has also been found with adult participants (Bayliss & Tipper, 2006), where negative information is better encoded and remembered (Bell & Buchner, 2012), suggesting that a precursor to forming a negative impression

about the person behind the gaze is encoding something about the face as negative. In Chapter 4, we expanded this focus to look more closely at what about the face was driving negative impressions; its behaviour or its effect on our outcome. Results robustly support a behaviour-based mechanism, whereby even when another's only inferred negative behaviour has a positive outcome, we still form a strong negative impression about the person, which in turn influences both social and financial decisions.

Inconsistent Information

We know all too well how unreliable or prone to change people can be, yet the extent to which we update initial impressions and the circumstances under which we do so are still contested; we do not fully understand how we process and incorporate inconsistencies in behaviour as we form our impressions. Indeed, this thesis was first motivated by the realisation that relatively little prior research had explored the impact of inconsistent behaviour during direct encounters, and those studies that had done so had often use highly evaluative stimuli, such as overt trust-based behaviour (i.e., reciprocity) in a rule-based interaction. The work in this dissertation, therefore, is principally concerned with how we incorporate changes in behaviour over time, when the behaviour in question is not the focus of our attention nor immediately trait diagnostic nor obviously intentional. Despite the differences in how person knowledge is acquired, do we update our impressions when we receive unreferenced social information in the same way as we do when social information is explicitly explained?

Across all studies (with the exception of the younger group of adolescents), when other people changed how they behaved, impressions were updated in line with the direction of their change (e.g., a decrease in helping behaviour resulted in distrust). This demonstrates that despite the fact that perceivers had no impression formation goal and the other person's behaviour was not relevant to the task, inconsistent behavioural information was attended to and impressions were updated. These findings, consistent across all three empirical chapters, are inconsistent with a primacy based explanation (Nisbett & Ross, 1980) (i.e., initial behaviours did not predict final impressions). Results also does not support the idea that inconsistent information receives less attention when we have no person-related goal (Ames & Fiske, 2013), which might have been expected given the

circumstances of the included tasks. However, in prior work that has found initial impressions are not updated, the other person's behaviour was not experienced directly, whereas in our work, the other person did (or could) affect our outcome, irrespective of their consistency. A change to a person's behaviour in the experiments included in this dissertation either had the concomitant effect of also changing our future outcome (Chapters 2,3) or required us to change our behaviour in order to maintain our performance (Chapter 4). As such, present findings suggest that inconsistent information is attended to and incorporated into overall impressions when our own outcome is influenced by the person's behaviour, regardless of whether or not we have an impression or person-related goal.

In the current studies, final impressions did not solely reflect others' recent behaviour, as both the valence of initial impressions and the direction of change influenced final impressions, depending on the circumstances. While this finding is consistent across experiments included here, it is somewhat different to the few other studies that have looked at inconsistent behaviour and impression updating using direct experience (in the trust game). When someone either behaves inconsistently with our prior expectation (based on indirect information) (e.g., King-Casas et al., 2005; Maurer et al., 2018; Zarolia et al., 2017), or changes how they behave towards us during a wholly direct interaction (e.g., Campellone & Kring, 2013), impressions tend to follow the valence of the person's most recent direct behaviour only. In our experiments, however, when gaze-cues were used and all behaviours were experienced directly, only recently positive behaviour did not result in impressions that matched those resulting from consistently positive behaviour, suggesting that initial behaviour likely continued to influence impressions. Results were mixed when recent behaviour was negative. In the gaze-cueing experiments (Chapters 2 & 3), final impressions were driven by recent negative behaviour alone, whereas in the card game (Chapter 4), consistently negative behaviour was perceived more negatively than only recently negative behaviour, again implying that initial information retained some influence. This inconsistency is likely the result of the difference in learning across the two paradigms. In the gaze-cueing tasks, learning is implicit while in the card-game, learning about the faces is explicit. In addition, the gaze-cueing experiments were powered for medium effect sizes and were not powered to detect small effects if they do exist. Thus, we cannot rule out the possibility that initial negative information continued to influence final impressions, even when social learning is implicit.

Although the explicit learning experience is comparable between the card game and the trust game, there are likely many reasons why inconsistent information may influence impressions differently in the two situations (Campellone & Kring, 2013). For instance, as we learn about someone who is initially 'good' in the card game, we improve our performance when we follow their cues, which are salient because we have to make a decision when we do not know the answer. In this way, they help us to learn the task and, eventually, allow us to stop attending to the task and rely only on their cues, making them genuinely helpful. Additionally, in the circumstances of the game, when they change their behaviour we can quickly recover our performance. Together, this may mean we do not fully over-write our once positive impression. In the trust game, however, the other person's behaviour determines our outcome, and if they change from sharing to stealing, they clearly signify a change in intention and we can no longer make any money with them, which may explain why we fully revise our impression of them. Conversely, someone who is initially 'bad' and does not directly help us at the start of the card game, does not actually impair our performance, even if they aren't trying to help. Further, when they change for the better and start to give us veridical cues, our performance is unaffected but they are now providing directly helpful information. Here, our impression becomes more positive, but the behavioural switch is not enough to fully recover from a bad start. In the trust game, on the other hand, we could not make money from someone who did not initially share, but once they change their behaviour and start consistently sharing, we start to earn money and can infer that they have changed their ways, which may be enough to over-ride a once negative impression. It is, however, surprising that consistently positive behaviour in the trust game does not garner a more favourable impression than someone who spent the first half of the encounter not sharing with us. Ultimately, our findings need to be replicated and more work is needed in this area, both to directly compare contextually different direct learning experiences and with a wider set of contexts and behavioural manipulations, so the mechanisms at work can be more precisely understood.

One of the most intriguing results in the dissertation is that, in the card game, pairs of conditions that ended by displaying the same behaviour were not perceived as the same. Differences between those final ratings could relate to the stable condition's consistency (e.g., always good/bad is the

most/least valued) or the inconsistent condition's change[13] (e.g., a once bad impression is not fully reversed), and it is not possible to know which explanation is the correct one using the current design. This is partially because the consistently helpful face was also the most helpful face, making it difficult to know if consistency is valued in and of itself, rather than simply the greatest amount of helpful behaviour. This possibility could be addressed by reducing the number of helpful cues the consistent face provides, such that they would be associated with some 'wrong' trials, and have the face that increases its number of veridical cues end at *higher* than the consistent face, resulting in them providing the same amount of helping behaviour overall. While a difference was found between a face that gave inconsistent but mostly helpful cues across the entire task and a face that changed behaviour from mostly unhelpful at the start to wholly helpful at the end in the traditional gaze-cueing task (Chapter 2, Experiment 1), this might not be true for the card game. In the card game, errors (and therefore, a loss of money) might be more salient, such that it is high helping behaviour and not consistency, per se, which people most value. However, this would not address the issue of updating an impression in response to inconsistent behaviour, which is a different question and one which the design did consider.

When updating incidental impressions that are learned from others' direct behavioural cues, our results suggest that positive impressions are downgraded (c. -£4 investment) more significantly in response to a negative change than negative impressions are upgraded (c. +£3 investment) with a positive change (Chapter 4, Experiment 2). While this may not match previous findings for other direct encounters, it does mirror the pattern of findings found when information about an individual's past (im)moral behaviours is given indirectly (Mende-Siedlecki, et al., 2013; Reeder & Coovert, 1986). The correspondence between our findings and those of this literature is particularly interesting given that the two scenarios provide both very different ways of learning about the other and different implications for the perceiver (i.e., there are no direct consequences for the perceiver when information is given indirectly). Together these two sets of findings do suggest that positive and negative impressions are updated asymmetrically, regardless of how they are acquired. In the card game, initial impressions were highly opposed to each other, and it makes

[13] It could also be due to more simple averaging, as the consistent person also helps the most overall, however, the two faces who averaged the same helping behaviour were not perceived to be the same, suggesting something more complicated.

intuitive sense that changes to the two might be processed and understood differently. If a helpful person suddenly switches their behaviour, we might infer they no longer wish to help us, indeed, they might wish to act against us now, so we severely downgrade our impression in response, lest they mean us harm. Conversely, when someone earns an initial negative impression, it may take much longer for us to trust them and update our impression, lest they then change back to their former behaviour. This interpretation is partially supported by a study in which trust impressions (measured through direct trust behaviour) were quickly updated in response to negative direct behaviour when the initial positive impression was conveyed indirectly, whereas people updated their behaviour more slowly in response to an actually trustworthy person who had been described as untrustworthy (Zarolia et al., 2017, exp.2). Further work will be required, however, to truly elucidate the specific commonalities and differences in impression updating across different modalities and environments.

One compelling future experiment might focus on learning environments. There has been a growing call in recent years for social cognition research to examine real-time social interactions from a first-person perspective ('online' cognition where the perceiver is affected by the other) rather than through third-party observations ('offline' cognition where a perceiver merely spectates) (De Jaegher & Di Paolo, 2007; Pfeiffer et al., 2013; Redcay & Schilbach, 2019; Schilbach et al., 2013). While the current studies' participants could be considered to be engaged in first-person interactions, and the design mitigated many of the issues associated with studying 'offline' social cognition, the task itself was still relatively contrived and rather different to how we are used to encountering new people. One way to counter this problem is to use more naturalistic stimuli, where other people are embedded in a richer, more natural context (Redcay & Moraczewski, 2020). The benefit of paradigms such as the trust game or the card game presented here is that the other person's behaviour can be mathematically manipulated and positive and negative behaviours can be essentially directly opposing, allowing for a high level of experimental control. Such mathematical control may suffer in trying to employ entirely natural scenes.

However, given the potential benefits to be gained from data that more accurately captures how we process social information in our everyday lives, it is important to consider how experimental designs can move towards more genuine environments. For example, people's (manipulated) gaze-

cue behaviour could be experienced within free-viewing dynamic scenes, where other people are also present and performing their own behaviours, such that rather than relying on the gaze-cueing effect to orient attention, cues will need to be detected in a busy scene. In such scenarios, the distractions are likely to be more similar to those experienced in the real world and the other person's behaviour would have to be detected, tracked, and interpreted within a complex scene. This additional information, which would comprise both meaningful signals and noise, could interact with how we process any one person's helpful or unhelpful cues. For instance, if we are in a free-viewing 'room' based scenario looking for, say, a list of objects, and a single person present displays misleading cues, might we perceive them more or less negatively than if the room had multiple people doing unrelated things, or if another person also present provides helpful cues? In other words, as a scene becomes more complex or behaviours become more/less salient, how might it interact with our detection and evaluation of meaningful social cues? The current studies suggest that, once detected, perceivers would attribute intent to the behaviours and form impressions based on the inferred valence of the behaviour; someone providing misleading cues would not be trusted. However, it is not known how differences within and across 'scenes' might interact with the strength of, and differences between, impressions formed. Further, it is not known how our emotional responses might be affected under these different circumstances (Hietanen, 2018). It is possible that social complexity, or an individual's distinctiveness from others in a scenario, interacts with how we respond to their behaviour. In Chapter 2, the final two experiments considered the influence of the social group on impressions of individuals. Though results were not unequivocal, they do suggest that this is an area worthy of future investigation. Additionally, in a free-viewing visual search scenario, if a previously helpful person changes to deceiving us, we would discover their switch immediately (as we would quickly look to the now wrong location). Conversely, it may take us longer to learn about a deceitful person's change to being helpful, as we would likely not look to their cued location for some time. It seems plausible, given the findings of Chapter 4, Experiment 2, that initial impressions would be updated unequally, however, potential differences in the perceiver's affective and behavioural experience may alter how inconsistent behaviour is processed. There are many ways we can experience people's behaviours and future impression formation research has many exciting new avenues to explore.

Conclusion

The current work adds to our knowledge of incidental impression formation by making three novel contributions. First, Chapter 2 showed that people are sensitive to others' nuanced social gaze behaviour and that incidental impressions are updated in response to inconsistent information. Where previous studies had only explored wholly consistent behaviour, the current studies establish that we are capable of processing even more complex behaviour in an implicit learning environment, revealing an impressive feat of social cognition. When new, inconsistent behaviour is experienced, incidental impressions are updated, however findings suggest that the degree of updating may differ for initially positive vs. negative impressions, as the valence of the change matters. Second, Chapter 3 cautiously finds that by 14 years of age, children can also detect and track others' complex gaze-based behaviour while younger adolescents may not. Finally, Chapter 4 demonstrated that it is the meaning behind a person's behaviour and not our own outcome which drives our social impressions and decisions. Together, they show that impressions can be formed and updated incidentally and that these impressions can influence our behaviour in much the same way as those formed more intentionally.

Appendices

Appendix A: Chapter 2: Preparatory work

Overview

Before exploring the influence of non-wholly predictive and inconsistent gaze-cue profiles on social decisions, it was important to establish the efficacy of the design, as there are four notable differences between this and previous experiments (e.g., Rogers et al., 2014).

Firstly, a target-location version of the gaze-cuing task was used, rather than an object-categorisation task as is commonly used in similar trust learning studies. In the object categorisation version, an image appears on the left or the right of the screen, and participants must identify the category to which the target belongs (for example, kitchen vs garage tools). In these designs, although the eyes provide a cue regarding where a target will appear, once the target is seen the participant must still additionally identify the correct category. Therefore, the judgement is orthogonal to the location of the target. Given the complexity of individual target's gaze-cue profiles in the present suite of experiments, it was decided that reducing the demands of the task would mitigate the effects of adjusting the validities and potentially facilitate the encoding of faces' behaviours; as they simply look towards (or away from) where the target later appears. In this version, although the faces are still task irrelevant (i.e., not necessary to complete the task), they do help (or hinder) participants' performance by identifying (or lying about) the target's future location.

Secondly, the number of faces participants experienced here was far fewer than are usually included in other trust learning studies, where targets can range from 16 (e.g., Rogers et al., 2014; Strachan & Tipper, 2017) to 40 (e.g., Bayliss & Tipper, 2006) in number. Again, in consideration of the fact that the gaze-cue contingencies were to be no longer wholly predictive, and therefore likely more difficult to encode, it was thought that the number of individual faces that participants were required to track should be substantially reduced. Consequently, as the tasks were not

dissimilar in length, the number of times each face was encountered was greatly increased. A third difference is that an incentive compatible design was not used. That is to say, the outcome of their decisions was not revealed to participants and they neither gained or lost money. Thus, we needed to check that participants would still invest more with faces who displayed valid versus invalid gaze cues, even when the monetary incentivisation element of the design was omitted. Finally, it was important to ensure that people are indeed capable of detecting (albeit implicitly) a range of nuanced gaze-cue behaviours beyond simply 100% valid or 100% invalid, as designs used included validity profiles that both varied among faces and over time[14].

To address these issues, both a conceptual replication and two pilot experiments were conducted. The conceptual replication experiment used wholly valid and invalid conditions in a similar manner to previous research (Rogers et al., 2014) but, as described above, both the task used was a target-location task and investments in the trust game carried no real risk. In addition, the first pilot design included faces displaying either 100%, 80%, 60%, 40%, 20% or 0% valid gaze-cues, while the second included 100%, 75%, 50%, 25% and 0% valid gaze-cues. These checked whether people were (i) able to track a range of behaviours and (ii) confirmed that the main effect of validity is still present when the design/environment is more complex. That is to say, that faces displaying mostly valid cues are trusted more than faces displaying mostly invalid cues. The format of trials was the same as for all other experiments reported in this chapter (see 'General Methods').

Method

Participants

The same eighteen volunteers (female = 10, mean age = 22.6, SD = 5.5) participated the in both the replication and first pilot experiments in return for course credit. The order of task completion was counterbalanced across participants. There were software issues during both tasks for one participant, resulting in no useable data being collected. This resulted in 17 participants being included in both the replication and pilot analyses. Ethical approval was granted by Bangor University's School of Psychology Ethics Committee.

[14] Results of the gaze-cueing studies (i.e., accuracy rates and response times) are not reported.

Procedure

Prior to completing the gaze cueing tasks, participants were introduced to the trust game and each participant made decisions about how much they would invest with each of two unfamiliar faces; one male, one female. Participants then completed either the replication or the pilot gaze-cueing task, after which they were shown each face displayed during the task and asked how much they would like to invest with them. The order in which faces were shown was randomised across participants. To 'reset' participants, and hopefully prevent any carryover effects, some questionnaires, taking around 10 minutes to complete, were administered between the two versions of the task. Which task was completed first was counterbalanced across participants and two sets of faces were used so that participants saw a different group of faces for the replication and the pilot experiments.

A1: Replication Study

Design

Six male faces were included, three providing wholly valid gaze cues and three wholly invalid gaze cues. To avoid effects of face, condition and face pairings were randomized across participants. Additionally, to avoid possible effects of race or sex, all conditions were represented by white males. Three blocks were included and the validity of the faces remained constant across all blocks. There were 48 trials per block meaning that there were 144 trials in total and each face was seen 24 times.

Results

Visual inspection of the data is highly suggestive of a robust effect of validity, with wholly valid faces seeming to be trusted more than wholly invalid faces (see **Figure A1**, left). Investments made with the three valid and three invalid faces were collapsed and a paired samples t-test was used to compare average investments made with wholly valid (mean £4.57, SD 1.94) compared to wholly invalid faces (mean £3.39, SD 1.56). This confirmed that participants invested more with the valid faces than the invalid faces $t(16) = 2.74$, $p = .015$, CI[0.27, 2.09], dz =.67 (see **Figure A1**, right).

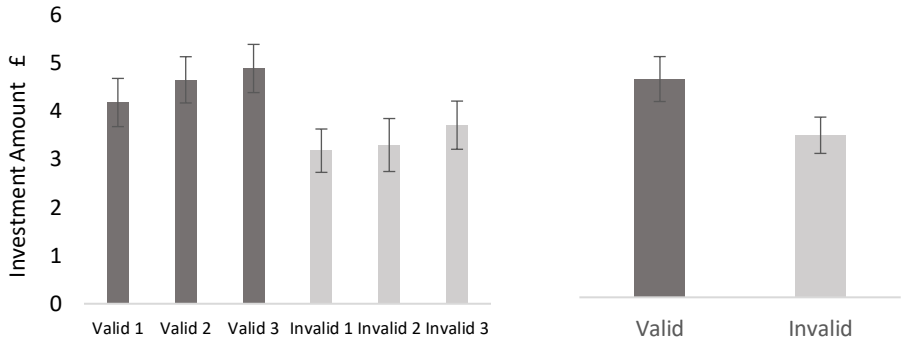


Figure A1: Investments made with the wholly valid and invalid faces at the individual face level (left) and average condition level (right) in the Replication Experiment. Bars represent SEM.

A2: Pilot 1

Design

Six different male faces were included, each providing a different gaze-cue profile that remained constant across the task. Faces provided either 100%, 80%, 60%, 40%, 20% or 0% valid gaze cues. The face associated with each condition was, again, varied across participants.

Three blocks, each with 60 trials, were displayed resulting in 180 trials in total and each face being seen 30 times. The additional number of trials and presentations per face was deliberately increased in consideration of the fact that the difference between some of the conditions was subtle (e.g., 100% compared to 80% helping behaviour) and the combination of conditions was more complex than that employed in the replication experiment.

Results

While the distribution of investments is not a perfect fit for validity, the pattern of results does provide a good indication that participants are capable of processing validities beyond wholly valid and wholly invalid (**Figure A2**, left). A one-way repeated measures Anova showed that there was a large difference in investments made between the conditions f(3.1,49.7)=5.18, p=.003, np^2=.24.

Follow up Sidak adjusted pairwise comparisons showed that the 100% valid and 80% valid faces received similar amounts to each other (p>.99, mean difference £0.53, CI[-1.66, 2.72]), though visualisation of the data does suggest that the wholly helpful condition attracted slightly higher

investments (Figure 2 – left). There was also no meaningful difference in investments made between the 20% valid and wholly deceitful faces (p>.99, mean difference -£0.41, CI[-2.10, 1.28]). Interestingly, despite the inclusion of other conditions, the magnitude of the difference between the wholly valid (100%) and wholly invalid (0%) conditions for this pilot (t(16)=2.62, *p*=.018, mean difference £1.76, CI[0.34, 3.19], dz=.64), was highly comparable to that of the replication experiment (dz=.67). Additionally, despite their occasional unhelpful behaviour, the 80% valid face was also trusted a lot more (£1.65) than the, occasionally helpful, 20% valid face, t(16)=2.48, *p*=.025, dz= .60.

A3: Pilot 2

Participants & Design

A further 24 volunteers (22 female, mean age 22.1, SD 4.5) were recruited. The gaze-cueing and investment tasks were identical to Pilot 1, the only changes made related to the validity of the profiles displayed by each of the six faces. Here, the difference between the faces was increased slightly (from 20 to 25%) with faces displaying either 100, 75, 50, 25 or 0% validity. The 50% condition was portrayed by two faces. This additional pilot served to confirm the findings of the first Pilot and introduce validity profiles more similar to those chosen for the final designs, which were brought about based on the mathematical restrictions of the designs.

Results

As can be seen from **Figure A2** (right), the conditions appeared to be perceived quite differently, which was confirmed by a simple 1*6 repeated measures Anova, f(3.2,74.5)=5.48, p<.001, np^2=.19. Again, there was a large difference between how much the wholly valid and wholly invalid faces received, t(23)=3.89, *p*=.001, dz=0.79, and the 75% valid face was also a little more (£1.17) than the 25% valid face, t(23)=1.81, *p*=.083, dz= .37.

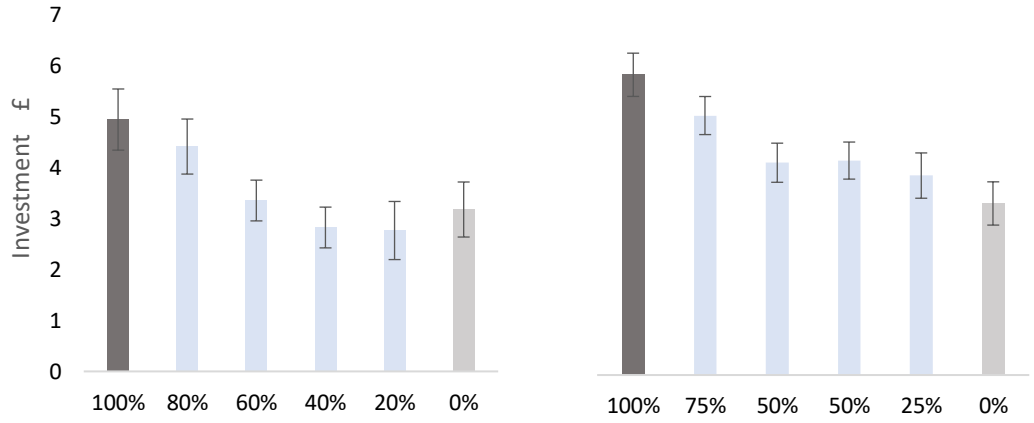


Figure A2: Investments made with each condition for Pilot 1 (left) and Pilot 2 (right). Bars represent SEM.

