Proto-trust and trust attribution: a theory of intuitive, affective forms of trust and the means by which trust decisions are made

Lauren Fell (l3.fell@qut.edu.au)
Faculty of Science, Queensland University of Technology
2 George St, Brisbane, QLD, Australia

Peter Bruza (p.bruza@qut.edu.au)
Faculty of Science, Queensland University of Technology
2 George St, Brisbane, QLD, Australia

Abstract

The purpose of this paper is to present a novel conceptualisation of an intuitive, primitive form of trust termed *proto-trust*. This concept is proposed in order to account for the many different senses, types and domains in which trust has traditionally been defined and theorised. A brief review of the literature on affective and intuitive trust is presented, informing the definition and formalisation of proto-trust. Following this, a preliminary empirical investigation of proto-trust is described, where intuitive trust assessments are compared to analytical trust decisions, under various attribution prompts. Results showed effects of attribution prompts on changes to trust assessments from intuitive to deliberative decisions. In addition, qualitative data are presented for the various reasons participants gave for their trust decisions. One of these reasons (emotional reaction) was found to affect the degree of difference between intuitive and deliberative trust assessments.

Keywords: Trust, Intuition, Affect, Dual Process.

Introduction

Trust is a concept that has been infamously difficult to define (McKnight & Chervany, 2001; Shapiro, 1987; PytlikZillig & Kimbrough, 2016), with definitions ranging from trust as belief (Hardin, 2002), disposition (Ben-Ner & Halldorsson, 2010), behaviour (Castelfranchi & Falcone, 2010), and even as a cognitive bias (Yamagishi & Yamagishi, 1994). Adding to this difficulty is the tendency for definitions to be developed independently in vastly different contexts. For instance, trust has been researched in the context of interactions with friends (Falcone & Castelfranchi, 2001), colleagues (Tan & Lim, 2009), romantic partners (Rempel, Holmes, & Zanna, 1985; Campbell & Stanton, 2019), governments (Levi & Stoker, 2000), institutions (La Porta, Lopez-de Silanes, Shleifer, & Vishny, 1997; Wesson, Lucey, & Cooper, 2019), organisations (Amaral, Sales, Guizzardi, & Porello, 2019), information (Sbaffi & Rowley, 2017), media (Kiousis, 2001; Turcotte, York, Irving, Scholl, & Pingree, 2015), and technology (Mcknight, Carter, Thatcher, & Clay, 2011; De Visser, Pak, & Shaw, 2018), to name a few. Additional subtleties in the conceptual differences between different types of trust (e.g., reliability vs. benevolence) and even forms of trust (e.g., lack of trust vs. mistrust vs. distrust) have further confounded the process of defining the concept more broadly.

Whilst the conscious deliberation of these (sometimes conflicting) types of trust may involve different processes and information, the implicit *feeling* of trust experienced is intuitively the same. In other words, in all forms of trust, an individual becomes aware of signals in the environment which lead them to the conclusion that a person or entity may, or may not, be trusted, e.g., to perform a certain task well (competence), or is not entirely truthful (honesty), or does not have good intentions (benevolence). The same intuition or feeling can be felt in situations where trust is applied to non-human aspects of an experience, such as in information contained within a news article (credibility), or the number or errors produced by an automated system (reliability). In the following, we argue that this pre-conscious "alert mechanism" underpins all senses or types of trust, but where the separation of these senses arise out of a secondary attribution process. This underlying, pre-conscious mechanism will be termed *proto-trust*, where 'proto' alludes to the 'first' or 'earliest form' senses of this term. Proto-trust is theorised to be based on connections between trust and affective/intuitive thinking.

This paper will be organised as follows. First, extant theories of intuitive and emotional trust will be introduced, along with an explanation of how these forms of trust may provide a more common platform from which divergent types of conscious, deliberative trust can arise. Following this, an initial attempt at developing and testing the new theory of proto-trust will be presented.

Background

Trust as Intuition

Trust as a rational process of expectation and judgement of subjective probabilities has been well researched. However, non-rational, intuitive forms of trust are often neglected (Stoltz & Lizardo, 2018). This is, perhaps, surprising given the relatively primitive nature of trust as an evolutionarily adaptive function compared to later developed higher-order rational thinking. A clear illustration of this is in the relative speed with which humans can make judgements about the trustworthiness of other humans based on facial information - as fast as just 33ms after exposure to a face (Todorov, Pakrashi, & Oosterhof, 2009).