Discussion

The results of both the replication and pilot experiments provide some assurance that the design is capable of detecting the effects of gaze-cue validity on trust decisions. Despite not using an incentive compatible design, effect sizes of validity on investment decisions were similar to that found by Rogers and colleagues (2014). This suggests that people still approach the decision in a similar manner, even when they do not actually win or lose money. While it is not known whether the same effect sizes would be obtained using this design if participants did stand to win or lose actual money, there is reasonable evidence that people are considering the risk associated with their decisions, as the different profiles elicited differing amounts. Interestingly, the size of the difference between wholly valid and wholly invalid faces was almost the same in the replication experiment (dz=0.67) as it was in Pilot 1 (dz=0.64) and Pilot 2 (dz=0.79). That this remains the same across three different tasks where faces display very different behavioural patterns bodes well for the main experiments.

Appendix B: Chapter 2: Niceness ratings and Gaze-cueing effects

B1: Niceness Ratings

Experiment 1

Effects of gaze-cue validity on niceness were more subtle than for trust based decisions (see **Figure B1** – Top, left). Faces look to have been perceived similarly, with the exception of the unhelpful decrease face who was least liked. Importantly, other faces all received ratings close to 'neutral' on the scale, suggesting helpful behaviours did not increase likeability.

A two-way repeated measures analysis of variance found a medium effect of overall valence, f(1,50)=4.19, p=.046, np^2=.08, with the least helpful faces (those averaging 25% helping behaviour; M 3.6, SD 1.2) being rated as less nice than the most helpful faces (averaging 75% helping behaviour; M 3.9, SD 0.9). This effect looks to be driven by the decrease condition in the unhelpful group being the least liked. The only (Sidak adjusted) contrast reaching nearing significance was between the stable-helpful and the decrease-unhelpful conditions (p=.054, CI[-1.5, 0.01]). There was some effect of order, f(2,100)=2.37, p=.099, np^2=.05, though this effect did not reach significance, and there was no interaction effect present f(2,102)=0.07, p=.928, np^2<.01.

Experiment 2

Paired samples t-tests showed that there were no differences between the two decrease (t(50)=1.30, p=.200, CI[-0.17, 0.80], dz=.18), neutral (t(50)=1.02, p=.313, CI[-0.29, 0.87], dz=.14) or increase (t(50)=1.01, p=.317, CI[-0.82, 0.27], dz=.14) faces, and they were collapsed to form three conditions. The conditions appeared to be rated similarly for niceness (see **Figure B1** – Top, right). This was confirmed by a one-way, repeated measures Anova f(1.8,89) =0.92, p =.391, np^2 =.02.

Experiment 3

A 1*5 repeated measures Anova found a medium to large effect of condition f(4,212)=8.01, p<.001, np^2=.13. Despite both conditions providing only helpful behaviour in the final two blocks, the condition that always helped throughout the task (mean 4.6, SD 1.6) was liked slightly more than the face that started by not helping at all and then switched to always helping (mean 4.0, SD

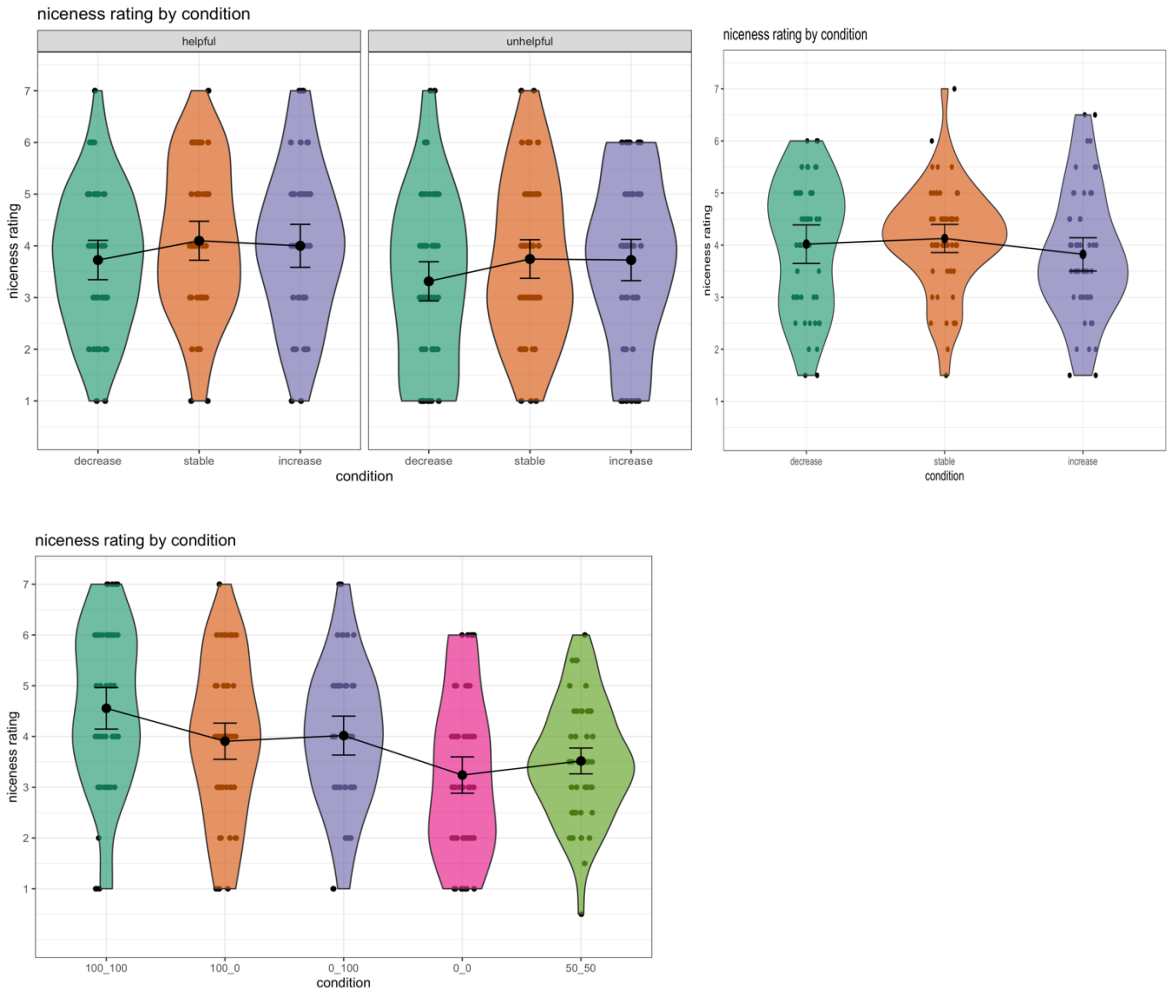


***Figure B1**: Niceness ratings for Experiment 1 (top, left), Experiment 2 (top, right) and Experiment 3 (bottom). Bars represent 95% CI.*

1.4) (p=.056, CI[-0.01,1.09], dz=.22). For the pair that ended by providing no valid cues, the face that never helped (mean 3.2, SD 1.6) was liked less than the face that initially always helped, but later never helped (mean 3.9, SD 1.6) (p=.010, CI[-1.16,- 0.17], dz=.37) (see **Figure B1** – Bottom). Further, the control condition, which helped 50% of the time across the whole task (mean 3.5, SD 1.1), was liked less than both the face that decreased its helping behaviour over time such that it never helped by the end (p=.052, CI[-0.78,0.00], dz=.27), and the face that started by not helping but always helped by the end (p=.037, CI[-0.97,-0.03], dz=.29), which does not support either recency or averaging effects. Interestingly, the faces that changed their helping behaviour over time, one increasing to always helping and the other decreasing to never helping (but who averaged the same amount of helping behaviour), were liked the same amount (p=.681, CI[-0.65,0.43],

dz=.06). For the faces that maintained their level of helping behaviour over time, the face that always helped was rated much more favourably than the control (neutral) face (p<.001, CI[0.56,1.51], dz=.59). The face the never helped, however, was viewed to be almost the same as the control face, (*p*=.178, CI[-0.69,0.13], dz=.19).

Experiment 4

Despite being encountered in very different environments, there was no difference in how the faces were rated in the improving (mean 4.1, SD 1.2) compared to the worsening (mean 4.1, SD 1.3) environment f(1,51)=0.01, p =.929, np^2 <.01 (see **Figure B2**, left). There was a difference between how the faces themselves were rated, however, with the face that consistently provided 75% helpful gaze cues (mean 4.5, SD 1.2) being rated as nicer than the face that consistently provided only 25% helpful gaze cues (mean 3.7, SD 1.2), f(1,51) = 12.80, p <.001, np^2 =.20. There was no interaction present f(1,53) =1.61, p =.211, np^2 =.03.

Experiment 5

Each inconsistent face, which changed its level of helping behaviour over time, appeared to be rated similarly for niceness, both across environments and as compared to each other (see **Figure B2**, right). This was confirmed by a repeated measures 2*2 Anova which found no effect of environment f(1,49) =1.20, p =.278, np^2 =.02, no effect of person f(1,49) =2.26, p =.139, np^2 =.04 and no interaction f(1,40) =.06, *p*=.805, np^2 <.01.

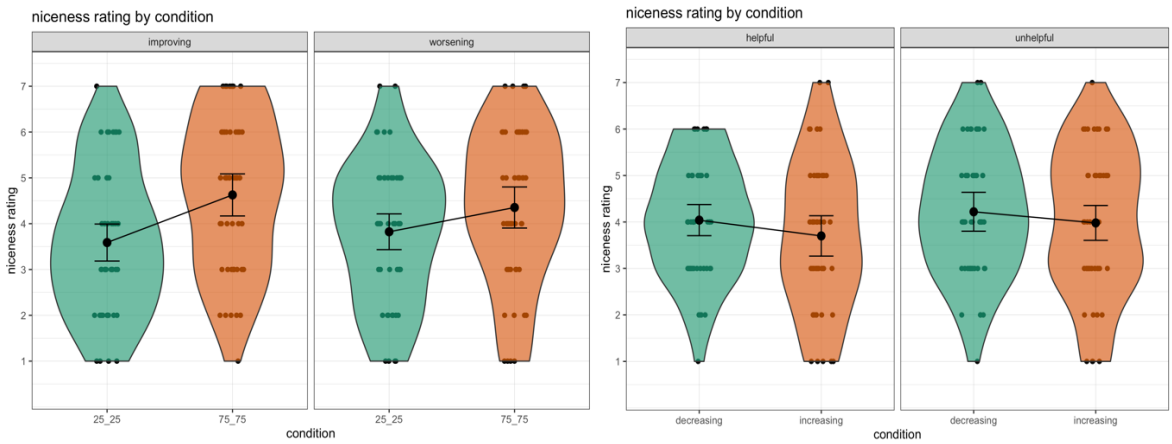


***Figure B2**: Niceness ratings for the target, consistent faces the changing environments in Experiment 4 (left) and for the inconsistent faces in the stable environments in Experiment 5 (right). Bars represent 95% CI.*

B2: Gaze-cueing task results

Experiment 1

Accuracy

Performance on the task was very good (accuracy 95%, trials excluded as too fast or too slow 4%, excluded as incorrect 1%). A two-way repeated measures Anova with factors Validity (2 levels: Valid and Invalid) and Block number (4 levels) found no obvious effect of validity on accuracy f(1,51)=2.47, p=.123, np^2=.05, with invalid trials (mean 95%, SD 3.0%) and valid trials (mean 95%, SD 3.6%) being similar. There was also no effect of Block number on accuracy f(2,80)=0.68, p=.565, np^2=.01, meaning accuracy levels were essentially the same over time. There was also no interaction present f(3,153)=0.41, p=.749, np^2=.01.

Reaction times

A two-way repeated measures Anova, again with factors Validity (2 levels: Valid and Invalid) and Block number (4 levels) revealed a large main effect of validity on reaction times f(1,51)=12.62, p=.001, np^2=.20. As expected in a gaze-cueing task, participants responded more quickly to valid trials (mean 344ms, SD 61ms) than invalid trials (mean 355ms, SD 70ms). There was a medium effect of Block f(1.8,90)=2.87, p=.069, np^2=.05. As anticipated, reaction times were highest in the first block (mean 357ms, SD 82ms), but then decreased in the second block (mean 341ms, SD 63) as participants practiced the task (p=.013, CI[3.5, 28.4). Somewhat unexpectedly, reaction times were then a little higher in the third block (mean 348, SD 66ms) than the second block (p=.071, CI[-0.6, 14.5]). This may reflect a sensitivity to the changes in four faces' gaze-cue validity that occurred between blocks two and three. Reaction times had stabilised by block four (mean 352ms, SD 63ms), with no notable difference between the final two blocks (p=.363, CI[-10.5, 3.9]). There was also a medium sized interaction effect f(2.5,129)=5.18, p=.002, np^2=.09, as the increase in reaction time in Block 3 from Block 2 occurred for the valid trials (p=.019, CI[2.45,25.70], dz=.34) but not in the invalid trials (p=.956, CI[-5.44,5.75], d<.01) (see **Figure B3**– left).

Experiment 2

Accuracy

Participants' overall accuracy was high (accuracy 95%, excluded too fast/slow 4%, incorrect 1%). Two-way repeated measures Anovas with factors Validity (2 levels: Valid and Invalid) and Block

number (4 levels) were performed for both accuracy and reaction time data. Unlike in Experiment 1, here there was an effect of validity on accuracy, f(1,51) =11.21, p=.002, np^2= .18, with participants being more accurate on valid trials (mean 95.8%, SD 3.0%) than invalid trials (mean 94.1%, SD 3.1%). There was no effect of Block number on accuracy f(3,125) =0.23, p=.838, np^2 =.01. There was a small to medium interaction effect present f(3,153) =.2.90, p=.037, np^2=.05. For valid trials, reaction times decreased between blocks one and two then increased between blocks three and four. Conversely, for invalid trials, reaction times increased between blocks one and two but decreased between blocks three and four.

Reaction times

Again, there was a large main effect of validity on reaction times f(1,51)=42.31, p<.001, np^2=.45, with valid trials (mean 333ms, SD 44) being responded to more quickly than invalid trials (mean 347ms, SD 50). There was also a medium effect of block f(2,96) =4.57, p=.004, np^2=.08. Pairwise comparisons showed that participants were slower to respond in Block 1 (mean 351ms, SD 69) than Block 2 (mean 338ms, SD 46) (p=.038, CI[0.74,25.05]), suggesting that participants' response times tended to improve with experience. Unlike the effect found in Experiment 1, here, there was no difference in reaction times between Blocks 2 and 3 (p=.621, CI[-6.69,11.09]. There was a medium sized interaction between validity type and Block number, f(3,153) =5.57, p=.001, np^2=.10. This was due to the difference between valid and invalid response times (the gaze-cueing effect) getting larger over time due to reaction times for valid trials decreasing over time (see **Figure B3** – middle).

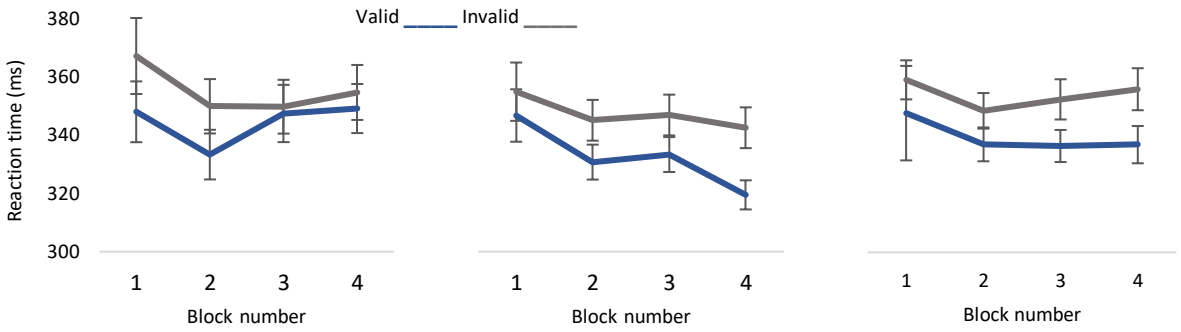


Figure B3: *Reaction times for Experiment 1 (left), Experiment 2 (middle) and Experiment 3 (right) for valid and invalid trials across each block of the gaze-cueing task. Bars represent SEM.*

Experiment 3

Accuracy

Overall accuracy for the gaze-cueing task was high and similar to the previous experiments (accuracy 94.7%, excluded too fast/slow 4.5%, incorrect 0.8%). Validity (2: Valid, Invalid) by Block (4 levels) repeated measures analyses of variance were performed for both accuracy and reaction times data. The accuracy data showed no effect of Validity f(1, 53) <.01, p =.972, np^2 <.01 and a medium effect of Block f(2, 110) =3.11, p =.046, np^2 =.06. Pairwise comparisons showed that accuracy in Block 4 was lower than in Block 1 (p=.004, CI[-3.49,- 0.68]) and Block 2 (p=.005, CI[-3.00.-0.55]), indicating that accuracy diminished over time (all other comparisons p>.10). There was no interaction f(2, 106) =1.13, p =.327, np^2 =.02 present.

Reaction times

The results were somewhat different for reaction times (see **Figure B3** – right). There was a large main effect of validity f(1, 53) =32.22, p <.001, np^2 =.39, with participants responding more slowly for invalid trials (mean 354ms, SD 47ms) than valid trials (mean 340ms, SD 40). This was accompanied by a medium sized main effect of Block f(2, 119) =3.26, p =.037, np^2 =.06. Pairwise comparisons revealed an expected difference between Blocks one (mean 353ms, SD 46) and two (mean 343ms, SD 42), with participants responding faster by the second block (p=.004, CI[-17.67,-3.37]). Reaction times were also faster for Block three (mean 344ms, SD 45), than for block one (p=.009, CI[-15.61,-2.36]). There was no difference in reaction times, however, between blocks two and three (p=.525, CI[-3.59,6.96]), perhaps owing to the fact that only two of the faces changed their behaviour. There was no meaningful interaction between validity and block number for reaction times f(3, 159) =1.90, p =.132, np^2 =.04.

Experiment 4

Accuracy

Improving environment

In the improving environment, the ratio of valid to invalid trials was only 1:2 ('initial') for the first two blocks of the cueing task, but *increased* to 2:1 ('final'), as four group members increased the number of valid gaze-cues they displayed over time. As there were unequal numbers of trial types in each block and this changed across blocks, accuracy data was analysed by comparing average

accuracy across the first two blocks (1/3 valid trials) to the final two blocks (2/3 valid trials) using a 2 (trial validity) * 2 (time: initial and final) repeated measures ANOVA.

Overall accuracy for the gaze-cueing task was high and similar to the previous experiments (accuracy 95%, excluded too fast/slow 4 %, incorrect 1%). In the improving environment (1:2 valid-to-invalid ratio increasing to 2:1), there was a medium effect of validity on accuracy, f(1,57)=5.33, p=.025, np^2=.09, with valid trials (mean 95.5, SE .37) being responded to more accurately than invalid trials (mean 94.4, SE .40). There was no effect of time f(1,57)=2.59, p=.114, np^2=.04, likely due to a medium interaction between validity and time f(1,57)=4.16, p=.046, np^2=.07. This interaction was due to accuracy decreasing over time for invalid trials but increasing slightly for valid trials (see **Figure B4** – top, left). As the number of valid cues increased, invalid cues became less frequent and so were primed for less. It is also worth noting, however, that an error when there are only 32 trials (1/3 of total trials) as compared to an error when there are 64 trails (2/3 of total trials) would have a larger effect on accuracy scores.

Worsening environment

In the worsening environment, the ratio of valid-to-invalid trials was the reverse of the improving environment (i.e., 2:1 valid-to-invalid ratio – 'initial' – *decreasing* to 1:2 – 'final'), There was a large effect of validity on accuracy f(1,53)=15.8, p<.001, np^2=.23, with valid trials (mean 95.6, se .48) being responsed to more accurately than invalid trials (mean 93.8, se .44). There was also a medium effect of time f(1,53)=6.04, p=.017, np^2=.10, with initial trials (mean 94.2, se 0.5) receiving more errors than final trials (mean 95.1, se 0.4). This was qualified by the presence of a small to medium interaction between validity and time f(1,53)=3.74, p=.059, np^2=.07. With the mirror opposite pattern to the improving environment, here in the worsening environment, again accuracy for valid trials was similar over time while accuracy for invalid trials improved over time as they became more frequent (**Figure B4** – top, right).

Reaction times

Improving environment

Reaction time data showed that there was a large effect of validity on response times f(1,57)=27.02, p<.001, np^2=.32, with valid trials (mean 330, SE 4.6) being responded to more quickly than invalid trials (mean 341, SE 5.0). As with accuracy, there was no effect of block on reaction times f(2.3,132)=1.96, p=.138, np^2=.03, but there was a large interaction between validity

and block f(3,171)=12.60, p<.001, np^2=.18. This was due to the fact that while response times for valid trials remained relatively stable over time, response times for invalid trials increased after block 2, following the change in validities and the decrease in the frequency of invalid trials (**Figure B4** – bottom, left). As invalid trials became less frequent, the cueing effect became larger as reaction times for invalid trials increased.

Worsening environment

For reaction times, there was a large effect of validity f(1,53)=59.97, p<.001, np^2=.53, with valid trials (mean 332, se 5.2) being responded to more quickly than invalid trials (mean 350, se 5.9). There was no effect of time f(2.2,118)=0.38, p=.708, np^2<.01, however there was a small to medium sized interaction between validity and block f(2.5,130)=3.65, p=.021, np^2=.06. With the opposite direction for effect to the improving environment, this was attributable to the gaze-cueing effect (valid less invalid) decreasing over time. Here, reaction times for invalid trials were steady over time, while responses for valid trials took longer as they became less frequent (**Figure B4** – bottom, right).

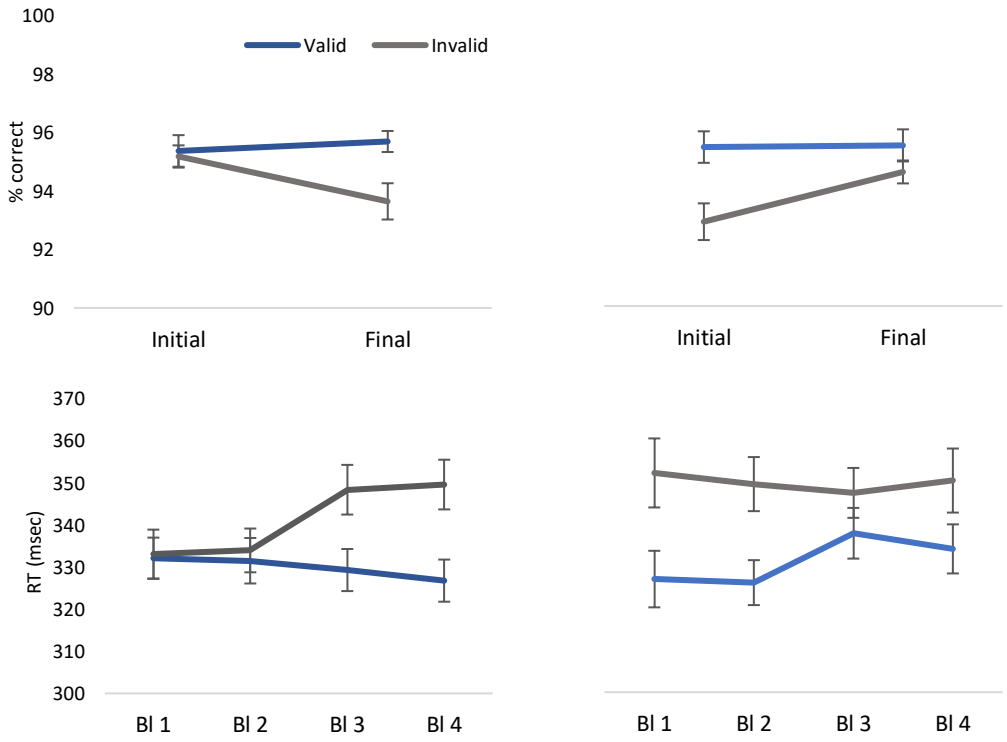


***Figure B4**: Average accuracy rates for* Initial *(blocks 1 and 2) and* Final *(blocks 3 and 4) (TOP) and Reaction times for each block (BOTTOM)* for valid and invalid trials *for Experiment* 4 *for the Improving* (*LEFT) and Worsening (RIGHT) environments. Bars represent SEM.*

Experiment 5

Accuracy

Consistently Deceitful group

In the unhelpful environment (1/3 valid trials), there was a medium main effect of validity f(1,51)=4.28, p=.044, np^2=.08, with accuracy being higher for valid (mean 95.4, SE .43) than invalid (mean 94.4, SE .43) trials. Accuracy was essentially the same across blocks, f(3,153)=0.47, p=.706, np^2=.01, and there was no interaction between validity and block f(3,153)=0.51, p=.677, np^2=.01 (**Figure B5** – top, left).

Consistently Helpful group

In the helpful environment (2/3 valid trials), there was a large main effect of
Validity f(1,53)=9.82, p=.003, np^2=.16, with accuracy being higher for valid (mean 95.2, SE.43) than invalid (mean 93.6, SE .65). Similar to the stable unhelpful environment, accuracy was essentially the same across blocks, f(1.6,87)=0.80, p=.432, np^2=.02, and there was no interaction between validity and block f(3,159)=1.73, p=.163, np^2=.03 (**Figure B5** – top, right).

Reaction times

Consistently Deceitful group

Reaction times looked similar for valid and invalid trials (**Figure B5** – bottom, left), producing only a small cueing effect (6 ms). This is likely due to the majority of trials being invalid. There was a medium main effect of validity present f(1,51)=6.26, p=.016, np^2=.10, with reaction times being slightly faster for valid (mean 334 ms, se 6 ms) than invalid (mean 340, se 7ms) trials. Response times looked to increase across blocks, but there was no meaningful effect of time f(2.2,111)=2.12, p=.120, np^2=.04, and there was no interaction between validity and block f(2.4,121)=0.31, p=.816, np^2<.01.

Consistently Helpful group

Results for reaction times when most trials are valid were a little different. As expected, when there were very few invalid cues, the cueing effect became larger (25 ms) and there was a very large main effect of validity f(1,53)=99.6, p<.001, np^2=.65, with reaction times being much faster for valid (mean 327 ms, se 5 ms) than invalid (mean 352, se 6ms) trials. There was no main effect of block f(1.8,96)=0.90, p=.401, np^2=.02, however, there was a medium interaction between

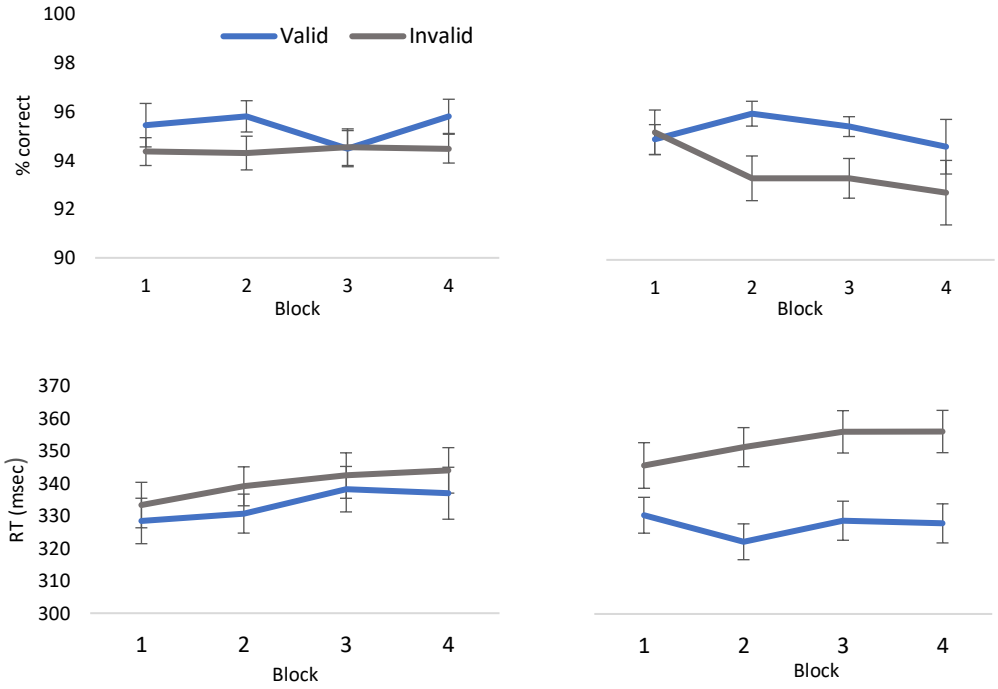


Figure B5: *Average accuracy rates (TOP) and Reaction times (BOTTOM) for each block separated for valid and invalid trials for Experiment 5 for the Deceitful (LEFT) and Helpful (RIGHT) environments. Bars represent SEM.*

validity and block f(2.5,133)=4.32, p=.010, np^2=.08. In general, reaction times for valid trials were stable over time, whereas they increased for invalid trials (**Figure B5** – bottom, right).

Appendix C: Chapter 4 – Designing the Card game

C1: Selection of faces

The faces used throughout the experiment were taken from the Oslo Face dataset (Chelnokova et al, 2014). The final faces chosen aimed to represent an 'average' group of people. As such, all white, male faces were ranked along three dimensions and those located near the average for all three measures were selected (see **Table C1**). Faces came pre-rated by a group of ~40 (20 female) judges.

Image	Attractiveness	Trustworthiness	Dominance
M006	4.0	6.1	4.4
M026	4.4	5.8	4.4
M029	4.2	5.8	3.9
M035	4.6	6.1	4.0
M038	4.6	5.4	5.1
M051	4.2	5.7	4.0
M053	3.6	5.4	4.5
M067	3.7	5.0	4.7
Average (SD)	**4.2 (0.4)**	**5.6 (0.4)**	**4.4 (0.4)**
Whole group (n=84)	3.8 (0.8)	5.3 (0.7)	4.2 (0.9)

Table C1: *Average ratings for each face used in the task together with average ratings for the dataset (male only) as a whole.*

C2: Card Game parameters

The following details how the card game was constructed so as to meet the conditions required. Not all playing card types were used in the game. No Aces, tens or picture cards were included. The face up card that was visible to players could take the form of 4, 5, 6 or 7. The unknown face down card could be either a 2, 3, 4, 5, 6, 7, 8 or 9. The face down card could never be the same as the face up card, meaning there were only ever seven possible outcomes. The sequence of face up cards presented was randomly generated and resulted in each card being shown approximately the same number of times.

The probability of whether a face down card is lower or higher than the face up card varies depending on the value of the face up card (see **Table C2**). This means that some outcomes are more likely than others. It was important for the game to 'feel real' for participants (i.e., that there was not some pattern in the trial outcomes – other than that relating to the gaze-cues' validities) and trial outcomes needed to conform, approximately, to what would be expected if cards really were being drawn at random. This meant that around 36% of trials' outcomes should be 'improbable', or, the least likely given the value of the face-up card. This needed to be achieved while also ensuring that how the trials were associated with each particular face did not impact upon participants' experience. In other words, one face could not be paired with more improbable trials than another.

There were four faces used in the card game and all faces, irrespective of their condition, were associated with the same numbers of improbable and probable trials. A 64/36 split between probable and improbable trials, respectively, was not possible with this design (see **Table C2**). This was because it would require a ratio of 18 and 10 trial types (or 28 in total), which is not divisible by four. Doubling the number of trials to 36 and 20 (totalling 56 trials per block) was not ideal as Experiment 2 required twice as many blocks as Experiment 1, which would have required 336 trials in total (6*56). Given that participants could take their time over each trial, it was decided that this would be too many trials. Instead, an alternative ratio of trials that could be split equally across four faces and still closely adhere to the outcomes that would be expected to occur by chance was chosen as follows. There were 12 (37.5% - as opposed to 36%) improbable trials per block and 20 probable (62.5% - as opposed to 64%) trials. This created blocks of 32 trials and resulted in each face being seen eight times and paired with three (4/12) improbable trials and five (4/20) probable trials per block.

Each face always looked at the correct (veridical cues) or incorrect (non-veridical cues) answer, regardless of the trial's probability, in any given block. This meant that for the majority of time a cooperative face would look at the most probable outcome (5/8) and a deceitful face would look at the least probable outcome (5/8). When associated with a trial with an improbable outcome (3/8), a cooperative face would look at the improbable outcome while a deceitful face would look towards the mostly likely outcome (3/8).

Face up card	**4**	**5**	**6**	**7**
Possible Lower cards	2, 3	2, 3, 4	2, 3, 4, 5	2, 3, 4, 5, 6
Total	2	3	4	5
Possible Higher cards	5, 6, 7, 8, 9	6, 7, 8, 9	7, 8, 9	8, 9
Total	5	4	3	2
Proportion Probable	5/7 = 0.71	4/7 = 0.57	4/7 = 0.57	5/7 = 0.71
Average	**18/28 = 0.64**			
Proportion Improbable	2/7 = 0.29	3/7 = 0.43	3/7 = 0.43	2/7 = 0.29
Average	**10/28 = 0.36**			

Table C2: *Each face down card has a limited number of possible outcomes that are either lower or higher than the face up card. This produces the probability associated with each face up card which, when taken together, determine the probability of a trial having a probable or improbable outcome.*

C3: Social Measures

Four versions of the card game were created. Each of the eight faces was only a card game face in two of the versions (see **Table C3**).

Version	**Veridical Face 1**	**Veridical Face 2^**	**Non-ver. Face 1^**	**Non-ver. Face 2**	**Unfamiliar Faces***
1	Face 1	Face 2	Face 3	Face 4	5, 6, 7, 8
2	Face 5	Face 6	Face 7	Face 8	1, 2, 3, 4
3	Face 3	Face 8	Face 5	Face 2	1, 4, 6, 7
4	Face 7	Face 4	Face 1	Face 6	2, 3, 5, 8

Table C3: *Table showing which condition each face represented in each version of the task. ^ In experiment 2, the Veridical face became Non-veridical and the Non-veridical face became Veridical, in Block 4. *These faces were used in social measures tasks that took place after the card game.*

Appendix D: Chapter 4 – Mixed effects model selection and results

The following describes how each model used in the final analyses was decided upon before proceeding to describe the outcomes for model comparisons, which form the bases of the results in the main text.

To measure the effects of a variable in a mixed model, a likelihood-ratio test between two models, where one contained the independent variable in question and the other did not, was performed. The general approach, then, was to create two models; one with the variable of interest and one without. If the model with the additional variable explained significantly more variance, it provided confidence that the variable helps predict the dependent variable (Barr et al., 2013; Matuschek et al., 2017).

There are different ways to approach mixed models, and there is currently no single agreed upon way as to how to 'fix' models or deal with errors when they arise. Throughout this paper, models have been formed and adjusted based on the advice of the current literature. Namely, baseline models should, where possible, be kept maximal (Barr et al., 2013). This means that both random slopes and intercepts should be included. On occasions where models did not converge, or an error was returned (e.g. A potential 'singularity' issue), they were simplified based on current best practice (see Matuschek et al., 2017). The present advice is to make a model more simple, meaning it stands a higher chance of converging, while also not sacrificing variance explained. Put another way, if a simple model explains roughly the same (or more) variance than a more complicated model, the simple model should be used.

D1: Experiment 1

D1-1: Card Game

The following tests were performed to test whether fixed effects of Time (block(s)) and Condition (faces' cue Validity: Invalid cue = non-veridical information / Valid cue = veridical information)

predicted performance (Accuracy and Reaction times) in the task. Here, Time represented all 96 trials in the card game.

Accuracy

For any given trial, participants can guess correctly or incorrectly. To test whether the participants performed better over time, general logistic mixed effect models were used.

Selection of Base Model

Participants (Subject) were included as a random factor in the model as it was expected that participants would vary in their performance, regardless of condition. The initial baseline model was kept maximal and took the form:

Maxmodel = Accuracy ~ (1 + Time * Validity | Subject)

Where 'Validity' represents whether a trial was valid or invalid which is based upon the cue validity of the particular face accompanying that trial. This model did not converge, so it was simplified by testing the following, progressively simpler models:

Model A = Accuracy ~ (1 + Time + Validity | Subject)
Model B = Accuracy ~ (1 + Time | Subject)
Model C = Accuracy ~ (1 | Subject)

Model A was the first model to converge. However, the model fit had singularities, suggesting that it was overfit (Matuschek 2017). The simpler models were run to see if they explained approximately as much variance as Model A. Model A and Model B explained similar levels of variance ($\chi^2(3)=1.467$, $p=.690$). Model B did, however, explain more variance than Model C ($\chi^2(2)=297.72$, $p< .001$). As such, Model B was chosen as the Base Model, since this was the simplest model which could not be simplified further without a large reduction in variance explained. This model includes a random factor of participant (Subject), with a random intercept and a random slope of time.

Base Model = Accuracy ~ (1 + Time | Subject)

Results

To test whether fixed effects of time and validity better predicted accuracy, we used several models of the form:

Model 1 = Accuracy ~ Time + (1 + Time | Subject)
Model 2 = Accuracy ~ Time + Validity + (1 + Time | Subject)
Model 3 = Accuracy ~ Time * Validity + Time + Validity + (1 + Time | Subject)

By comparing Model 1 (main effect of time) to the Base Model, it was found that individuals did increase their accuracy over time ($\chi^2(1)=40.07$, $p< .001$). Adding a fixed effect of Time to the Base Model significantly increased model fit. Comparing Model 2 (adding a main effect of Validity to main effect of Time) to Model 1 showed that there was no main effect of Validity ($\chi2(1)=0.05$, $p =0.822$), meaning there was no difference in accuracy rates for valid and invalid trials over time. Finally, comparing Model 3 (interaction between time and validity) to Model 2 showed there was no evidence of an interaction between Time and Validity ($\chi^2(1)=1.55$, $p=.214$).

Reaction Time

To test the effects of Time and Validity on reaction time, linear mixed effect models were used. As above, several models were used to explore main effects and interactions.

Selection of Base Model

Again, a random effect of participant (Subject) was included in the baseline model, which initially took the maximal model form:

Maxmodel = Reaction Time ~ (1 + Time * Validity | Subject)

This model converged, but had singularities which might represent model overfit. Thus, it was simplified by testing the following, progressively simpler models:

Model D = Reaction Time ~ (1 + Time + Validity | Subject)
Model E = Reaction Time ~ (1 + Time | Subject)
Model F = Reaction Time ~ (1 | Subject)

Model D was compared to the maxmodel and found that the maxmodel predicted significantly more variance than Model D ($\chi^2(4)=36.62$, $p< .001$). The maxmodel was thus selected as the Base Model, since it could not be simplified further without significantly reducing the fit of the model.

Base Model = Reaction Time ~ (1 + Time * Validity | Subject)

Results

To test whether reaction times were affected by Time and trial Validity, a series of models were created as follows:

Model 1a = reactiontime ~ Time + (1 + Time * Validity + Time + Validity | Subject)

Model 2a = reactiontime ~ Time + Validity + (1 + Time * Validity + Time + Validity | Subject)

Variables 'Time' and 'Reaction Time' were normalised to get the models to converge. Comparing Model 1a to the Base model showed that participants reduced their reaction time over Time ($\chi^2(1)=24.77$, $p< .001$). To test for the main effect of Validity, Model 2a was created, however, it did not converge. Thus, as per Barr (2013), the random effects in these models were simplified until they did converge. The new models took the form shown in Model 1b, 2b, and 3b (the random slope of the interaction between Time and Validity was removed).

Model 1b = reactiontime ~ Time + (1 + Time + Validity | Subject)

Model 2b = reactiontime ~ Time + Validity + (1 + Time + Validity | Subject)

Model 3b = reactiontime ~ Time * Validity + Time + Validity + (1 + Time + Validity | Subject)

Comparing Model 1b to 2b demonstrated no main effect of Validity ($\chi^2(1)=0.50$, $p=.823$). However, comparing Model 2b to 3b, showed that there was an interaction between Time and Validity ($\chi^2(1)=24.98$, $p< .001$). When the first four trials are removed from analyses, the interaction goes away. Comparing Model 2b with Model 3b now shows no significant interaction ($\chi^2(1)=0.10$, $p=.751$). As such, the original interaction was likely an artefact of starting the game

with two valid trials, and there is no actual interaction effect between validity and time when predicting reaction time.

D1-2: Social Measures

Niceness

A test was performed to see whether Condition (including four card game faces and four unfamiliar faces) predicted niceness ratings. For the purposes of the models, there were two observations each for valid and invalid trials and four for unfamiliar faces. These were treated as three Conditions: Valid, Invalid and Unfamiliar.

Selection of Base Model

Participants were expected to vary both in terms of their general niceness ratings and how they perceived each of the eight faces, regardless of Condition. In other words, some faces will be liked more or less than others based purely on individuals' own preference, regardless of what condition they represent in the task. With this in mind, random effects of both participant (Subject) and stimulus (Face) were included and the base model was kept maximal by including random slopes and intercepts as follows:

Maxmodel = Niceness ~ 1 + (1 + Condition | Subject) + (1 + Condition | Face)

The maximal model warned of singularities, but it still converged. Simpler models were checked to see if they could explain a similar amount of variance. Two models were run, each of which removed one of the random slopes:

No random slope Subject = Niceness ~ 1 + (1 | Subject) + (1 + Condition | Face)
No random slope Face = Niceness ~ 1 + (1 + Condition | Subject) + (1 | Face)

When removing either the random slope of Condition on the random-effect of Subject ($\chi^2(5)=97.75$, $p< .001$) or Face ($\chi^2(5)=27.25$, $p< .001$), model fit was significantly reduced compared to the maxmodel. As such, the maxmodel was chosen as the Base Model:

Base Model: Niceness ~ 1 + (1 + Condition | Subject) + (1 + Condition | Face)

Results

The main effect of Condition was tested by comparing the Base Model to Model 1:

Model 1: Niceness ~ Condition + (1 + Condition | Subject) + (1 + Condition | Face)

There was a main effect of condition for niceness ratings ($\chi^2(2)=29.11$, $p< .001$).

Trust Decisions

A test was performed to see whether Condition (including four card game faces and four unfamiliar faces) predicted trust decisions. For the purposes of the models, there were two observations each for valid and invalid trials and four for unfamiliar faces. These were treated as three Conditions: Valid, Invalid and Unfamiliar.

Selection of Base Model

The same approach as Niceness, above, was used and the results returned were similar. The maximal model once again warned of singularities, but it still converged. The same simple models were run, and the same results were found; neither explained variance better than the maxmodel. Removing either the slope of condition from the Subject ($\chi^2(5)=95.02$, $p< .001$) or slope from the random effect of Face ($\chi^2(5)=17.65$, $p=.003$) significantly reduced model fit. So, the maxmodel was used as the Base Model:

Base Model: Trust ~ 1 + (1 + Condition | Subject) + (1 + Condition | Face)

Results

The main effect of Condition on trust decisions was tested by comparing the Base Model to Model 1:

Model 1: Trust ~ Condition + (1 + Condition | Subject) + (1 + Condition | Face)

Model 1 explained significantly more variance and there was a large effect of condition on trusting decisions ($\chi^2(2)=24.08$, $p< .001$).

Ultimatum Game

Participants could either accept or reject each offer. Logistic mixed effect models were used to see if acceptance rates in the Ultimatum games differed depending on the face making the offer (Condition) or the amount being offered (Amount).

Selection of Base Model

The maximal baseline model included random-effects of both participant (Subject) and stimulus (Face). Both of these random effects included random slopes of the interaction of Condition as well as the Amount that was offered:

Max Model = Acceptance ~ 1 + (1 + Condition * Amount | Subject) + (1 + Condition * Amount | Face)

This model did not converge, so increasingly simpler models were tested:

Model G: Acceptance ~ 1 + (1 + Condition + Amount | Subject) + (1 + Validity + Amount | Face)

Model H: Acceptance ~ 1 + (1 + Condition | Subject) + (1 + Condition + Amount | Face)

Model I: Acceptance ~ 1 + (1 + Condition + Amount | Subject) + (1 + Condition | Face)

Model J: Acceptance ~ 1 + (1 + Amount | Subject) + (1 + Condition + Amount | Face)

Model K: Acceptance ~ 1 + (1 + Condition + Amount | Subject) + (1 + Amount | Face)

None of the above converged. Finally, Model L was tested:

Model L: Acceptance ~ 1 + (1 + Condition | Subject) + (1 + Condition | Face)

This converged and an attempt to simplify it without sacrificing variance explained was made. Removing either the random slope of Condition on the random-effect of Subject ($\chi^2(5)=103.30$, $p< .001$) or Face ($\chi^2(5)=11.62$, $p=.045$), significantly reduced model fit compared to Model L. As such, Model L, which included random effects of participant (Subject) and stimulus (Face), each with random intercepts and random slopes of validity, was used as the Base Model:

Base Model = Acceptance ~ 1 + (1 + Condition | Subject) + (1 + Amount | Face)

Results

To test the effects of Condition (3 levels) and Amount (3 levels), together with any interaction between the two, on acceptance rates, models of the following forms were tested:

Model 1a: Acceptance ~ Condition + (1 + Condition | Subject) + (1 + Condition | Face)

Model 2a: Acceptance ~ Condition + Amount + (1 + Condition | Subject) + (1 + Condition | Face)

Comparing Model 1a (main effect of Condition) to the Base Model, showed that Condition did affect acceptance rates ($\chi^2(2)=22.50$, $p< .001$). To test for the main effect of Amount offered, Model 1a was compared to Model 2a (adding a main effect of amount), however, Model 2a did not converge. The random effects structure of the models was simplified to the form shown in Models 1b, 2b and 3b below, since retaining any random slope caused the model to not converge.

Model 1b = Acceptance ~ Validity + (1 | Subject) + (1 | Face)

Model 2b = Acceptance ~ Validity + Amount +(1 | Subject) + (1 | Face)

Model 3b = Acceptance ~ Validity * Amount + Validity + Amount +(1 | Subject) + (1 | Face)

Now, comparing Model 1b to 2b demonstrates a main effect of Amount ($\chi^2(1)=28.84$, $p< .001$), as significantly more variance is explained compared to using Condition alone. When controlling for Condition, the higher the amount offered, the more likely an individual is to accept the offer. Finally, comparing Model 2b to 3b showed that there was no interaction between Condition and Amount ($\chi^2(2)=0.39$, $p=.822$). While condition did predict acceptance rates, the influence of amount was similar across the three conditions.