Automatic trust appraisals of faces Social judgements based on facial features have a long history in psychological literature. Trustworthiness has been identified as a key dimension upon which many other social impressions can be based. Oosterhof and Todorov (2008) conducted a comprehensive analysis of a range of trait judgements about facial

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

features (e.g., confidence, intelligence, attractiveness, etc.), finding that two dimensions were sufficient to explain the variance across all of these traits. These two dimensions were those represented by judgements of valance and dominance. The first factor was later changed from *valence* to *trustworthiness* based on high correlations between the trustworthiness trait and the "valence" dimension [1]. Judgements of trustworthiness have been shown to be made implicitly outside conscious control (Engell, Haxby, & Todorov, 2007; Swe et al., 2020). Judgements of trustworthiness from faces are an important adaptation for survival (Hou & Liu, 2019), and inform important behaviours such as decisions to approach or avoid a person who may represent a potential threat (Slepian, Young, & Harmon-Jones, 2017).

In studies of perceived trustworthiness of faces, contrary to most survey-based trust measures (e.g., Rempel et al. (1985)), trustworthiness is generally assessed using a single question about the degree to which someone trusts a photographed person or computer-generated face. Simple binary yes/no questions are often used (e.g., "Is this person trustworthy?") (Todorov et al., 2009; Castle et al., 2012; Günaydin, Zayas, Selcuk, & Hazan, 2012), as well as Likert scales with ratings associated with high trustworthiness to low trustworthiness (Sutherland, Young, & Rhodes, 2017; Todorov, Baron, & Oosterhof, 2008). The question "How trustworthy [is this person/does this person look]?" has also been used by Little, Roberts, Jones, and DeBruine (2012), Tingley (2014), Alrajih and Ward (2014), Oh, Dotsch, Porter, and Todorov (2019), and Lambert, Declerck, and Boone (2014).

By asking a general question of trust, rather than a series of questions specific to various types, senses and scenarios where trust might be assessed, the experimenters are allowing participants to interpret trust in a way that is most natural to them. In fact, these studies generally ask participants to answer based on their "gut feeling". A general decision of trust based on intuitive first impressions can be seen as representing a holistic culmination of all possible senses of trust. For instance, a general trust decision may arise from a holistic inference of how a person might behave in a range of future circumstances, thus adaptively priming a decision maker to respond with according caution. This general, holistic trust accords with the concept of proto-trust proposed later in this paper.

Intuitive trust as a disposition Another compelling reason to include intuition in theories of trust is related to the fact that trust often occurs without rational cause. For instance, what rational thought process could reasonably lead a young child to trust their mother? Or an oft betrayed friend to continue to trust the person responsible for their betrayal? Or indeed a person to keep believing that a piece of technology will reliably meet expectations despite many signs that it will not? This "faith" style concept of trust has appeared in various places in the literature, and has been given various terms such as *faith in humanity* (McKnight & Chervany, 2000), *default trust* (A. M. Evans, Dillon, Goldin, & Krueger, 2011), or *generalised trust* (Uslaner, 2002). These theories tend to describe it as a dispositional trust toward strangers which is said to facilitate societal interaction. Although these types of trust are often contrasted with deliberative, rational forms of trust (Freitag & Traunmüller, 2009), they still lack a clear explanation of true intuition-driven trust, particularly as it presents in situations outside interpersonal interactions.

Broader views of intuitive trust Perhaps the closest to providing a broader picture of the type of fast, intuitive trust exemplified in face literature are theories of trust that draw from dual process models of cognition. Dual process models arose in a trend in behavioural economics and (to a lesser extent) psychology, which has seen cognition split into two distinct processing styles: one fast and intuitive (System 1) and one slow and deliberate (System 2) (J. Evans, 2010; Kahneman, 2003). One example is Hermes, Behne, and Rakoczy (2018)'s use of a dual process perspective to explain early learning behaviours that require irrational trust in young children. Another is Stoltz and Lizardo (2018)'s dual process theory of trust, where a distinction is drawn between two types of trust appraisals which lead to reliance of a trustor on a trustee: intuitive and deliberative. These follow the general dual process accounts of reasoning, describing deliberative trust as involving conscious, rational decision-making based on expectations of risk and management of informational uncertainty (covered by the majority of extant trust literature), and intuitive trust as embodied, unconscious and reflexive, involving associative learning and proprioception of internal physiological and affective states. Stoltz and Lizardo (2018) refer to intuitive trust as being ever-present, even when deliberative trust is not. As they put it, "Even when certain entities in a situation invite our effortful scrutiny, our decision to deliberately trust will always incorporate and rest upon further entities, a diffuse social penumbra, which we will take on faith" (Stoltz & Lizardo, 2018, p. 244).

In everyday language, intuitions tend to share common language with affect (emotion). We tend to describe intuitions as something we "feel", for example, when describing that something "feels right" or when we might have a "funny feeling" about something. Castelfranchi (1999) describes the emotional aspect of intuitive trust as "a spontaneous, non reasonable or reasoned upon (non-rational) reliance, and a feeling of confidence in a given environment or in a person" (p. 86). This type of "feeling" or sensing of intuitive trust will help inform the basis for the theory presented in this paper.