Effect of learning rate (accuracy) on person perception (niceness and trust)

Since it has been established that face validity (Condition) predicted niceness ratings and trust decisions (see Model 1 for both Niceness and Trust above), participants' learning rates (i.e. The percentage of trials they guessed correctly) were added to Model 1 to create Model 2 as follows:

Model 2: Niceness (or Trust) ~ Condition + percentcorrect + (1 + Condition | Subject) + (1 + Condition | Subject).

By comparing Model 1 to Model 2, it was found that there was no main effect of learning for either niceness ratings($\chi^2(1)=0.45$, $p=.500$) or trust decisions ($\chi^2(1)=0.19$, $p=.660$). This could be because differences between conditions were averaged out when collapsed across conditions. To test whether there was an interaction between faces' validity and learning rate, Model 2 was compared to Model 3:

Model 3 = Trustworthiness (or Niceness) ~ Condition * percentcorrect + Condition + percentcorrect + (1 + Condition | PID) + (1 + Condition | PID).

The interaction of condition and learning rate did predict both niceness ($\chi^2(2)=15.23$, $p< .001$) and trust ($\chi^2(2)=12.33$, $p=.002$). This demonstrates that those who learned the validity of faces' gaze-cues early on in the task went on to have more extreme impressions of the respective faces as compared to those who did not learn the predictive nature of the cues. Indeed, for those people who used a probabilistic strategy, there was no discernible effect of validity at all. Having said that, the question concerned learning, and those who played the probabilistic strategy did not learn. Those who scored above, say, 70%+ overall likely did learn the association between gaze-cues and trials outcomes eventually. Those at 62.5% or below, however, probably just used the cards. If the participants (n=6) who scored less than or equal to 62.5% are excluded and the models are re-run with only those people who did learn, albeit at different rates, then the interaction remains significant for niceness ratings ($\chi^2(2)=7.54$, $p=.023$) but is now only trending for Trust ($\chi^2(2)=4.99$, $p=.082$).

D2: Experiment 2

As per Experiment 1, mixed models were used to explore fixed effects while controlling for random effects and, in general and wherever possible, the same approach was used. Analyses for Experiment 2 were, though, a little different, as each card game face here played a different condition, meaning there was only one observation per condition.

D2-1: Card Game

Subsets of the data were analysed based on blocks as follows:

- First 96 trials (Blocks 1-3) – The three blocks prior to strategy change (same as Study 1)
- Last 96 trials (Blocks 4-6) – The three blocks after strategy change (unique to Study 2)
- Trials 97 – 129 (Block 4) – The block immediately following strategy changes
- Last 64 trials (Blocks 5-6) – The last two blocks, after the participants had some time to learn the change in strategies

Each of these groupings was anlaysed separately to see how accuracy (performance) was predicted by time and condition (the four card game faces).

Accuracy

Linear mixed models were used to assess the influence of time (block subsets) and condition (faces' validity) on accuracy during the card game.

Selecting a base model

As per Barr and others (2013), a maximal Base Model was initially created which took the form:

Maxmodel = Accuracy ~ (1 + Time * Condition | Subject)

This model included random effects of condition and participant (subject) with random slopes and intercepts. The model did not converge, so it was simplified by testing the following, progressively simpler models:

Model A = Accuracy ~ (1 + Time + Condition | Subject)
Model B = Accuracy ~ (1 + Time | Subject)
Model C = Accuracy ~ (1 | Subject)

Model B was the first model to converge. As per Matuschek and others (2017), it was tested whether simpler models explained approximately as much variance as the more complicated models, so Model B was compared to Model C. Model B explained considerably more variance than Model C ($\chi^2(2)=775.66$, $p< .001$). As such, (and as per Matuschek et al., 2017), Model B was

selected for the base model, since Model 3, though simpler, explained much less of the variance. For each of the time subsets, the baseline model was of the form:

Base Model = Accuracy ~ (1 + Time | Subject)

Results

To test whether fixed effects of time (blocks) and condition (faces) predicted accuracy, several models were used as follows:

Model 1 = Accuracy ~ Time + (1 + Time | Subject)
Model 2 = Accuracy ~ Time + Condition + (1 + Time | Subject)
Model 3 = Accuracy ~ Time * Condition + Time + Condition + (1 + Time | Subject)

Model 1 was compared to the Base Model to test whether accuracy was predicted by time, of subsets thereof, as was the case here. It was found that individuals increased their accuracy over time for the first three blocks ($\chi^2(1)=73.74$, $p< .001$), the last three blocks ($\chi^2(1)=68.18$, $p< .001$), and the fourth block ($\chi^2(1)=82.90$, $p< .001$). However, accuracy rates did not increase over the 5th and 6th blocks ($\chi^2(1)=0.76$, $p=.384$). Model 1 demonstrated that the was a main effect of time. By adding a fixed effect of condition to the model (Model 2), it could be identified whether accuracy was better explained by having both factors included. A comparison of Model 2 to Model 1 showed that there was no main effect of condition on accuracy rates during the first three blocks ($\chi^2(3)=4.478$, $p=.214$). This is unsurprising, given the results of Experiment 1. There was an effect of condition for the last three blocks ($\chi^2(3)=42.83$, $p< .001$). By the time the participants played the last two blocks (5-6), accuracy rates were not significantly predicted by condition ($\chi^2(3)=6.25$, $p=.100$). Further, there was certainly no effect of validity for the last block ($\chi^2(3)=4.07$, $p=.253$). This demonstrates that by the end of the task, the accuracy rates of each facial strategy were generally comparable. This is further illustrated in **Figure 18**. By the end of the task, participants were close to ceiling, regardless the validity. Model 3 was designed to see if there was any interaction between time and condition. Unfortunately, the model would not converge with any random slopes and no further analyses were performed here.

Reaction time

Linear mixed effect models were used to see how time (block subsets) and conditions (faces' validity in the card game) predicted reaction times. 65 participants were included in the analyses.

Selecting a Base Model

Again, the initial Baseline Model was kept maximal (Barr et a., 2013) and took the form:

Base Model = Reaction Time ~ (1 + Time * Condition + Time + Condition | Subject)

Results

Several models were used to test whether fixed effects of time (blocks) and condition (faces) predicted reaction times.

Model 1a = Reaction Time ~ Time + (1 + Time * Condition + Time + Condition | Subject)
Model 2a = Reaction Time ~ Time + Condition (1 + Time * Condition + Time + Condition | Subject)
Model 1b = Accuracy ~ Time + (1 + Time | Subject)
Model 2b = Accuracy ~ Time + Condition + (1 + Time | Subject)

Model 1a was compared to the Base Model to test whether time predicted reaction times across the various subsets. Time predicted reaction time when considering all trials ($\chi^2(1)=20.51$, $p< .001$) and the first three blocks ($\chi^2(1)=41.41$, $p< .001$), but reaction time did not change over time from Block 4-6 ($\chi^2(1)=0.32$, $p< .572$). Time did, however, predict a *decrease* in reaction time for Block 4 in isolation ($\chi^2(1)=5.11$, $p=.024$), suggesting that response times initially increased then decreased again during the block. The main effect of condition was tested by adding Condition to the model (Model 2a), but the models failed to converge. Removing the random slope of Condition was required for convergence, generating the 'b' models above. Model 1b was compared to Model 2b. No effect of condition was found for the first three blocks ($\chi^2(3)=4.72$, $p=.193$), however, there was an effect for the final three blocks ($\chi^2(3)=14.09$, $p=.003$). There was no main effect of condition on reaction times when Block 4 was considered in isolation ($\chi^2(3)=3.44$, $p=.329$).

D2-2: Social Measures

Niceness Ratings

Linear mixed models were used to see if condition (card game faces plus four unfamiliar faces) predicted niceness ratings.

Selection of Base Model

Due to the design of the experiment, there was only one observation for each card game condition (as all faces displayed different gaze-cue patterns), while there were four observations for unfamiliar faces. Participants were expected to vary both in terms of their niceness ratings in general and how they perceived each face, regardless of its condition. As no slope can be formed from only one observation, random effects of slope were excluded from the initial maximal baseline model, which included random intercepts only for participant (Subject) and stimulus (Face):

Base Model = Trust ~ 1 + (1 | Subject) + (1 | Face)

Results

To test whether condition (four card game faces plus four unfamiliar faces) predicted niceness ratings, the Base Model was compared to Model 1:

Model 1: Niceness ~ Condition + (1 | Subject) + (1 | Face)

There was a large main effect of condition on niceness ratings ($\chi^2(4)= 104.47$, $p< .001$).

Trust Decisions

Linear mixed models were again used to see if condition (card game faces plus four unfamiliar faces) also predicted trust decisions.

Selection of Base Model

To select a base model for trust decisions, the same approach as the aforementioned section 'Niceness Ratings' was taken.

Base Model = Trust ~ 1 + (1 | Subject) + (1 | Face)

Results

To test whether condition (four card game faces plus four unfamiliar faces) predicted trust decisions the Base Model was compared to Model 1:

Model 1 = Niceness ~ Condition + (1 | Subject) + (1 | Face)

There was a large main effect of condition on trust decisions ($\chi^2(4)=128.18$, $p< .001$).

Acceptance rates in the Ultimatum Game

Logistic mixed models were used to see if acceptance of offers differed depending on the face making the offer (condition) and the offer amount.

Selection of Base Model

The maximal model was used as follows:

Base model: Acceptance (yes or no) ~ (1 | Subject) + (1 | Face)

Results

To test the effects of condition and amount on acceptance rate, several models were used:

Model 1a = Acceptance (yes or no) ~ Condition + (1 | Subject) + (1 | Face)

Model 1b = Acceptance (yes or no) ~ Condition + (1 | Subject)

Model 2a = Acceptance (yes or no) ~ Condition + Amount + (1 | Subject) + (1 | Face)

Model 2b = Acceptance (yes or no) ~ Condition + Amount + (1 | Subject)

Model 3 = Acceptance (yes or no) ~ Condition * Amount + Condition + Amount +(1 | Subject)

Comparing Model 1a to the Base model, showed that condition affected acceptance rates ($\chi^2(4)=49.76$, $p< .001$). To test for a main effect of amount being offered, Model 1a was compared to Model 2a, but Model 2a did not converge. Both models were simplified by excluding the random effect of Face. This was acceptable since Model 1b and Model 1a explained an almost identical

amount of variance ($\chi^2(1)<.01$, $p=.999$). In other words, there was no random effect of Face (stimulus) present in the data, meaning participants perceived the faces (ignoring condition) similarly. Comparing Model 1b to Model 2b found a main effect of amount ($\chi^2(1)=40.08$, $p< .001$); higher offers were accepted more than lower offers when condition was ignored. Finally, a model created to test the interaction between Condition and Amount (Model 3b) did not converge.

Influence of learning rates (accuracy) on social judgements (niceness ratings) and decisions (trust)

To explore how learning during the card game influenced subsequent social judgements and decisions, two subsets of the card game data were selected. For parity with Experiment 1, the first three blocks (1-3) were selected. To see if disruption to accuracy rates had an effect, Block 4 was considered in isolation. It was established that Condition predicts niceness ratings and trust decisions when each of their Base Models was compared to the following Model:

Model 1 = Niceness / Trust ~ Condition + (1 | Subject) + (1 | Face)

This Model was thus extended to include accuracy scores ("Accuracy"), done separately for average accuracy across the first three Blocks (1-3) and fourth Block alone, as follows:

Model 2 = Niceness / Trust ~ Accuracy + Condition + (1 | Subject) + (1 | Face)

When these two models were compared there was no effect of learning (Accuracy) for Niceness ratings when considering either Blocks 1-3 ($\chi^2(1)=1.77$, $p=.183$) or Block 4 ($\chi^2(1)=0.48$, $p=.487$). The same was found for trust, where there was no effect for Blocks 1-3 ($\chi^2(1)=1.00$, $p=.297$) or Block 4 ($\chi^2(1)=0.13$, $p= 0.719$). To test whether there was an interaction between Accuracy and Condition, a final model was created:

Model = Niceness / Trust ~ Accuracy*Condition + Accuracy + Condition + (1 + Validity | Subject) + (1 + Validity | Subject)

For niceness ratings, there was a significant interaction of Accuracy and Condition when considering Blocks 1-3 accuracy rates ($\chi^2(4)=18.74$, $p =0.008839$) and Block 4 accuracy rates ($\chi^2(4)=16.973$, $p =0.001957$). Results were similar for trust ratings, where there was a significant

interaction present when considering Blocks 1-3 accuracy rates ($\chi^2(4)=23.75$, $p< .001$) and Block 4 accuracy rates ($\chi^2(4)=18.61$, $p< .001$).

Appendix E: Chapter 4 – Traditional analyses

Where useful or appropriate, additional and more traditional statistical analyses were performed on the card game and social measures' data. Findings from these additional analyses do not contradict anything reported in the main text. For the sake of simplicity, veridical and non-veridical cue types are referred to under the umbrella term 'validity'.

E1: Experiment 1

E1-1: Card Game

Accuracy

Mean scores in the first block were similar to those that might be expected if people were only using the cards (a probabilistic strategy) to determine their answer. Accuracy then increased over time as participants learned to use the faces' gaze-cues as a means to identifying the correct answer, irrespective of probability. Accuracy rates did not appear to vary depending on whether the gaze-cue provided veridical or non-veridical information (see **Figure E1** – left). This suggests that it was the predictive nature of the face and not its validity that was most salient.

A 3 (Block: 1- 3) by 2 (Condition: veridical, non-veridical) repeated measures Anova on accuracy was performed. There was a large main effect of block $f(2,96)=43.42$, $p< .001$, $np^2= .48$. Unadjusted pairwise comparisons showed that accuracy increased over time such that scores (% correct) were higher in block 2 than block 1 ($p<.001$, CI[9, 17]) and in block 3 than block 2 ($p=.001$, CI[3, 10]). As can be seen in **Figure E1** (left), there was no effect of condition $f(1,48)< .01$, $p=.962$, $np^2 < .01$, and no interaction $f(2,77)=0.23$, $p=.797$, $np^2 < .01$ present.

While all gaze-cues were veridical or non-veridical, the outcome to each trial could be probable or improbable. That is, the outcome could be either the most likely (probable) or the least likely (improbable) given the value of the face up card. To score highly then, the improbable option would, on occasion, have to be chosen. This could be achieved by learning the veracity of each

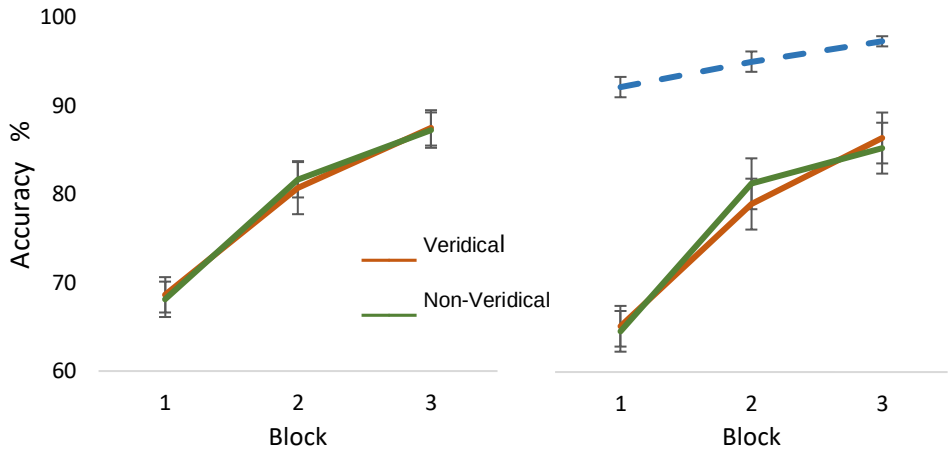


Figure E1*: Exp.1 Accuracy rates by block split across all veridical and non-veridical trial types (left) and then further into probable only (dash line) and improbable (least likely) split between whether the trial was associated with a veridical or non-veridical face (right). Bars represent SEM.*

faces' gaze-cues. If a veridical face looks at the improbable option, it should be selected. If, on the other hand, a non-veridical face looks at the probable answer (which it rarely does), then the improbable option should be selected. In the absence of the faces, choosing the improbable option might be considered 'irrational', which means that getting these trials correct should be contingent upon having detected the faces' contingencies. These improbable trials were extracted from the total number of trials so as to explore learning effects and identify whether people learned to select improbable options at different rates for different faces (see **Figure E1** – right).

A 3 (Block: 1- 3) by 2 (Condition: veridical, non-veridical) repeated measures Anova for just trials with improbable outcomes was performed (36/96 trials – or 12/block). There was a large main effect of block f(2,76)=47.27, p< .001, np^2= .50. Unadjusted pairwise comparisons showed that accuracy (% correct) increased over time such that scores were higher in block 2 than block 1 (p< .001, CI[19, 34]) and in block 3 than block 2 (p=.002, CI[4, 16]). There was no effect of validity f(1,48)<.01, p=.964, np^2< .01, and no interaction f(2,77)=1.30, p=.277, np^2=.03 present. This can be seen in **Figure E1** (right), which also includes accuracy scores for the remaining 'probable' trials (60/96 – or 20/block), and demonstrates that scores were consistently higher for these types of trials. Despite the differences in non-veridical and veridical cue providers' behaviours, there

was no difference in learning rates for improbable trials, though accuracy rates had still not reached ceiling level by the end of the game.

Despite there only being four faces, each of whom provided wholly consistent (in terms of within and across time) cues throughout the entire game, group level accuracy was at less than 90% by the end of the final block. This suggests that some players either learned the association between faces and the informativeness of their cues slowly or they did not learn their behaviours at all. This likely resulted in a disparity between the experiences of the card game players. For instance, those who learned quickly would have both earned more money than those who learned slowly or only ever played the cards. Further, faster learning meant more trials for witnessing the faces' behaviours and experiencing them as either 'cooperative' (veridical) or 'deceitful' (non-veridical). It is possible that this experiential difference may have affected how the players came to perceive the faces later on. For example, we would expect a player who learned to use the faces' cues quickly to have a more positive impression of a correct cue provider than someone who had not detected that they were even providing useful cues, or any type of information at all. The relationship between learning and impressions is considered in separate analyses (see 'Learning rates & impressions below).

Reaction times

There were no time restrictions in the card game, so players could take their time when making their decisions. Though reaction time was not, therefore, of direct importance, it is useful to see how it might have changed over time. Indeed, while players were not urged to go quickly, a decrease in decision making time over time would suggest that they have both become accustomed to the task and, more than likely, learned to use the gaze-cues instead of the value of the face up card when making their decision. It is true that people may become more adept at calculating the probability of a higher or lower card on any given trial as the game progresses, but this would still require longer to calculate than simply following (or not) a face's gaze-cue. This appears to be what occurred (see **Figure E2** – left).

For parity with the mixed effect models analyses, the first four trials were omitted from the following analyses. This removed two veridical and two non-veridical trials, and all trials had 'probable' outcomes, thus ensuring equality across trial types removed. To understand how time (block) and condition (validity) affected reaction times, a 3 (Block; 1, 2, 3) by 2 (Condition: veridical, non-veridical) repeated measures Anova was performed. There was a large main effect of block, f(1,71)=39.93, p< .001, np^2= .45. Unadjusted pairwise comparisons showed that response times decreased over time such that they were lower in block 2 than block 1 (p< .001, mean difference -0.76 sec, CI[-1.0, -0.5]) and in block 3 than block 2 (p=.046, mean difference -0.14, CI[-0.3, -0.0]). There was also an effect of validity f(1,48)=7.94, p=.007, np^2= .14, with cooperative trials (M 1.85 sec, CI[1.62, 1.96]) being responded to more quickly than deceitful trials (M 1.92 sec, CI[1.71, 2.12]) (see **Figure E2**– left). There was no interaction f(2,96)=0.07, p=.932, np^2< .01 present. This suggests that, while there was no difference in accuracy between veridical and non-veridical trials, more time was spent on trials when the cue was non-veridical (and 5/8 times/block looking towards the *least* probable outcome).

Again, improbable trials as they were associated with either veridical and non-veridical faces, were extracted and looked at separately to see whether response times differed for choosing a response based on whether the gaze-cue was veridical or non-veridical. A 3 (Block; 1, 2, 3) by 2 (Condition: veridical, non-veridical) repeated measures Anova on reaction times for trials with improbable outcomes (36/96 trials) was performed. There was a large main effect of block f(1,65)=39.74, p< .001, np^2= .45 (see **Figure E2** – right). Unadjusted pairwise comparisons showed that response times decreased over time such that they were lower in block 2 than block 1 (p< .001, mean difference -1.0 sec, CI[-1.4, -0.7]) and in block 3 than block 2 (p=.057, mean difference -0.2 sec, CI[-0.3, -0.0]). There was no effect of validity present f(1,48)=0.66, p=.420, np^2< .01, and there was no interaction f(2,96)=0.94, p=.395, np^2= .02. This can be seen in **Figure E2** (right) which includes response times for the remaining 'probable' trials (again, excluding the first four trials) for comparison. Interestingly, while response times were initially faster for the probable trials, it seems that over time, selecting the improbable answer when it is looked at by a veridical face (blue line) is done just as quickly.

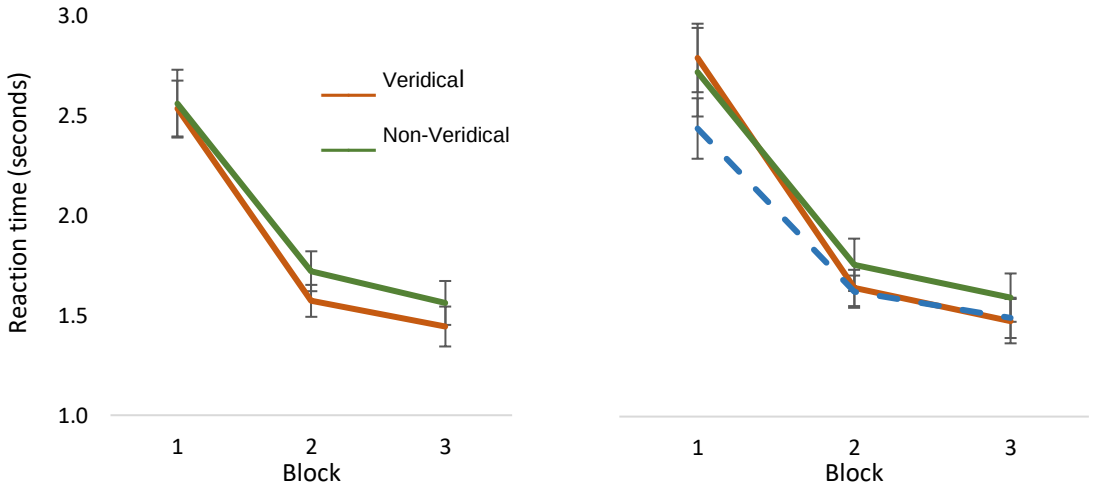


Figure E2*: Exp.1 Reaction times by block split between veridical and non-veridical trial types (left) and then further into probable* only (dash line) and improbable trials split between whether the trial was associated with a veridical or non-veridical face (right). * the first four trials of block 1 were excluded in both figures. Bars represent SEM.*

E1-2: Social Measures

Niceness Ratings

A 1*8 repeated measures Anova explored differences in niceness ratings between all eight faces. There was a clear difference between how the faces were judged $f(4,174) = 30.1$, $p < .001$, $np^2 = .39$ (see **Figure E3** – left). Importantly, the was no difference between how the two veridical ($p > .99$) and non-veridical ($p > .99$) faces were seen. Both of the veridical faces were liked considerably more than both of the non-veridical faces (all p's< .001). Unfamiliar faces were liked less than veridical faces (all p's= .001 or less) but more than non-veridical faces (all p's< .01 - with the exception of 'U4' which was presented last; p< .06). As faces within each condition were viewed very similarly, they were collapsed to explore overall differences between conditions. A 1*3 repeated measures Anova again found a large difference between the faces $f(1,64) = 54.9$, $p < .001$, $np^2 = .53$. As expected, veridical faces were liked more than non-veridical ($p < .001$, $dz = 1.16$, CI[1.9, 3.6]) and unfamiliar ($p < .001$, $dz = 1.14$, CI[1.0, 2.0]) faces, while non-veridical faces were liked less than unfamiliar faces ($p < .001$, $dz = 0.76$, CI[-1.9, -0.7]).

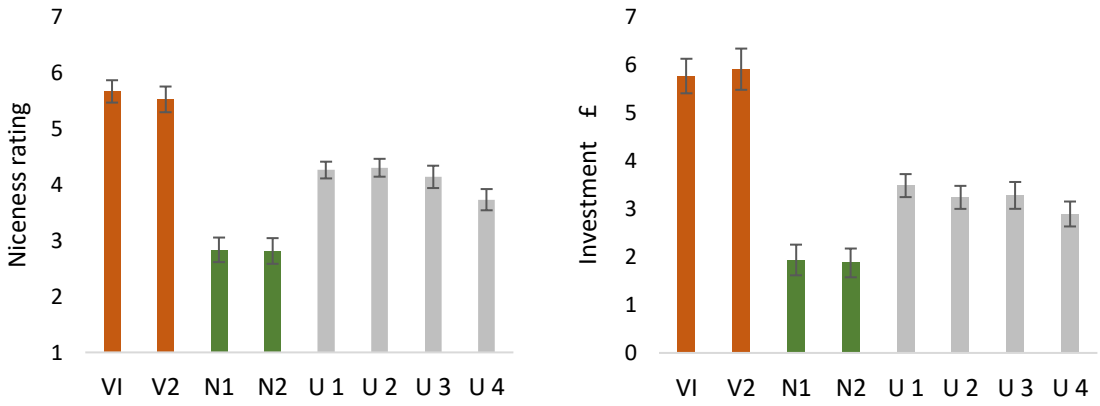


Figure E3*: Exp.1 Niceness ratings (left) and Investment amounts (right) for each face (V = veridical, N = non-veridical, U = unfamiliar). Bars represent SEM.*

Trust Decisions / Investments

Results for trusting decisions closely followed those of niceness judgements (**Figure E3** – right). A 1*8 repeated measures Anova found a large difference between the investments made with each face f(3,159) =30.1, p< .001, np^2= .39. Again, there was no difference between the veridical (p> .99), non-veridical (p> .99) and unfamiliar (all p's> .40) faces. After collapsing across conditions, a 1*3 repeated measures Anova found a large difference between the conditions f(1,60) =53.5, p< .001, np^2= .53. Again, veridical faces were trusted more than non-veridical (p< .001, dz= 1.10, CI[2.7, 5.2]) and unfamiliar (p< .001, dz= 1.76, CI[1.8, 3.5]) faces. Non-veridical faces were trusted less than unfamiliar faces (p< .001, dz= 0.73, CI[-2.0, -0.7]).

Forced Choice Tasks

There were eight pair combinations for each round of the forced choice tasks. Each pair was tested to see whether one condition was selected at an above chance (50%) level using 2-tailed binomial t-tests. The results of the tests are shown in **Table E1.** Veridical faces were clearly chosen over both non-veridical and unfamiliar faces across both tasks. Despite the differences between the tasks, the non-veridical faces were not chosen over unfamiliar faces for either, suggesting that a dislike for them contributed more to the decision than the utility of their gaze-cues in the card game.

	More rounds of the Card Game		More rounds of the Trust Game	
	Proportion	**P value**	**Proportion**	**P value**
V 1 > N 1	.92	<.001	.88	<.001
V 1 > N 2	.92	<.001	.88	<.001
V 2 > N 1	.78	<.001	.82	<.001
V 2 > N 2	.86	<.001	.78	<.001
V 1 > U	.86	<.001	.84	<.001
V 2 > U	.86	<.001	.84	<.001
N 1 > U	.33	.021	.35	.044
N 2 > U	.39	.152	.24	<.001

Table E1: *Outcome of binomial t-tests for forced choice tasks concerning partner choice for more rounds of the card game (left) and more rounds of the trust game (right) for each pair choice. Proportion refers to the proportion of times the condition cited on the left was selected over the condition cited on the right. (V=veridical, N=non-veridical, U=unfamiliar).*

Ultimatum Game

Each of the card game faces offered £3, £4 and £5 while the four unfamiliar faces offered either £2, £3, £4 or £5. For simplicity, the single £2 offer has been excluded from analyses.

Acceptance rates

Participants could either accept or reject each offer in a binary nature, as such, binomial t-tests were used to see if offers were accepted above chance (50%) level (**Table E2**). Irrespective of amount, offers from both veridical and unfamiliar faces were mostly accepted; though those from veridical faces were accepted most often. For non-veridical faces, however, lower offers were more likely to be rejected and higher offers were only accepted at chance level. This suggests that people will turn down offers from people who they do not like or trust even though this means losing money for themselves.

	V 1	V 2	N 1	N 2	Unfamiliar
Offer = £3	(.73) .001	(.80) <.001	(.39) .152	(.33) .021	(.67) .021
Offer = £4	(.86) <.001	(.86) <.001	(.51) >.99	(.43) .392	(.76) <.001
Offer = £5	(.92) <.001	(.90) <.001	(.55) .568	(.61) .152	(.80) <.001

Table E2: *Outcome of binomial t-tests for acceptance rates in the Ultimatum games. With (proportion accepted) and p values provided by condition and amount offered (V=veridical, N=non-veridical, U=unfamiliar).*

Minimum Acceptable Offers (MAO)

Participants could either accept or reject an offer, making it difficult to directly compare differences in acceptance rates between conditions. Owing to the fact that each face offered more than one amount, data was transformed into a continuous variable by determining a participant's minimum acceptable offers (MAO). That is, the lowest amount that each condition (or face) could offer for a player to accept. Each participant was offered £3, £4 and £5 by each face representing a condition (with the exception of unfamiliar faces who each offered only one amount). A MAO could, therefore be any of these amounts or, if all were rejected, an amount of £6 could was inserted instead. While it is not known whether £6 would be accepted or not, it is of higher value than any offers that could be rejected and on the same scale as those that were offered, meaning its inclusion should not skew results. Not all offers are accepted or rejected in a logical fashion and players may display some 'irrational' behaviour. For example, a veridical player could offer £4 which is rejected but £3 which is accepted. From these decisions, it is not clear what the MAO for that player for that condition would actually be. In instances where a player made an 'irrational' decision for one or more condition, their data was removed completely. This resulted in 18 of the 49 players being removed and 31 remaining (see **Figure E4** – left). Some players (n=36), however, made wholly rational decisions as they pertained to the card game faces, but were less consistent with the unfamiliar faces. This is not surprising given that each unfamiliar face only made a single offer and it is plausible that a higher offer may have been rejected if it was made by the least liked of the group. For this reason, additional analyses were performed that excluded the unfamiliar condition altogether and only removed participants who accepted offers from card game faces in an irrational fashion. This resulted in 13 faces being removed and 36 analysed (see **Figure E4** – right).

A 1*5 repeated measures Anova looked at differences in MAOs for veridical, non-veridical and unfamiliar faces and found a large effect of condition $f(2,64)=9.23$, $p< .001$, $np^2= .24$. In order for offers made by non-veridical faces to be accepted, they had to be around £1.00 higher than those made by veridical faces (p's< .003) and around £0.75 higher than those made by unfamiliar faces (p's< .015). There was no meaningful difference in MAO for veridical and unfamiliar faces (p's> .08). This suggests that it is the non-veridical faces whose offers are rejected most often.

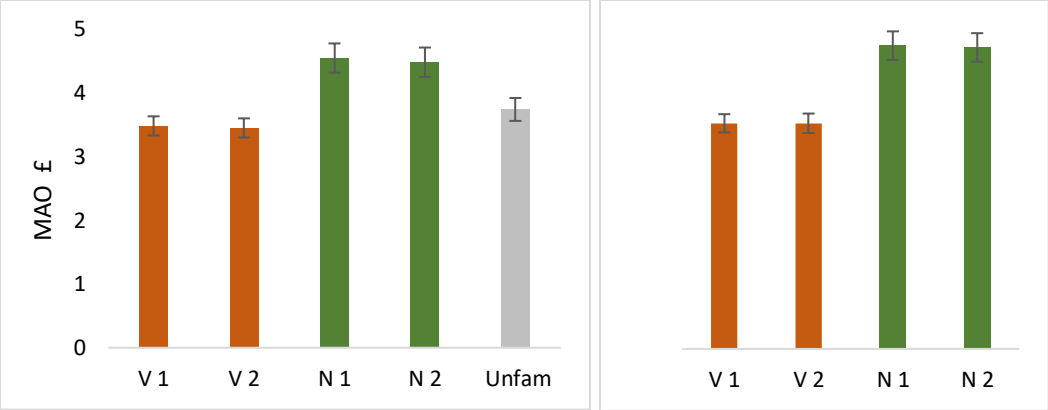


Figure E4*: MAOs for each condition for players who demonstrated rational acceptance rates across all five conditions (left – n=31) and those who only displayed rational acceptance rates for the card game conditions only (right – n=36). Bars represent SEM.*

In the 1*4 repeated measures Anova, the unfamiliar faces were excluded, resulting in an additional five participants being included in the analysis. As can be seen from **Figure E4** (right), MAOs barely changed and again there was another large effect of condition f(2,60)=16.48, p< .001, np^2= .32.

Learning Rates & Impressions

The overarching aim of all these analyses was to explore whether impressions differed depending on learning in the trust game. There are a number of ways that learning rates and the strength of impressions can be calculated and measured. Mixed models found that effects are the product of learning and validity: those who learn the association between faces and their behaviours have greater positive or negative impressions of veridical and non-veridical faces, respectively, than those who do not learn. Learning rate, however, did not impact on ratings. Here, a number of regression and between subject analyses were conducted for trusting decisions (investments) so as to better understand the findings from the mixed models.

When considering the effect of learning on impressions during the writing of the pre-registration, the point was to seek to identify an index for learning. It was thought that people's beta values (i.e. The slope of their accuracy over blocks 1 to 3) might be a useful proxy for learning. However, due

to the speed with which some participants learned this is no longer appropriate, as the function is non-linear. Further, slopes for fast and non-learners might have different intercepts, but they would be similarly flat, making the betas functionally meaningless. Simple regressions between each condition (average cooperative and deceitful) and accuracy scores (percentage correct) can check to see whether accuracy scores are related to investments made. Results showed that investments made to cooperative faces increased as accuracy scores increased (r= .44, p=.002) and investments made to deceitful faces decreased as accuracy increased (r= -.34, p=.018). Accuracy had no correlation with investments made with unfamiliar faces (r= -.04, p=.792).

The relationship between learning rates and the magnitude, or strength, of impressions was also of interest. That is to say, the size of the differences between investments made with veridical and non-veridical faces. Veridical faces were expected to be trusted more positively than non-veridical faces and it was hypothesised that the size of the difference between the two may be positively correlated with learning speed. Here, the strength impressions formed can be measured as the difference between how much is invested (on average) with veridical faces and how much is invested with non-veridical faces (on average). Substantiating the effects found separately for each condition, this analysis shows that the strength of the effect increases as accuracy increases (r= .47, p=.001).

Using a between subjects' approach, people's performance in the card game could also be classified into three categories. There were the majority of players who learned to use the faces very quickly, some who learned, but not immediately, and a further group who appeared to either not learn at all, or were learning very slowly and perhaps unwittingly. Using the visual effects found in the mixed models that explored the influence of learning on trust as a guide, three exploratory performance groups were created (see **Table E3**). As **Figure 17** shows, after around 80% overall accuracy, investments amounts seem to plateau for these 'fast learners'. Using an upper limit of 65% (as opposed to the actual expected accuracy of 62.5%) for players deemed to have used the cards to inform their decisions allowed room for a little trial-and-error in behaviour, as there were also players who scored <62.5% (but above 57%) included in this group. All other players were put into a middle category called 'slow learners'.

	65% or less (card players)	66 - 79% (slow learners)	80% or more (fast learners)	Total Players
Overall (1-3)	11	12	26	49
Block 3 only	9	5	35 *	49

Table E3*: Table showing how many players were allocated to each group based on both their overall performance in the card game (top) and their performance just in Block 3, the final block (bottom). *In this category, 32 of the players scored over 90%.*

Accuracy rates were looked at both in terms of overall accuracy and accuracy in the final block of the game. While overall accuracy is a suitable index for *learning* rates, because it includes all trials, it does not adequately capture whether the contingency has been *learned* by the time the task finishes. As stated previously, in order for updated impressions to be measured, it was important to check that faces' initial behaviours had been sufficiently learned. By looking at accuracy in just the final block of the game, the proportion of players who have fully learned the faces' validities can be found. As can be seen from **Table E3**, many players shifted groups based on the trials being referenced. This reflects the fact that most are performing at almost ceiling level accuracy by the end of the game and only a small number are still learning. Perhaps most interesting of all was the finding that nearly 20% (9 participants) of the players did not use the faces' gaze-cues at all. Given their central placing on every trial and their eye movements, which are known to capture attention, this is highly surprising.

Differences in investments made between learning groups, as determined by overall accuracy scores, were then explored using a 3 (between – group; fast, slow, cards) by 3 (within - condition; veridical, non-veridical, unfamiliar) Anova. There was the expected effect of trust $f(1,61)=40.03$, $p< .001$, $np^2= .47$, where more money was invested with veridical faces. There was no effect of group $f(2,46)=1.14$, $p=.327$, $np^2=.05$, however, this was qualified by an interaction between group and condition $f(4,92)=6.55$, $p< .001$, $np^2= .22$, as can be seen in **Figure E5** (top). The direction of this finding mirrors that found by the regressions stated earlier; differences in investments made with each condition were more pronounced in the group of fast learners. To explore this interaction further, a series of Anovas 1*3 Anovas were performed.

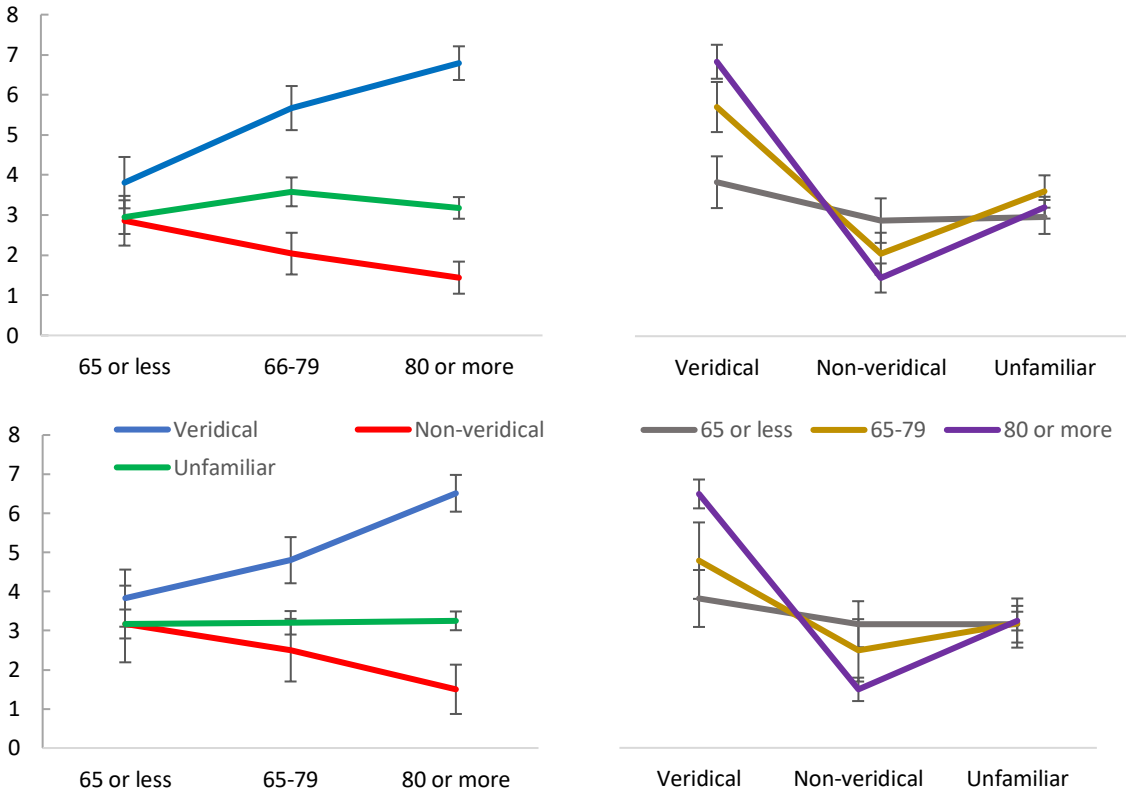


Figure E5*: Investments made by each group based on overall accuracy; plotted by condition (top left) and group (top right). This is compared to investments made by groups based on accuracy in Block 3 only, also by condition (bottom left) and group (bottom right). Bars represent SEM.*

Repeated measures Anovas looked at differences in amounts invested with each condition within each group. In the group of fast learners (80%+), there was a large and clear difference between investments made $f(1,31)=60.45$, $p< .001$, $np^2= .71$. A difference was also found in the group of slower learners (66-79%), where there was just as clear a difference between investments made with each condition $f(1,15)=8.48$, $p=.007$, $np^2= .44$ and, interestingly, in the group that did not learn the gaze-cue contingencies at all (65% or less) $f(2,20)=4.99$, $p=.017$, $np^2= .33$ (see **Figure E5** – top, left). See **Table E4** for pairwise comparisons.

Between subjects Anovas looked at the effect of group on investments made separately with each condition. Veridical faces received different amounts depending on group $f(2,48)=7.55$, $p=.001$. The fast learners invested more than the card players ($p< .001$, CI[1.43, 4.52]) but no more than the slow learners ($p=.139$, CI[-0.38,2.62]). Slow learners also invested more than the card players($p=.044$, CI[0.06, 3.64]). However, the difference was less obvious for non-veridical faces,

f(2,48)=2.42, p =.101 and unfamiliar faces, who received similar amounts regardless of group f(2,48)=0.63, p=.540 (see **Figure E5**– top, right). Here, the only notable difference was that the fast learners invested considerably less with the non-veridical face than the card players (p=.036, CI[-2.73, -0.11]).

Initial results were somewhat similar when group membership was determined by performance in only the final block of the game (see **Figure E5** – bottom, left). There was another interaction present f(4,92)=6.31, p< .001, np^2= .22, which was again explored using a series of one-way Anovas. Repeated measures Anovas again found differences investments made by the fast f(1,44)=59.01, p< .001, np^2= .63 and slow f(2,8)=8.99, p=.009, np^2= .69 learners. Now, however, there was a smaller difference in investments made with each condition by the group of card players f(2,16)=2.13, p=.151, np^2= .21. Between subject Anovas found that differences in investments made with each condition by each group were more pronounced (with the exception of unfamiliar faces) as compared with looking at overall accuracy (see **Figure E5** – bottom, right). Now, both veridical f(2,48)=6.00, p=.005, and non-veridical f(2,48)=3.45, p=.040 faces received different amounts depending on group while the lack of differences between investments made to unfamiliar faces had become even smaller f(2,48)=0.01, p=.986. This demonstrates that most people are performing at ceiling by the end of the first three blocks and provides support for the fact that (ignoring card players) even those small number of people who are still in the learning phase in block three are showing a clear difference between how they view veridical and non-veridical faces.

	<65% (n=11)		66 – 79% (n=12)		80%+ (n=26)	
	Mean difference [95% CI]	P	Mean difference [95% CI]	P	Mean difference [95% CI]	P
V and N	0.96 [0.24,1.67]	.014	3.63 [1.13,6.12]	.008	5.35 [3.98,6.71]	<.001
V and U	0.86 0.24,1.65]	.034	2.08 [0.19,3.97]	.034	3.61 [2.75,4.46]	<.001
I and U	-0.09 [-0.82,0.64]	.787	-1.54 [-2.79,-0.29]	.020	-1.74 [-2.48,-1.00]	-<.001

Table E4*: Contrasts between conditions (V-veridical, N=non-veridical, U=unfamiliar) for each group based on overall accuracy.*

E2: Experiment 2

E2-1: Card Game

As with Experiment 1, it was important to check that participants had learned to use the gaze-cues when making their decisions in the card game. More to the point, it was important to confirm that this happened *before* two of the faces reversed the validity of their cues. It was also of interest to understand how the faces' changes impacted upon participants' performance and how quickly they updated their behaviour to reflect the new contingencies. Reaction times were also explored to see how a change in validity may have affected response times.

Accuracy

Accuracy levels (scores over time) were explored to establish that participants learned to use the faces' gaze-cues over the probabilities associated with the face-up cards. As can be seen from **Figure E6** (left), participants' performance improved significantly over the first three blocks and clearly exceeded levels beyond that which would be achieved if only the cards' values were used to inform decisions. In the fourth block, following a change in two of the faces' behaviour, performance declined. This occurred as the players initially applied the original rule associated with each face before detecting that the rule has changed for two; one face now helps, another does not. Performance soon increases such that by the end of the game, scores are approaching ceiling level.

A 6 (block) by 2 (condition) repeated measures Anova was performed. As is clearly illustrated by **Figure E6** (left), there was a large effect of block $f(3,186)=66.43$, $p< .001$, $np^2= .51$. There was no meaningful difference between accuracy rates for veridical and non-veridical trials, $f(1,65)=3.11$, $p=.082$, $np^2= .05$, though there is the potential for accuracy to have been higher for non-veridical trials than veridical trials in the final block, $t(65)=1.85$, $p=.070$, $dz=0.23$. There was no interaction present $f(3,211)=0.21$, $p=.901$, $np^2< .01$.

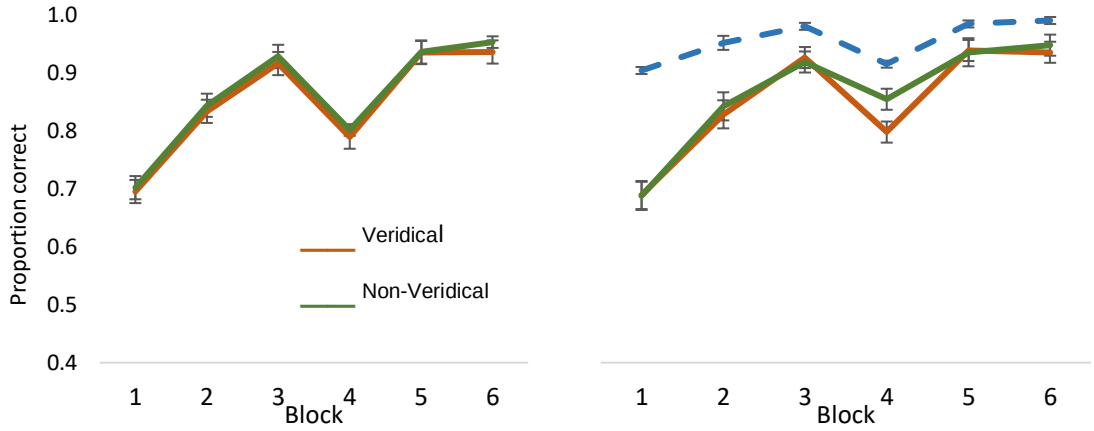


Figure E6: *Exp.2 Reaction times by block split between veridical and non-veridical trial types (left) and then further into probable* only (dash line) and improbable trials split between whether the trial was associated with a veridical or non-veridical face (right). * the first four trials of block 1 were excluded in both figures. Bars represent SEM.*

To further explore how the validity of trials was related to learning, analyses were performed for only those trials associated with improbable outcomes (i.e. The face-down card was the least likely given the face-up card). As can be seen from **Figure E6** (right), accuracy levels were similar for veridical and non-veridical cues with the exception of the block immediately after two faces changed the validity of their cues. Here, accuracy was most negatively affected for improbable trials presented with a veridical cue. That is to say, those trials where the face looked at the least likely option, which would transpire to be the correct outcome. This occurred because a face that had once only provided non-veridical cues changed to provide only veridical cues, and a greater decrease in accuracy suggests that it took longer for players to trust its cue. A similar decrease in accuracy for both trial types would have indicated that people were responding with a primed action based on historic cue validity. The fact that error rates differed for these trial types, however, implies that there is something tangibly different when a face changes from veridical to non-veridical than from non-veridical to veridical behaviour. It is possible that a face's past non-veridical cue behaviour, which has been shown in Exp. 1 to be associated with distrust, interferes with updating behaviour such that it continues to be distrusted for longer than a face who did provides veridical cues (and is trusted) then changes to non-veridical. A 6 (block) by 2 (trial type) repeated measures confirmed the effect of block, $f(3,169)=53.39$, $p< .001$, $np^2= .45$ and validity, $f(1,65)=4.06$, $p=.048$, $np^2= .06$, on accuracy. There was also an interaction, $f(4,228)=4.66$, $p=.002$, $np^2= .07$, reflecting the difference in scores during the fourth block, $t(65)=5.42$, $p< .001$, $dz=0.67$.

Reaction Time

Reaction times were examined to identify whether participants varied in their response times for veridical and non-veridical trials (see, **Figure E7** left). A 6 (block) by 2 (validity) repeated measures Anova was performed. There was a clear effect of block, f(3,186)=23.10, p< .001, np^2= .27, with trials in the first block being responded to much more slowly than later trials in later blocks (all p's< .001). There was no real difference in response times based on validity, f(1,63)=0.13, p=.722, np^2< .01, and there was no interaction f(4,233)=1.32, p=.267, np^2= .02.

Analyses were also performed on just the improbable trials. As can be seen from **Figure E7** (right), trials where the outcome was improbable (least likely) and paired with a veridical cue, players appeared to return slightly longer reaction times initially (block 1) and after the changes to cue behaviour took place (block 4). With the exception of an obvious effect of block, f(3,182)=24.26, p< .001, np^2= .28, there was no difference between improbable trials associated with veridical or non-veridical gaze cues, f(1,63)=0.29, p=.593, np^2< .01, and there was no interaction present f(3,179)=1.37, p=.236, np^2=.02. When including probable trials (but excluding trials 1-4) and looking at block 1 only, f(1,94)=5.09, p=.014, np^2= .08, it can be seen that both veridical improbable (p=.003, mean difference 0.45 seconds, CI [0.16, 0.73]) and veridical improbable (p=.012, mean difference 0.25 seconds, CI [0.06, 0.44]) were responded to more slowly than probable trials. There was no meaningful difference, however, in response times for improbable veridical and improbable non-veridical trials (p=.245, mean difference 0.20 seconds, CI [-0.14,0.54]) in the first block. This suggests that while people may not have detected and reacted to the interaction of gaze-cues and probability (as can be seen in low accuracy levels for improbable trials in block 1), they were attending to them nonetheless, as response times were longer for these trials as compared with probable trials, when their outcome, of course, was not known to participants at the time they made their decision.

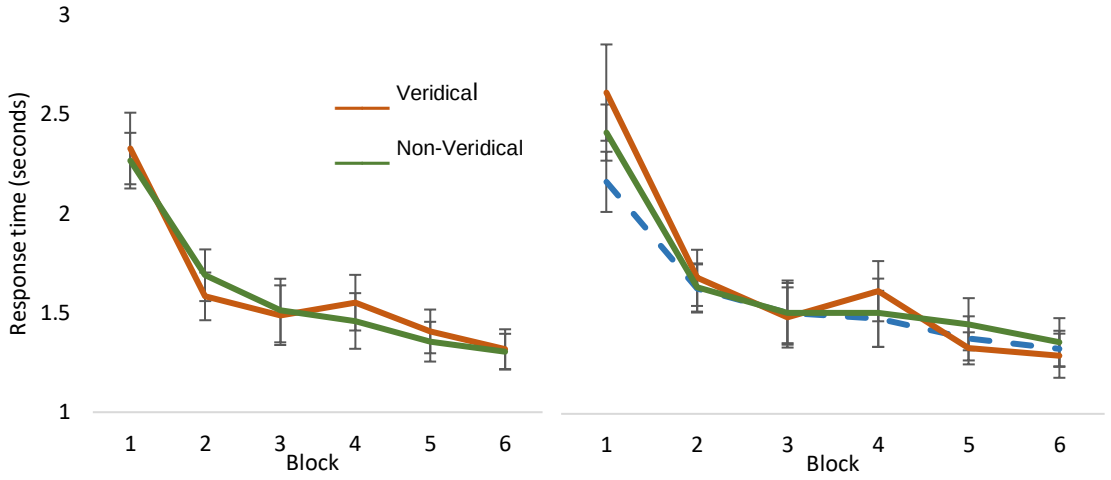


Figure E7*: Exp.2 Reaction times by block split between veridical and non-veridical trial types (left) and then further into probable* only (dash line) and improbable trials split between whether the trial was associated with a veridical or non-veridical face (right). * the first four trials of block 1 were excluded in both figures. Bars represent SEM.*

E2-2: Social Measures

Niceness ratings & Trust decisions (investments)

Two 1*8 repeated measures Anova found large effects of condition on both niceness ratings f(5,335)=32.27, p< .001, np^2= .33 and investments made in the trust game f(5,309)=37.38, p< .001, np^2= .37 (see, **Figure E8** left and right respectively). The wholly veridical condition was liked and trusted far more than all other faces, while the wholly non-veridical face was liked and trusted the

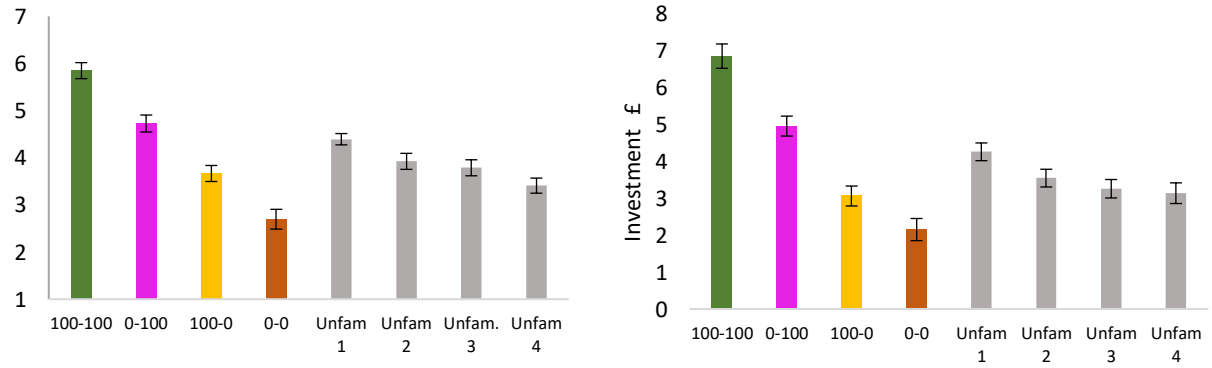


Figure E8*: Niceness ratings (left) and investment decisions (right) for each of the card game faces and the unfamiliar faces. Bars represent SEM.*

least (see **Table E5** for comparisons). Though there was some variation, the unfamiliar faces were rated relatively similarly. Worth noting is the fact that the 'Unfamiliar face 1' condition was seen first by all participants, which may have contributed towards its slightly higher ratings.

Forced Choice Tasks

There were eight pair combinations for each round of the forced choice tasks. Each pair was tested to see whether one condition (left in **Table E6**) was selected over the other condition at an above chance (50%) level using 2-tailed binomial t-tests. While the face that was chosen within a pair clearly depended on the pair being presented, the selection of the face was the same whether participants were choosing a future partner for more rounds of the card game or more rounds of the investment game. In Experiment 1, unfamiliar faces had been chosen in preference to consistently deceitful faces, here, the distinction was less pronounced. Interestingly, the face that became deceitful over time was not selected over unfamiliar faces either, but was selected over consistently non-veridical faces.

	Niceness ratings (scale: 1-7)				**Investment decisions (scale: £0-10)**			
	Mean diff.	Sig.	CI diff. Lower	CI diff. Upper	Mean diff.	Sig.	CI diff. Lower	CI diff. Upper
100-100 & 0-100	1.12	<.001	0.36	1.89	1.89	<.001	0.76	3.03
0-100 & 100-0	1.06	.002	0.24	1.89	1.89	<.001	0.86	2.93
100-0 & 0-0	0.97	<.001	0.30	1.64	0.91	.081	-0.05	1.87
100-100 & 100-0	2.18	<.001	1.30	3.07	3.79	<.001	2.38	5.20
0-100 & 0-0	2.03	<.001	1.23	2.83	2.83	<.001	1.65	3.95
100-100 & 0-0	3.15	<.001	2.14	4.17	4.70	<.001	3.08	6.32

Table E5*: Results of pairwise comparisons (Sidak adjusted) for niceness ratings (left) and investment decisions (right).*

	More rounds of the Card Game		**More rounds of the Trust Game**	
	Proportion	P value	Proportion	P value
100-100 > Unfamiliar	.91	<.001	.94	<.001
100-100 > 0-100	.83	<.001	.83	<.001
100-100 > 0-0	.88	<.001	.86	<.001
100-0 > Unfamiliar	.59	.175	.52	.902
100-0 > 0-100	.21	<.001	.23	<.001
100-0 > 0-0	.64	.036	.68	.004
0-100 > Unfamiliar	.79	<.001	.80	.044
0-0 > Unfamiliar	.47	.712	.41	.175

Table E6: *Outcome of binomial t-tests for forced choice tasks for more rounds of the card game (left) and more rounds of the trust game (right) for each pair choice. The proportion selected refers to the condition listed first.*

Ultimatum Game

There were 16 offers made across the eight faces; three each for the card game faces and one each for the unfamiliar faces. For simplicity, the results of the £2 offer made by an unfamiliar face are excluded from analyses. The acceptance rate for each offer was examined using a binomial t-test set at chance level (50%). The results of which can be found in **Table E7**. In general, higher offers were accepted more than lower offers and offers made by faces that ended by providing veridical cues and those from unfamiliar faces were accepted more often than those made by faces that ended by providing wholly non-veridical gaze-cues.

Minimum acceptable offers (MAOs) for each condition were explored using repeated measures Anovas. After players who made 'irrational' decisions (see Ex. 1 analyses) were removed, there were 37 who made rational decisions for all five conditions (card game plus the unfamiliar) and 46 who made rational decisions for all card game conditions. When considering all five conditions, MAOs did vary f(3,111)=4.18, p=.007, np^2= .10. MAOs were higher for the two conditions that ended by providing non-veridical gaze-cues and similar for the two conditions that ended with veridical gaze-cues and the unfamiliar face (see **Table E8**, left, for pairwise comparisons). This suggests that it is the non-veridical gaze-cues that are driving the effect (see **Figure E9** – left). To see whether the effects changed substantially when looking at MAOs for only the card game conditions, a further test was performed. For this group of players who made consistent acceptance decisions across each of the card game faces, the effects were the same, only more pronounced f(2,107)=7.71, p< .001, np^2= .15 (see **Table E8** and **Figure E9** – right).

	100-100 (consistently Verid.)	0-100 (increase)	100-0 (decrease)	0-0 (consistently Non-Verid.)	Unfamiliar
Offer = £3	(.74) .001	(.62) .064	(.39) .109	(.50) >.99	(.67) .009
Offer = £4	(86.) <.001	(.88) <.001	(.74) <.001	(.59) .175	(.73) <.001
Offer = £5	(.91) <.001	(.92) <.001	(.82) <.001	(.64) .036	(.67) .009

Table E7: *Outcome of binomial t-tests for acceptance rates in the Ultimatum games. With (proportion accepted) and p values provided by condition and amount offered.*

	All five conditions				Card Game conditions only			
	Mean diff.	Sig.	CI diff. Lower	CI diff. Higher	Mean diff.	Sig.	CI diff. Lower	CI diff. Higher
100-100 & 100-0	-0.38	.051	-0.76	0.00	-0.41	.033	-0.79	-0.34
100-100 & 0-0	-0.76	.003	-1.24	-.028	-0.85	.001	-1.32	-0.38
100-0 & 0-0	-0.38	.090	-0.82	0.06	-0.44	.042	-0.85	-0.02

Table E8*: Results of unadjusted pairwise comparisons for the repeated measures Anovas examining MAOs in the Ultimatum Game for all five conditions (left) and card game conditions only (right).*

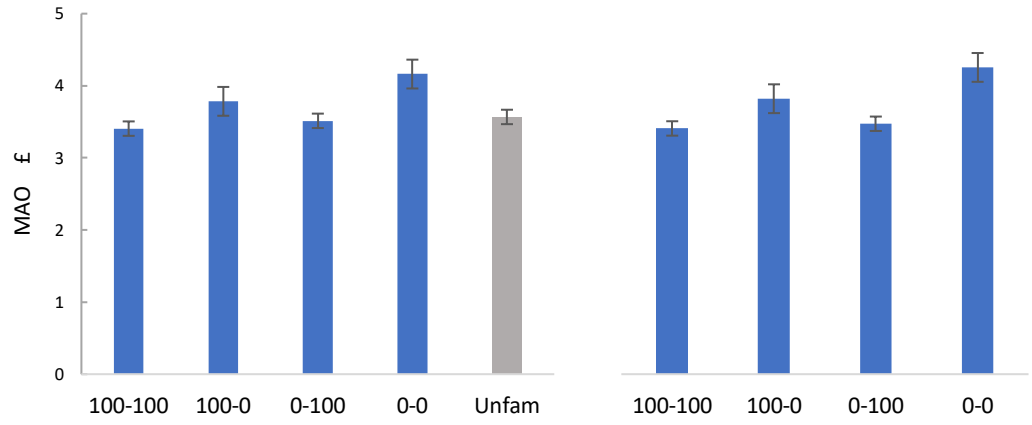


Figure E9: *Charts showing the minimum acceptable offers (MAOs) for all conditions when only those players who responded rationally are included (left) or all of those who responded rationally to only the card game faces' offers (right). Bars represent SEM.*

Learning rates and impressions

Participants were grouped using the same approach as Experiment 1. As this study was three blocks longer, 'overall accuracy' incorporated six blocks in total (see **Table E9**). The relationship between performance in the card game and the strength of impressions was first explored by considering accuracy in the first three blocks (prior to the faces changing) only to see if learning rates in general impacted on social decisions. After, scores from the whole game were used.

Blocks 1-3 only

A 4 (within: condition) * 3 (between: group) Anova was performed for each of niceness ratings and investments in the trust game. For niceness ratings, there was a main effect of condition, as expected, f(3,158)=31.50, p< .001, np^2= .33. There was no main effect of group, f(2,63)=2.85, p=.065, np^2= .08, but this was qualified by an interaction, f(6,189)=2.74, p=.014, np^2= .08, as

	65% or less (card players)	66-79% (slow learners)	80% or more (fast learners)	Total Players
Block 1-3	11	11	44	66
Overall Accuracy (Blocks 1-6)	7	2	57	66

Table E9*: *Participant groupings based on performance across different proportions of the card game.

ratings for each condition differed across groups (see **Figure E10** – Top, left). Ratings did not, however, appear to vary within the group of card players. Ratings between slow and fast learners appeared similar. For investments made, there was again a main effect of condition, $f(3,189)=35.35$, $p< .001$, $np^2= .36$. Similarly, there was no main effect of group, $f(2,63)=0.06$, $p=.943$, $np^2< .01$, but there was an interaction, $f(6,189)=4.24$, $p< .001$, $np^2= .12$, as investments differed substantially across groups based on condition (see **Figure E10** – Top, right).

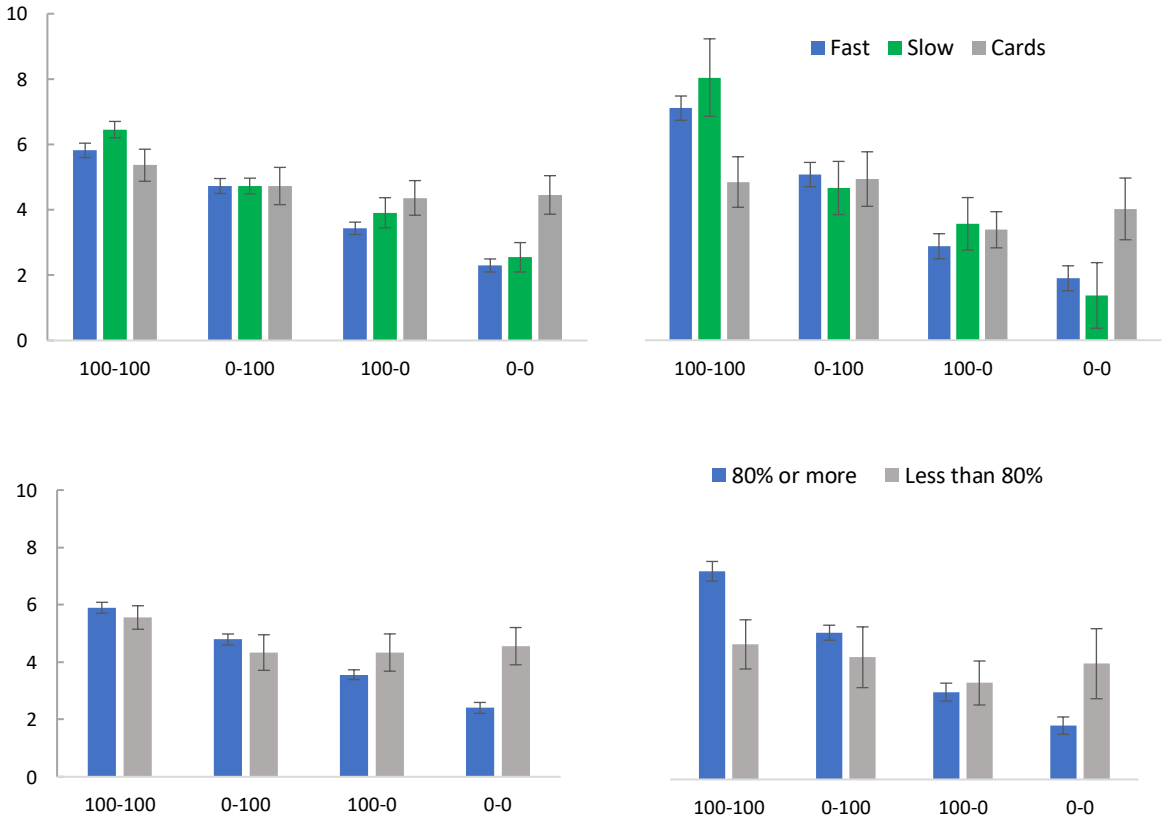


***Figure E10**: Average niceness ratings (left) and investment decisions (right) given performance in the card game. The top row shows players grouped according to their average accuracy score in the first three blocks of the game (Blocks 1-3 – top) and the bottom row shows players grouped by their overall performance across all six blocks (Average accuracy – bottom). Bars represent SEM.*

As can be seen from **Figure E10** (top) both niceness ratings (left) and investment decisions (right) appear to be influenced by learning rates such that those who do not learn to use the faces do not show an effect of condition. Interestingly, there is no clear difference in the strength of effects between fast and slow learners, suggesting that knowledge of the cues' veracity (over time spent using the validity) is sufficient to drive impressions and this is not moderated by the speed with which someone learns.

Overall accuracy (blocks 1-6)

Next, the effect of overall accuracy on social judgements was assessed. By the end of the six blocks, almost all players were performing at ceiling level (see **Table E9**). There looked to be only seven people who consistently played the cards and only two who had learned the gaze-cues contingencies late in the game. This meant there were essentially only two groups; those who learned both pre and post gaze-cues validities and those who either didn't or likely only learned one. These were analysed using a 4 (within: condition) * 2 (between: group) Anova.

For niceness ratings, there was a main effect of condition, $f(2,158)=15.92$, $p< .001$, $np^2= .20$. As before, there was no main effect of group, $f(1,64)=2.47$, $p=.121$, $np^2=.04$, and this was again qualified by an interaction, $f(3,192)=4.47$, $p=.005$, $np^2= .07$ (see **Figure E10** – Bottom, left). Findings were similar for investments in the trust game, where there was a main effect of condition, $f(2,148)=14.17$, $p< .001$, $np^2= .18$, no main effect of group, $f(1,64)=0.12$, $p=.727$, $np^2< .01$, and another interaction, $f(3,148)=6.26$, $p< .001$, $np^2= .09$ (see **Figure E10** – Bottom, right). For those deemed to have learned to use the faces (80%+ accuracy), effects mirrored those observed when group was not a factor, namely, each of the conditions was perceived differently. There were only nine participants in the group of card players, which makes it difficult to place any reliance on the results, however, there does not appear to be any difference in how each of the conditions was viewed.

References

Ambady, N., Bernieri, F. J., & Richeson, J. A. B. T.-A. in E. S. P. (2000). *Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream* (Vol. 32, pp. 201–271). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2601(00)80006-4

Ames, D. L., & Fiske, S. T. (2013). Outcome dependency alters the neural substrates of impression formation. *NeuroImage*, *83*, 599–608. https://doi.org/10.1016/j.neuroimage.2013.07.001

Ames, D. L., & Fiske, S. T. (2015). *formation*. 599–608. https://doi.org/10.1016/j.neuroimage.2013.07.001.Outcome

Ames, D. L., Fiske, S. T., & Todorov, A. T. (2011). Impression formation: A focus on others' intents. In *The Oxford handbook of social neuroscience.* (pp. 419–433). Oxford University Press.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. In *Journal of Experimental Psychology* (Vol. 70, Issue 4, pp. 394–400). American Psychological Association. https://doi.org/10.1037/h0022280

Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression formation. In *The Journal of Abnormal and Social Psychology* (Vol. 63, Issue 2, pp. 346–350). American Psychological Association. https://doi.org/10.1037/h0046719

Anderson, N. H., & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior*, *2*(5), 379–391. https://doi.org/https://doi.org/10.1016/S0022-5371(63)80039-0

Argyle, M., & Cook, M. (1976). Gaze and mutual gaze. In *Gaze and mutual gaze.* Cambridge U Press.

Argyle, M., Lefebvre, L., & Cook, M. (1974). The meaning of five patterns of gaze. *European Journal of Social Psychology*, *4*(2), 125–136. https://doi.org/https://doi.org/10.1002/ejsp.2420040202

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390

LP – 1396. https://doi.org/10.1126/science.7466396

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. In *Psychological Bulletin* (Vol. 137, Issue 4, pp. 594–615). American Psychological Association. https://doi.org/10.1037/a0023489

Barbato, M., Almulla, A. A., & Marotta, A. (2020). The Effect of Trust on Gaze-Mediated Attentional Orienting . In *Frontiers in Psychology* (Vol. 11, p. 1554).

Baron-Cohen, S., Cosmides, L., & Tooby, J. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. Bradford Books.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/https://doi.org/10.1016/j.jml.2012.11.001

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-

Bayliss, A. P., Griffiths, D., & Tipper, S. P. (2009). Predictive gaze cues affect face evaluations: The effect of facial emotion. *European Journal of Cognitive Psychology*, *21*(7), 1072–1084. https://doi.org/10.1080/09541440802553490

Bayliss, A. P., Paul, M. A., Cannon, P. R., & Tipper, S. P. (2006). Gaze cuing and affective judgments of objects: I like what you look at. *Psychonomic Bulletin and Review*, *13*(6), 1061–1066. https://doi.org/10.3758/BF03213926

Bayliss, A. P., & Tipper, S. P. (2006). Predictive gaze cues and personality judgments: Should eye trust you? *Psychological Science*, *17*(6), 514–520. https://doi.org/10.1111/j.1467-9280.2006.01737.x

Beaudoin, C., & Beauchamp, M. H. (2020). Chapter 21 - Social cognition. In A. Gallagher, C. Bulteau, D. Cohen, & J. L. B. T.-H. of C. N. Michaud (Eds.), *Neurocognitive Development: Normative Development* (Vol. 173, pp. 255–264). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-444-64150-2.00022-8

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

Bell, R., & Buchner, A. (2012). How Adaptive Is Memory for Cheaters? *Current Directions in Psychological Science*, *21*(6), 403–408. https://doi.org/10.1177/0963721412458525

Belmore, S. M. (1987). Determinants of attention during impression formation. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *13*(3), 480–489. https://doi.org/10.1037/0278-7393.13.3.480

Berg, J., Dickhaut, J., & McCabe, K. (1995). Berg et al 1995.pdf. In *Games and Economic Behavior* (Vol. 10, pp. 122–142). https://doi.org/10.1006/game.1995.1027

Birmingham, E., & Kingstone, A. (2009). Human social attention: A new look at past, present, and future investigations. *Annals of the New York Academy of Sciences*, *1156*, 118–140. https://doi.org/10.1111/j.1749-6632.2009.04468.x

Blakemore, S.-J., & Choudhury, S. (2006). Development of the adolescent brain: implications for executive function and social cognition. *Journal of Child Psychology and Psychiatry*, *47*(3–4), 296–312. https://doi.org/https://doi.org/10.1111/j.1469-7610.2006.01611.x

Blakemore, S.-J., & Mills, K. L. (2014). Is Adolescence a Sensitive Period for Sociocultural Processing? *Annual Review of Psychology*, *65*(1), 187–207. https://doi.org/10.1146/annurev-psych-010213-115202

Bliss-Moreau, E., Barrett, L. F., & Wright, C. I. (2008). Individual Differences in Learning the Affective Value of Others Under Minimal Conditions. *Emotion*, *8*(4), 479–493. https://doi.org/10.1037/1528-3542.8.4.479

Bos, W. Van Den, Westenberg, M., Dijk, E. Van, & Crone, E. A. (2010). Cognitive Development Development of trust and reciprocity in adolescence. *Cognitive Development*, *25*(1), 90–102. https://doi.org/10.1016/j.cogdev.2009.07.004

Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology*, *82*, 64–73. https://doi.org/https://doi.org/10.1016/j.jesp.2019.01.003

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. *Advances in Experimental Social Psychology*, *64*(January), 187–262. https://doi.org/10.1016/bs.aesp.2021.03.001

Calder, A. J., Lawrence, A. D., Keane, J., Scott, S. K., Owen, A. M., Christoffels, I., & Young, A. W. (2002). Reading the mind from eye gaze. *Neuropsychologia*, *40*(8), 1129–1138. https://doi.org/https://doi.org/10.1016/S0028-3932(02)00008-8

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences*, *7*(5), 225–231. https://doi.org/https://doi.org/10.1016/S1364-6613(03)00094-9

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M.,

Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campellone, T. R., & Kring, A. M. (2013). Who do you trust? The impact of facial emotion and behaviour on decision making. *Cognition and Emotion*, *27*(4), 603–620. https://doi.org/10.1080/02699931.2012.726608

Capozzi, F., & Ristic, J. (2018). How attention gates social interactions. *Annals of the New York Academy of Sciences*, *1426*(1), 179–198. https://doi.org/https://doi.org/10.1111/nyas.13854

Carlson, D. E., & Mae, L. (2003). The accidental tourist: Capturing incidental (versus intentional) impressions. In *Foundations of Social Cognition: A Festschrift in Honor of Robert S. Wyer, Jr.* (pp. 97–130). Lawrence Erlbaum Associates Publishers.

Carlston, D. E., McCall, T. C., McCarty, M. K., & Tay, L. (2015). On being judged by the company you keep: The effects of group consensus and target behavior on impressions of individual group members. *Journal of Experimental Social Psychology*, *60*, 173–182. https://doi.org/10.1016/j.jesp.2015.06.001

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. https://doi.org/10.1016/j.cogpsych.2010.03.001

Chelnokova, O., Laeng, B., Eikemo, M., Riegels, J., Løseth, G., Maurud, H., Willoch, F., & Leknes, S. (2014). Rewards of beauty: the opioid system mediates social motivation in humans. *Molecular Psychiatry*, *19*(7), 746–747. https://doi.org/10.1038/mp.2014.1

Choudhury, S., Blakemore, S.-J., & Charman, T. (2006). Social cognitive development during adolescence. *Social Cognitive and Affective Neuroscience*, *1*(3), 165–174. https://doi.org/10.1093/scan/nsl024

Cline, M. G. (1956). The Influence of Social Context on the Perception of Faces. *Journal of Personality*, *25*(2), 142–158. https://doi.org/10.1111/j.1467-6494.1956.tb01294.x

Colombatto, C., Chen, Y. C., & Scholl, B. J. (2020). Gaze deflection reveals how gaze cueing is tuned to extract the mind behind the eyes. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(33), 19825–19829. https://doi.org/10.1073/PNAS.2010841117

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. In *Advances in experimental social psychology*. (pp. 131–199). Elsevier Academic Press.

Cox, S. M. L., Andrade, A., & Johnsrude, I. S. (2005). Learning to Like: A Role for Human Orbitofrontal Cortex in Conditioned Reward. *The Journal of Neuroscience*, *25*(10), 2733 LP – 2740. https://doi.org/10.1523/JNEUROSCI.3360-04.2005

Crano, W. D. (1977). Primacy versus recency in retention of information and opinion change. In *The Journal of Social Psychology* (Vol. 101, Issue 1, pp. 87–96). Heldref Publications. https://doi.org/10.1080/00224545.1977.9923987

Crone, E. A., & Dahl, R. E. (2012). Understanding adolescence as a period of social–affective engagement and goal flexibility. *Nature Reviews Neuroscience*, *13*(9), 636–650. https://doi.org/10.1038/nrn3313

Cuddy, A. J. C., Fiske, S. T., & Glick, P. B. T.-A. in E. S. P. (2008). *Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map* (Vol. 40, pp. 61–149). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2601(07)00002-0

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dalmaso, M., Castelli, L., & Galfano, G. (2020). Social modulators of gaze-mediated orienting of attention: A review. *Psychonomic Bulletin and Review*, *27*(5), 833–855. https://doi.org/10.3758/s13423-020-01730-x

De Bruin, E. N. M., & Van Lange, P. A. M. (1999). Impression formation and cooperative behavior. *European Journal of Social Psychology*, *29*(2–3), 305–328. https://doi.org/https://doi.org/10.1002/(SICI)1099-0992(199903/05)29:2/3<305::AID-EJSP929>3.0.CO;2-R

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507. https://doi.org/10.1007/s11097-007-9076-9

De Neys, W., Hopfensitz, A., & Bonnefon, J.-F. (2015). Adolescents gradually improve at detecting trustworthiness from the facial features of unknown adults. *Journal of Economic Psychology*, *47*, 17–22. https://doi.org/https://doi.org/10.1016/j.joep.2015.01.002

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate

the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618. https://doi.org/10.1038/nn1575

Denrell, J. (2005). Why Most People Disapprove of Me: Experience Sampling in Impression Formation. In *Psychological Review* (Vol. 112, Issue 4, pp. 951–978). American Psychological Association. https://doi.org/10.1037/0033-295X.112.4.951

Derks, J., Lee, N. C., & Krabbendam, L. (2014). Adolescent trust and trustworthiness: Role of gender and social value orientation. *Journal of Adolescence*, *37*(8), 1379–1386. https://doi.org/https://doi.org/10.1016/j.adolescence.2014.09.014

Doherty, M. J. (2006). The development of mentalistic gaze understanding. *Infant and Child Development*, *15*(2), 179–186. https://doi.org/https://doi.org/10.1002/icd.434

Dotsch, R., Hassin, R. R., & Todorov, A. (2017). Statistical learning shapes face evaluation. *Nature Human Behaviour*, *1*(1), 1–6. https://doi.org/10.1038/s41562-016-0001

Dreben, E. K., Fiske, S. T., & Hastie, R. (1979). The independence of evaluative and item information: Impression and recall order effects in behavior-based impression formation. In *Journal of Personality and Social Psychology* (Vol. 37, Issue 10, pp. 1758–1768). American Psychological Association. https://doi.org/10.1037/0022-3514.37.10.1758

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, *6*(5), 509–540. https://doi.org/10.1080/135062899394920

Ecker, U. K. H., & Rodricks, A. E. (2020). Do False Allegations Persist? Retracted Misinformation Does Not Continue to Influence Explicit Person Impressions. *Journal of Applied Research in Memory and Cognition*, *9*(4), 587–601. https://doi.org/https://doi.org/10.1016/j.jarmac.2020.08.003

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. Neuroscience and Biobehavioral Reviews. *Neuroscience and Biobehavioral Reviews*, *24*, 581–604.

Engell, A. D., Haxby, J. V, & Todorov, A. (2007). Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *Journal of Cognitive Neuroscience*, *19*(9), 1508–1519. https://doi.org/10.1162/jocn.2007.19.9.1508

Erber, R., & Fiske, S. T. (1984). Outcome dependency and attention to inconsistent information. In *Journal of Personality and Social Psychology* (Vol. 47, Issue 4, pp. 709–726). American

Psychological Association. https://doi.org/10.1037/0022-3514.47.4.709

Evans, A. M., Athenstaedt, U., & Krueger, J. I. (2013). The development of trust and altruism during childhood. *Journal of Economic Psychology*, *36*, 82–95. https://doi.org/https://doi.org/10.1016/j.joep.2013.02.010

Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2015). Perceived trustworthiness of faces drives trust behaviour in children. *Developmental Science*, *18*(2), 327–334. https://doi.org/https://doi.org/10.1111/desc.12218

Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The Robustness of Learning about the Trustworthiness of Other People. *Social Cognition*, *33*(5), 368–386. https://doi.org/10.1521/soco.2015.33.5.368

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, *6*(OCT), 1–17. https://doi.org/10.3389/fnins.2012.00148

Farroni, T., Johnson, M. H., Brockbank, M., & Simion, F. (2000). Infants' use of gaze direction to cue attention: The importance of perceived motion. *Visual Cognition*, *7*(6), 705–718. https://doi.org/10.1080/13506280050144399

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and How Implicit First Impressions Can Be Updated. *Current Directions in Psychological Science*, *28*(4), 331–336. https://doi.org/10.1177/0963721419835206

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org/10.1037/0022-3514.38.6.889

Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from Daguerreotype to laserphoto. In *Journal of Personality and Social Psychology* (Vol. 63, Issue 6, pp. 877–889). American Psychological Association. https://doi.org/10.1037/0022-3514.63.6.877

Fiske, S. T. (1993). Social Cognition and Social Perception. *Annual Review of Psychology*, *44*(1), 155–194. https://doi.org/10.1146/annurev.ps.44.020193.001103

Forgas, J. P. (2011). Can negative affect eliminate the power of first impressions? Affective

influences on primacy and recency effects in impression formation. *Journal of Experimental Social Psychology*, *47*(2), 425–429. https://doi.org/10.1016/j.jesp.2010.11.005

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational Priors Magnify Striatal Responses to Violations of Trust. *The Journal of Neuroscience*, *33*(8), 3602 LP – 3611. https://doi.org/10.1523/JNEUROSCI.3086-12.2013

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. In *Psychological Review* (Vol. 118, Issue 2, pp. 247–279). American Psychological Association. https://doi.org/10.1037/a0022327

Freire, A., Eskritt, M., & Lee, K. (2004). Are Eyes Windows to a Deceiver's Soul? Children's Use of Another's Eye Gaze Cues in a Deceptive Situation. In *Developmental Psychology* (Vol. 40, Issue 6, pp. 1093–1104). American Psychological Association. https://doi.org/10.1037/0012-1649.40.6.1093

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin and Review*, *5*(3), 490–495. https://doi.org/10.3758/BF03208827

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694–724. https://doi.org/10.1037/0033-2909.133.4.694

Frischen, A., & Tipper, S. P. (2004). Orienting attention via observed gaze shift evokes longer term inhibitory effects: implications for social interactions, attention, and memory. *Journal of Experimental Psychology. General*, *133*(4), 516–533. https://doi.org/10.1037/0096-

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain Research*, *1079*(1), 36–46. https://doi.org/https://doi.org/10.1016/j.brainres.2005.12.126

Frith, C. D., & Frith, U. (2007). Social Cognition in Humans. *Current Biology*, *17*(16), R724–R732. https://doi.org/https://doi.org/10.1016/j.cub.2007.05.068

Frith, C. D., & Frith, U. (2008). Implicit and Explicit Processes in Social Cognition. *Neuron*, *60*(3), 503–510. https://doi.org/https://doi.org/10.1016/j.neuron.2008.10.032

Frith, C. D., & Frith, U. (2011). Mechanisms of Social Cognition. *Annual Review of Psychology*, *63*(1), 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Frith, C. D., & Singer, T. (2008). The role of social cognition in decision making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3875–3886.

https://doi.org/10.1098/rstb.2008.0156

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Galfano, G., Dalmaso, M., Marzoli, D., Pavan, G., Coricelli, C., & Castelli, L. (2012). Eye gaze cannot be ignored (but neither can arrows). *Quarterly Journal of Experimental Psychology*, *65*(10), 1895–1910. https://doi.org/10.1080/17470218.2012.663765

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. In *Psychological Bulletin* (Vol. 117, Issue 1, pp. 21–38). American Psychological Association. https://doi.org/10.1037/0033-2909.117.1.21

Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition*, *136*, 359–364. https://doi.org/https://doi.org/10.1016/j.cognition.2014.11.040

Goldberg, M. C., Mostow, A. J., Vecera, S. P., Larson, J. C. G., Mostofsky, S. H., Mahone, E. M., & Denckla, M. B. (2008). Evidence for Impairments in Using Static Line Drawings of Eye Gaze Cues to Orient Visual-Spatial Attention in Children with High Functioning Autism. *Journal of Autism and Developmental Disorders*, *38*(8), 1405–1413. https://doi.org/10.1007/s10803-007-0506-x

Goodwin, G. P. (2015). Moral Character in Person Perception. *Current Directions in Psychological Science*, *24*(1), 38–44. https://doi.org/10.1177/0963721414550709

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Goodwin, G., Piazza, J., & Rozin, P. (2013). Moral Character Predominates in Person Perception and Evaluation. *Journal of Personality and Social Psychology*, *106*. https://doi.org/10.1037/a0034726

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295x.102.1.4

Gregory, S. E. A., & Jackson, M. C. (2017). Joint attention enhances visual working memory. In

Journal of Experimental Psychology: Learning, Memory, and Cognition (Vol. 43, Issue 2, pp. 237–249). American Psychological Association. https://doi.org/10.1037/xlm0000294

Güroğlu, B., van den Bos, W., & Crone, E. A. (2014). Sharing and giving across adolescence: an experimental study examining the development of prosocial behavior . In *Frontiers in Psychology* (Vol. 5, p. 291).

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388. https://doi.org/https://doi.org/10.1016/0167-2681(82)90011-7

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, *88*, 103948. https://doi.org/https://doi.org/10.1016/j.jesp.2019.103948

Hamilton, A. F. de C. (2016). Gazing at me: the importance of social meaning in understanding direct-gaze cues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150080. https://doi.org/10.1098/rstb.2015.0080

Harris, L. T., & Fiske, S. T. (2010). Neural regions that underlie reinforcement learning are also active for social expectancy violations. *Social Neuroscience*, *5*(1), 76–91. https://doi.org/10.1080/17470910903135825

Heerey, E. A., & Velani, H. (2010). Implicit learning of social predictions. *Journal of Experimental Social Psychology*, *46*(3), 577–581. https://doi.org/https://doi.org/10.1016/j.jesp.2010.01.003

Hendrick, C. (1972). Effects of salience of stimulus inconsistency on impression formation. In *Journal of Personality and Social Psychology* (Vol. 22, Issue 2, pp. 219–222). American Psychological Association. https://doi.org/10.1037/h0032599

Hendrick, C., & Costantini, A. F. (1970). Effects of varying trait inconsistency and response requirements on the primacy effect in impression formation. In *Journal of Personality and Social Psychology* (Vol. 15, Issue 2, pp. 158–164). American Psychological Association. https://doi.org/10.1037/h0029203

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly Punishment Across Human Societies. *Science*, *312*(5781), 1767 LP – 1770. https://doi.org/10.1126/science.1127333

Hermens, F., & Walker, R. (2010). Gaze and Arrow Distractors Influence Saccade Trajectories Similarly. *Quarterly Journal of Experimental Psychology*, *63*(11), 2120–2140. https://doi.org/10.1080/17470211003718721

Hietanen, J. K. (2018). Affective Eye Contact: An Integrative Review . In *Frontiers in Psychology* (Vol. 9, p. 1587).

Hirozawa, P. Y., Karasawa, M., & Matsuo, A. (2020). Intention matters to make you (im)moral: Positive-negative asymmetry in moral character evaluations. *The Journal of Social Psychology*, *160*(4), 401–415. https://doi.org/10.1080/00224545.2019.1653254

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. https://doi.org/https://doi.org/10.1016/0010-0285(92)90002-J

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's Eyes Trigger Shifts of Visual Attention in Human Infants. *Psychological Science*, *9*(2), 131–134. https://doi.org/10.1111/1467-

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive and Affective Neuroscience*, *12*(1), 49–60. https://doi.org/10.1093/scan/nsw147

Itier, R. J., & Batty, M. (2009). Neural bases of eye and gaze processing: the core of social cognition. *Neuroscience and Biobehavioral Reviews*, *33*(6), 843–863. https://doi.org/10.1016/j.neubiorev.2009.02.004

Jones, E. E., & Davis, K. E. (1965). *From Acts To Dispositions The Attribution Process In Person Perception11Much of the research reported herein was supported by National Science Foundation Grants 8857 and 21955 to the first author.* (L. B. T.-A. in E. S. P. Berkowitz (ed.); Vol. 2, pp. 219–266). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2601(08)60107-0

Jones, E. E., & Goethals, G. R. (1987). Order effects in impression formation: Attribution context and the nature of the entity. In *Attribution: Perceiving the causes of behavior.* (pp. 27–46). Lawrence Erlbaum Associates, Inc.

Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. In *Journal of Personality and Social Psychology* (Vol. 10, Issue 4, pp. 317–340). American Psychological Association. https://doi.org/10.1037/h0026818

Kail, R. V. (2003). Information processing and memory. In *Well-being: Positive development across the life course.* (pp. 269–279). Lawrence Erlbaum Associates Publishers.

Kaisler, R. E., & Leder, H. (2016). Trusting the Looks of Others: Gaze Effects of Faces in Social Settings. *Perception*, *45*(8), 875–892. https://doi.org/10.1177/0301006616643678

Kaland, N., Smith, L., & Mortensen, E. L. (2007). Response Times of Children and Adolescents with Asperger Syndrome on an 'Advanced' Test of Theory of Mind. *Journal of Autism and Developmental Disorders*, *37*(2), 197–209. https://doi.org/10.1007/s10803-006-0152-8

Kilford, E. J., Garrett, E., & Blakemore, S.-J. (2016). The development of social cognition in adolescence: An integrated perspective. *Neuroscience & Biobehavioral Reviews*, *70*, 106–120. https://doi.org/https://doi.org/10.1016/j.neubiorev.2016.08.016

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science*, *308*(5718), 78 LP – 83. https://doi.org/10.1126/science.1108062

Kinzler, K. D., & Shutts, K. (2008). Memory for "mean" over "nice": The influence of threat on children's face memory. *Cognition*, *107*(2), 775–783. https://doi.org/https://doi.org/10.1016/j.cognition.2007.09.005

Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance? In *Journal of Personality and Social Psychology* (Vol. 111, Issue 5, pp. 655–664). American Psychological Association. https://doi.org/10.1037/pspa0000062

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147. https://doi.org/https://doi.org/10.1016/S1364-6613(00)01452-2

Kuhn, G., & Kingstone, A. (2009). Look away! Eyes and arrows engage oculomotor responses automatically. *Attention, Perception, & Psychophysics*, *71*(2), 314–327. https://doi.org/10.3758/APP.71.2.314

Kylliäinen, A., & Hietanen, J. K. (2004). Attention orienting by another's gaze direction in children with autism. *Journal of Child Psychology and Psychiatry*, *45*(3), 435–444. https://doi.org/https://doi.org/10.1111/j.1469-7610.2004.00235.x

Lammers, J., Gast, A., Unkelbach, C., & Galinsky, A. D. (2018). Moral Character Impression Formation Depends on the Valence Homogeneity of the Context. *Social Psychological and Personality Science*, *9*(5), 576–585. https://doi.org/10.1177/1948550617714585

Landy, J. F., Piazza, J., & Goodwin, G. P. (2018). Morality traits still dominate in forming impressions of others. *Proceedings of the National Academy of Sciences*, *115*(25), E5636 LP-E5636. https://doi.org/10.1073/pnas.1807096115

Langton, S. R. H., & Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Visual Cognition*, *6*(5), 541–567. https://doi.org/10.1080/135062899394939

Lee, K., Eskritt, M., Symons, L. A., & Muir, D. (1998). Children's use of triadic eye gaze information for "mind reading." In *Developmental Psychology* (Vol. 34, Issue 3, pp. 525–539). American Psychological Association. https://doi.org/10.1037/0012-1649.34.3.525

Lee, N. C., Jolles, J., & Krabbendam, L. (2016). Social information influences trust behaviour in adolescents. *Journal of Adolescence*, *46*, 66–75. https://doi.org/https://doi.org/10.1016/j.adolescence.2015.10.021

Lee, V. K., & Harris, L. T. (2013). How social cognition can inform social decision making. *Frontiers in Neuroscience*, *7*, 259. https://doi.org/10.3389/fnins.2013.00259

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, *53*, 107–117. https://doi.org/https://doi.org/10.1016/j.jesp.2014.03.005

Li, M., Mai, Z., Wang, S., Feng, T., Van Overwalle, F., & Ma, N. (2021). Warmth is more influential than competence: an fMRI repetition suppression study. *Brain Imaging and Behavior*, *15*(1), 266–275. https://doi.org/10.1007/s11682-019-00254-w

Li, T., Liu, X., Pan, J., & Zhou, G. (2017). The interactive effect of facial appearance and behavior statement on trust belief and trust behavior. *Personality and Individual Differences*, *117*, 60–65. https://doi.org/https://doi.org/10.1016/j.paid.2017.05.038

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska directed emotional faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet*, *91*(630), 2.

Macrae, C. N., Hood, B. M., Milne, A. B., Rowe, A. C., & Mason, M. F. (2002). Are you looking at me? Eye gaze and person perception. *Psychological Science*, *13*(5), 460–464. https://doi.org/10.1111/1467-9280.00481

Macrae, C. N., & Quadflieg, S. (2010). Perceiving people. In *Handbook of social psychology, Vol. 1, 5th ed.* (pp. 428–463). John Wiley & Sons, Inc.

https://doi.org/10.1002/9780470561119.socpsy001012

Mailath, G. (1987). Incentive Compatibility in Signaling Games with a Continuum of Types. *Econometrica*, *55*(6), 1349–1365.

Malle, B. F. (1999). How People Explain Behavior: A New Theoretical Framework. *Personality and Social Psychology Review*, *3*(1), 23–48. https://doi.org/10.1207/s15327957pspr0301_2

Malle, B. F. (2004). How the mind explains behavior: Folk explanations, meaning, and social interaction. In *How the mind explains behavior: Folk explanations, meaning, and social interaction.* (pp. viii, 314–viii, 314). MIT Press.

Malle, B. F. (2011). Attribution theories: How people make sense of behavior. In *Theories in social psychology.* (pp. 72–95). Wiley Blackwell.

Malle, B. F., & Knobe, J. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121. https://doi.org/https://doi.org/10.1006/jesp.1996.1314

Malle, B., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, *102 4*, 661–684.

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. In *Journal of Personality and Social Psychology* (Vol. 108, Issue 6, pp. 823–849). American Psychological Association. https://doi.org/10.1037/pspa0000021

Manssuer, L. R., Pawling, R., Hayes, A. E., & Tipper, S. P. (2016). The role of emotion in learning trustworthiness from eye-gaze: Evidence from facial electromyography. *Cognitive Neuroscience*, *7*(1–4), 82–102. https://doi.org/10.1080/17588928.2015.1085374

Manssuer, L. R., Roberts, M. V, & Tipper, S. P. (2015). The late positive potential indexes a role for emotion during learning of trust from eye-gaze cues. In *Social Neuroscience* (Vol. 10, Issue 6, pp. 635–650). Taylor & Francis. https://doi.org/10.1080/17470919.2015.1017114

Marotta, A., Lupiáñez, J., Román-Caballero, R., Narganes-Pineda, C., & Martín-Arévalo, E. (2019). Are eyes special? Electrophysiological and behavioural evidence for a dissociation between eye-gaze and arrows attentional mechanisms. *Neuropsychologia*, *129*(March), 146–152. https://doi.org/10.1016/j.neuropsychologia.2019.03.017

Marotta, A., Román-Caballero, R., & Lupiáñez, J. (2018). Arrows don't look at you: Qualitatively different attentional mechanisms triggered by gaze and arrows. *Psychonomic*

Bulletin & Review, *25*(6), 2254–2259. https://doi.org/10.3758/s13423-018-1457-2

Mason, M., Hood, B., & Macrae, C. N. (2004). Look into my eyes: Gaze direction and person memory. *Memory*, *12*(5), 637–643. https://doi.org/10.1080/09658210344000152

Mattavelli, G., Sormaz, M., Flack, T., Asghar, A. U. R., Fan, S., Frey, J., Manssuer, L., Usten, D., Young, A. W., & Andrews, T. J. (2014). Neural responses to facial expressions support the role of the amygdala in processing threat. *Social Cognitive and Affective Neuroscience*, *9*(11), 1684–1689. https://doi.org/10.1093/scan/nst162

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/https://doi.org/10.1016/j.jml.2017.01.001

Maurer, C., Chambon, V., Bourgeois-Gironde, S., Leboyer, M., & Zalla, T. (2018). The influence of prior reputation and reciprocity on dynamic trust-building in adults with and without autism spectrum disorder. *Cognition*, *172*, 1–10. https://doi.org/https://doi.org/10.1016/j.cognition.2017.11.007

McCarthy, A., & Lee, K. (2009). Children's knowledge of deceptive gaze cues and its relation to their actual lying behavior. *Journal of Experimental Child Psychology*, *103*(2), 117–134. https://doi.org/https://doi.org/10.1016/j.jecp.2008.06.005

Mende-Siedlecki, P. (2018). Changing our minds: the neural bases of dynamic impression updating. *Current Opinion in Psychology*, *24*, 72–76. https://doi.org/10.1016/j.copsyc.2018.08.007

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience*, *33*(50), 19406–19415. https://doi.org/10.1523/JNEUROSCI.2334-13.2013

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631. https://doi.org/10.1093/scan/nss040

Mende-Siedlecki, P., & Todorov, A. (2016). Neural dissociations between meaningful and mere inconsistency in impression updating. *Social Cognitive and Affective Neuroscience*, *11*(9), 1489–1500. https://doi.org/10.1093/scan/nsw058

Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. In *Developmental Psychology* (Vol. 49, Issue 3, pp. 404–418). American

Psychological Association. https://doi.org/10.1037/a0029500

Moore, C. D. (2015). Impression Formation. In *The Blackwell Encyclopedia of Sociology*. https://doi.org/https://doi.org/10.1002/9781405165518.wbeosi025.pub2

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. In *Journal of Personality and Social Psychology* (Vol. 53, Issue 3, pp. 431–444). American Psychological Association. https://doi.org/10.1037/0022-3514.53.3.431

Newman, L. S., & Uleman, J. S. (1990). Assimilation and Contrast Effects in Spontaneous Trait Inference. *Personality and Social Psychology Bulletin*, *16*(2), 224–240. https://doi.org/10.1177/0146167290162004

Neys, W. De, Hopfensitz, A., & Bonnefon, J. (2015). Adolescents gradually improve at detecting trustworthiness from the facial features of unknown adults. *JOURNAL OF ECONOMIC PSYCHOLOGY*, *47*, 17–22. https://doi.org/10.1016/j.joep.2015.01.002

Nisbett, R. E., & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment* (Issue 3). Prentice-Hall.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, *12*(7), 265–272. https://doi.org/https://doi.org/10.1016/j.tics.2008.03.006

North, M. S., & Fiske, S. T. (2012). A history of social cognition. In *Handbook of the history of social psychology*. (pp. 81–99). Psychology Press.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. https://doi.org/https://doi.org/10.1016/j.tics.2014.09.007

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087 LP – 11092. https://doi.org/10.1073/pnas.0805664105

Open, S. C. (2015). Psychology. Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pakrashi, M., Oosterhof, N. N., & Todorov, A. (2009). Evaluating Faces on Trustworthiness After Minimal Time Exposure. *Social Cognition*, *27*(6), 813–833.

Patton, G. C., Sawyer, S. M., Santelli, J. S., Ross, D. A., Afifi, R., Allen, N. B., Arora, M., Azzopardi, P., Baldwin, W., Bonell, C., Kakuma, R., Kennedy, E., Mahon, J., McGovern,

T., Mokdad, A. H., Patel, V., Petroni, S., Reavley, N., Taiwo, K., … Viner, R. M. (2016). Our future: a Lancet commission on adolescent health and wellbeing. *The Lancet*, *387*(10036), 2423–2478. https://doi.org/10.1016/S0140-6736(16)00579-1

Pfeiffer, U. J., Vogeley, K., & Schilbach, L. (2013). From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, *37*(10, Part 2), 2516–2528. https://doi.org/https://doi.org/10.1016/j.neubiorev.2013.07.017

Pfeiffer, U., Timmermans, B., Vogeley, K., Frith, C., & Schilbach, L. (2013). Towards a neuroscience of social interaction . In *Frontiers in Human Neuroscience* (Vol. 7, p. 22).

Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010). Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences*, *107*(29), 13099 LP – 13104. https://doi.org/10.1073/pnas.1008137107

Posner, M. I. (1980). Orienting of attention. *Journal of Experimental Psychology*, *32*(July 1979), 3–25.

Rabin, M., & Schrag, J. L. (1999). First Impressions Matter: A Model of Confirmatory Bias*. *The Quarterly Journal of Economics*, *114*(1), 37–82. https://doi.org/10.1162/003355399555945

Redcay, E., & Moraczewski, D. (2020). Social cognition in context: A naturalistic imaging approach. *NeuroImage*, *216*, 116392. https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.116392

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Reeder, G. (2009). Mindreading: Judgments About Intentionality and Motives in Dispositional Inference. *Psychological Inquiry*, *20*(1), 1–18. https://doi.org/10.1080/10478400802615744

Reeder, G. D., & Coovert, M. D. (1986). Revising an Impression of Morality. *Social Cognition*, *4*(1), 1–17. https://doi.org/10.1521/soco.1986.4.1.1

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional attribution: Multiple inferences about motive-related traits. In *Journal of Personality and Social Psychology* (Vol. 86, Issue 4, pp. 530–544). American Psychological Association. https://doi.org/10.1037/0022-3514.86.4.530

Renfrew, C., Frith, C., Malafouris, L., & Frith, C. D. (2008). Social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1499), 2033–2039. https://doi.org/10.1098/rstb.2008.0005

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.

Ricciardelli, P., Bricolo, E., Aglioti, S. M., & Chelazzi, L. (2002). My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual's gaze. *NeuroReport*, *13*(17).

Ricciardelli, P., Carcagno, S., Vallar, G., & Bricolo, E. (2013). Is gaze following purely reflexive or goal-directed instead? Revisiting the automaticity of orienting attention by gaze cues. *Experimental Brain Research*, *224*(1), 93–106. https://doi.org/10.1007/s00221-012-3291-5

Richey, M. H., Koenigs, R. J., Richey, H. W., & Fortin, R. (1975). Negative Salience in Impressions of Character: Effects of Unequal Proportions of Positive and Negative Information. *The Journal of Social Psychology*, *97*(2), 233–241. https://doi.org/10.1080/00224545.1975.9923343

Rilling, J. K., & Sanfey, A. G. (2010). The Neuroscience of Social Decision-Making. *Annual Review of Psychology*, *62*(1), 23–48. https://doi.org/10.1146/annurev.psych.121208.131647

Rogers, R. D., Bayliss, A. P., Szepietowska, A., Dale, L., Reeder, L., Pizzamiglio, G., Czarna, K., Wakeley, J., Cowen, P. J., & Tipper, S. P. (2014). I want to help you, but i am not sure why: Gaze-cuing induces altruistic giving. *Journal of Experimental Psychology: General*, *143*(2), 763–777. https://doi.org/10.1037/a0033677

Ross, L., & Nisbett, R. E. (1991). The person and the situation: Perspectives of social psychology. In *The person and the situation: Perspectives of social psychology.* (pp. xvi, 286–xvi, 286). Mcgraw-Hill Book Company.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, *23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. In *Journal of Personality and Social Psychology* (Vol. 104, Issue 3, pp. 409–426). American Psychological Association. https://doi.org/10.1037/a0031050

Rule, N. O., Slepian, M. L., & Ambady, N. (2012). A memory advantage for untrustworthy faces. *Cognition*, *125*(2), 207–218. https://doi.org/10.1016/j.cognition.2012.06.017

Sanfey, A. G. (2007). Social decision-making: insights from game theory and neuroscience. *Science (New York, N.Y.)*, *318*(5850), 598–602. https://doi.org/10.1126/science.1142996

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/DOI: 10.1017/S0140525X12000660

Senju, A., & Johnson, M. H. (2009). The eye contact effect: mechanisms and development. *Trends in Cognitive Sciences*, *13*(3), 127–134. https://doi.org/https://doi.org/10.1016/j.tics.2008.11.009

Senju, A., Tojo, Y., Dairoku, H., & Hasegawa, T. (2004). Reflexive orienting in response to eye gaze and an arrow in children with and without autism. *Journal of Child Psychology and Psychiatry*, *45*(3), 445–458. https://doi.org/https://doi.org/10.1111/j.1469-7610.2004.00236.x

Shen, X., Mann, T. C., & Ferguson, M. J. (2020). Beware a dishonest face?: Updating face-based implicit impressions using diagnostic behavioral information. *Journal of Experimental Social Psychology*, *86*(July 2019), 103888. https://doi.org/10.1016/j.jesp.2019.103888

Shepherd, S. V. (2010). Following gaze: Gaze-following behavior as a window into social cognition. *Frontiers in Integrative Neuroscience*, *4*(MARCH 2010), 1–13. https://doi.org/10.3389/fnint.2010.00005

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756. https://doi.org/10.1038/s41562-018-0425-1

Simpson, D. D., & Ostrom, T. M. (1976). Contrast effects in impression formation. *Journal of Personality and Social Psychology*, *34*(4), 625–629. https://doi.org/10.1037/0022-3514.34.4.625

Sims, T. B., Van Reekum, C. M., Johnstone, T., & Chakrabarti, B. (2012). How reward

modulates mimicry: EMG evidence of greater facial mimicry of more rewarding happy faces. *Psychophysiology*, *49*(7), 998–1004. https://doi.org/https://doi.org/10.1111/j.1469-8986.2012.01377.x

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain Responses to the Acquired Moral Status of Faces. *Neuron*, *41*(4), 653–662. https://doi.org/10.1016/S0896-6273(04)00014-5

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Smith, E. R., & Semin, G. R. (2004). Socially Situated Cognition: Cognition in its Social Context. In *Advances in experimental social psychology, Vol. 36.* (pp. 53–117). Elsevier Academic Press. https://doi.org/10.1016/S0065-2601(04)36002-8

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. In *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Vol. 35, Issue 3, pp. 780–805). American Psychological Association. https://doi.org/10.1037/a0015145

Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences*, *9*(2), 69–74. https://doi.org/https://doi.org/10.1016/j.tics.2004.12.005

Stewart, R. H. (1965). Effect of continuous responding on the order effect in personality impression formation. *Journal of Personality and Social Psychology*, *1*(2), 161–165. https://doi.org/10.1037/h0021641

Strachan, J., & Tipper, S. (2015). Using scalar ratings to track changes in apparent trustworthiness induced by helpful and misleading gaze cues. *Journal of Vision*, *15*(12), 1216. https://doi.org/10.1167/15.12.1216

Strachan, J. W. A., Guttesen, A. á V., Smith, A. K., Gaskell, M. G., Tipper, S. P., & Cairney, S. A. (2020). Investigating the formation and consolidation of incidentally learned trust. In *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Vol. 46, Issue 4, pp. 684–698). American Psychological Association. https://doi.org/10.1037/xlm0000752

Strachan, J. W. A., Kirkham, A. J., Manssuer, L. R., Over, H., & Tipper, S. P. (2017). Incidental learning of trust from eye-gaze: Effects of race and facial trustworthiness. *Visual Cognition*, *25*(7–8), 802–814. https://doi.org/10.1080/13506285.2017.1338321

Strachan, J. W. A., Kirkham, A. J., Manssuer, L. R., & Tipper, S. P. (2016). Incidental learning of trust: Examining the role of emotion and visuomotor fluency. In *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Vol. 42, Issue 11, pp. 1759–1773). American Psychological Association. https://doi.org/10.1037/xlm0000270

Strachan, J. W. A., & Tipper, S. P. (2017). Examining the durability of incidentally learned trust from gaze cues. *Quarterly Journal of Experimental Psychology*, *70*(10), 2060–2075. https://doi.org/10.1080/17470218.2016.1220609

Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior*, *59*(2), 364–382. https://doi.org/https://doi.org/10.1016/j.geb.2006.07.006

Swettenham, J., Condie, S., Campbell, R., Milne, E., & Coleman, M. (2003). Does the perception of moving eyes trigger reflexive visual orienting in autism? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1430), 325–334. https://doi.org/10.1098/rstb.2002.1203

Tanida, S., Shimoma, E., Mashima, R., Ma, L., & Yamagishi, T. (2003). Photographic face recognition of cooperators vs. defectors. In *Japanese Journal of Psychology* (Vol. 74, Issue 2, pp. 148–155). Japanese Psychological Assn. https://doi.org/10.4992/jjpsy.74.148

Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. In *Judgment and Decision Making* (Vol. 11, Issue 5, pp. 527–536). Society for Judgment and Decision Making.

Thomas, L. A., De Bellis, M. D., Graham, R., & LaBar, K. S. (2007). Development of emotional facial recognition in late childhood and adolescence. *Developmental Science*, *10*(5), 547–558. https://doi.org/https://doi.org/10.1111/j.1467-7687.2007.00614.x

Tipples, J. (2008). Orienting to counterpredictive gaze and arrow cues. *Perception & Psychophysics*, *70*(1), 77–87. https://doi.org/10.3758/PP.70.1.77

Todorov, A. (2008). Evaluating Faces on Trustworthiness. *Annals of the New York Academy of Sciences*, *1124*(1), 208–224. https://doi.org/https://doi.org/10.1196/annals.1440.012

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*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Olson, I. R. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social Cognitive and Affective Neuroscience*, *3*(3), 195–203. https://doi.org/10.1093/scan/nsn013

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. https://doi.org/https://doi.org/10.1016/j.tics.2008.10.001

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/https://doi.org/10.1016/j.psychres.2008.05.006

Tulving, E. (2008). On the law of primacy. In *Memory and mind: A festschrift for Gordon H. Bower.* (pp. 31–48). Lawrence Erlbaum Associates Publishers.

Tummeltshammer, K.S., Wu, R., Sobel, D.M. and Kirkham, N.Z., 2014. Infants track the reliability of potential informants. *Psychological Science*, *25*(9), pp.1730-1738.

Uleman, J. S., Blader, S. L., & Todorov, A. (2005). Implicit impressions. In *The new unconscious.* (pp. 362–392). Oxford University Press.

Uleman, J. s., & Kressel, L. m. (2013). A Brief History of Theory and Research on Impression Formation. In *The Oxford Handbook of Social Cognition*.

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. In *Advances in experimental social psychology, Vol. 28.* (pp. 211–279). Academic Press. https://doi.org/10.1016/S0065-2601(08)60239-7

Uleman, J. S., Newman, L., & Winter, L. (1992). Can personality traits be inferred automatically? Spontaneous inferences require cognitive capacity at encoding. *Consciousness and Cognition*, *1*(1), 77–90. https://doi.org/https://doi.org/10.1016/1053-8100(92)90049-G

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*, 329–360. https://doi.org/10.1146/annurev.psych.59.103006.093707

Uleman, J. S., Saribay, S. A., Uleman, J. S., & Saribay, S. A. (2018). Initial Impressions of

Others. In *The Oxford Handbook of Personality and Social Psychology* (Issue October). https://doi.org/10.1093/oxfordhb/9780190224837.013.15

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. https://doi.org/https://doi.org/10.1016/j.cognition.2008.07.002

van den Bos, W., van Dijk, E., & Crone, E. A. (2011). Learning whom to trust in repeated social interactions: A developmental perspective. *Group Processes & Intergroup Relations*, *15*(2), 243–256. https://doi.org/10.1177/1368430211418698

van den Bos, W., Westenberg, M., van Dijk, E., & Crone, E. A. (2010). Development of trust and reciprocity in adolescence. *Cognitive Development*, *25*(1), 90–102. https://doi.org/https://doi.org/10.1016/j.cogdev.2009.07.004

Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. B. T.-A. in E. S. P. (2010). Chapter 2 - An Interpersonal Approach to Emotion in Social Decision Making: The Emotions as Social Information Model. In *Advances in Experimental Social Psychology* (Vol. 42, pp. 45–96). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2601(10)42002-X

van Rooijen, R., Junge, C., & Kemner, C. (2018). No Own-Age Bias in Children's Gaze-Cueing Effects . In *Frontiers in Psychology* (Vol. 9, p. 2484).

Vogeley, K. (2017). Two social brains: neural mechanisms of intersubjectivity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1727), 20160245. https://doi.org/10.1098/rstb.2016.0245

Vonk, R. (1994). Trait inferences, impression formation, and person memory: Strategies in processing inconsistent information about persons. *European Review of Social Psychology*, *5*(1), 111–149. https://doi.org/10.1080/14792779543000039

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*(3), 655–684. https://doi.org/https://doi.org/10.1111/1467-8624.00304

Williams, K. D., & Jarvis, B. (2006). Cyberball: A program for use in research on interpersonal ostracism and acceptance. *Behavior Research Methods*, *38*(1), 174–180. https://doi.org/10.3758/BF03192765

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. https://doi.org/10.1111/j.1467-

9280.2006.01750.x

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. In *Journal of Personality and Social Psychology* (Vol. 47, Issue 2, pp. 237–252). American Psychological Association. https://doi.org/10.1037/0022-3514.47.2.237

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the Dominance of Moral Categories in Impression Formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001

Wu, X., Hua, R., Yang, Z., & Yin, J. (2018). The influence of intention and outcome on evaluations of social interaction. *Acta Psychologica*, *182*, 75–81. https://doi.org/https://doi.org/10.1016/j.actpsy.2017.11.010

Wyer, N. A. (2010). You Never Get a Second Chance to Make a First (Implicit) Impression: The Role of Elaboration in the Formation and Revision of Implicit Impressions. *Social Cognition*, *28*(1), 1–19. https://doi.org/10.1521/soco.2010.28.1.1

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235 LP – 8240. https://doi.org/10.1073/pnas.0701408104

Yu, M., Saleem, M., & Gonzalez, C. (2014). Developing trust: First impressions and experience. *Journal of Economic Psychology*, *43*, 16–29. https://doi.org/10.1016/j.joep.2014.04.004

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. In *American Psychologist* (Vol. 35, Issue 2, pp. 151–175). American Psychological Association. https://doi.org/10.1037/0003-066X.35.2.151

Zaki, J., Hennigan, K., Weber, J., & Ochsner, K. N. (2010). Social Cognitive Conflict Resolution: Contributions of Domain-General and Domain-Specific Neural Systems. *The Journal of Neuroscience*, *30*(25), 8481 LP – 8488. https://doi.org/10.1523/JNEUROSCI.0382-10.2010

Zarolia, P., Weisbuch, M., & McRae, K. (2017). Influence of indirect information on interpersonal trust despite direct information. *Journal of Personality and Social Psychology*, *112*(1), 39–57. https://doi.org/10.1037/pspi0000074

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120. https://doi.org/DOI: 10.1017/S0140525X1700197