Trust and Affect

Literature dealing with affect and trust tend to deal with the connection in one of two ways: a) affective trust as a type of trust distinct from "cognitive" trust, and b) the correlation between internal and external affective states and trust. The former tends to define affective trust in terms of a relational trust between people, rather than specifically drawing on the

[1] This highlights the strong links between intuitive thinking, valence (emotion) and trust

more general emotional content of trust appraisals. The latter represents evidence that internal emotional states impact propensity to trust (e.g., the effect of emotional valance on trust (J. R. Dunn & Schweitzer, 2005)), and that external emotional states (e.g., emotions displayed by others via facial expressions) often overlap with impressions of others' trustworthiness (Todorov, 2008).

Theories of affective trust Despite early attempts by influential trust researchers to integrate a more general emotional component into theories of trust (e.g., Lewis and Weigert (1985)), subsequent research has tended to focus on emotional features of interpersonal relationships when defining and investigating affect-based trust. This type of trust is often seen as involving mutual care and concern (McAllister, 1995), involving antecedents of shared values, long-term contact and open sharing of personal information (Chowdhury, 2005). This conceptualisation of affective trust has been used in studies in managerial trust (Chua, Ingram, & Morris, 2008), knowledge sharing (Chowdhury, 2005), team dynamics (Webber, 2008), e-commerce (Punyatoya, 2019), and human-robot interaction (Gompei & Umemuro, 2018).

A somewhat broader conceptualisation of affective trust is offered by Castelfranchi and Falcone (2010), who describe the role of emotion in trust as being an "activator" of trust-related goals. They view the affective form of trust as a "feeling, an affective response arousing from a given more or less explicit perception and appraisal of the world" (Castelfranchi & Falcone, 2010, p.141), which then serves to activate the goals relevant to trust behaviours. In a similar way to Stoltz and Lizardo (2018)'s theory of intuitive trust, this view adopts a dual process perspective. However, the nature of the interaction between affective and intuitive trust and deliberative trust has not yet been adequately addressed. The primary source of interaction between these forms of trust proposed in this paper is via an attribution process.

Attribution Conscious attributions are an important aspect of the experience of emotion. In appraisal-based theories of emotion, cognitive processing of the source and meaning of an emotionally arousing event is vital to the experience of emotion (Ellsworth & Scherer, 2003). This necessarily requires some attribution of a state to a particular stimulus. In another prominent theory of emotion, the Two-Factor Theory, attribution is explicitly outlined as necessary for physiological arousal to become an emotional state (Schachter & Singer, 1962; Shaked & Clore, 2017). Attribution and similar concepts of causal appraisals are integral to many theories of emotion. When an emotion is felt, it is almost always attributed to some event or entity in one's environment which is presumed to have lead to such a state.

In addition, the attribution process provides another link between emotion and intuition. B. D. Dunn et al. (2010) found that interoception ability (the extent to which people can perceive internal bodily functions such as heart rate) affects the capacity for intuitive reasoning, and that this effect is mediated by the emotion-based attributions made regarding what is perceived interoceptively. Considering affective and intuitive trust, it is easy to imagine situations in which one "feels" a sense of distrust, and subsequently attributes this feeling to a person or element in their immediate environment (even if this person or element was not the cause of their intuitive sense of distrust). An example of trust-based attributions in the literature is in the effect that attributing blame/distrust in AI-mediated conversations has on trust (Hohenstein & Jung, 2020).

Proto-trust

The benefit of further developing intuitive trust as a theory is in providing a common framework from which trust as a general concept can be described. A recent theory has made first attempts to do this by describing trust in terms of a fundamental predictive mechanism (Fell, Gibson, Bruza, & Hoyte, 2020). This theory, termed the Cognitive Predicting Theory of Trust (CPTT), posits that individuals have certain expectations about the world based on values and standards. These might be general, such as ethical or moral standards, or specific, such as the expectation that a person will perform a certain action or that a chair will not collapse upon being sat on. Based on these expectations, as well as expectations about the surrounding context, CPTT proposes that humans make implicit and continuous predictions about the environment with which they interact. The process of prediction is increasingly recognised as a vital part of cognition, with some theories proposing that prediction is one of the central functions governing cognition (Hohwy, 2013; Huang & Rao, 2011; Clark, 2015; Friston & Kiebel, 2009).

The reliability of predictions made via interactions with an environment can be seen as what an individual experiences as trust. When predictions produce error, trust is affected and has an increased chance of being brought to conscious awareness. If a person recognises (or attributes) these errors as stemming from the actions of a particular element of their environment, for example, a person or entity, their explicit trust in that person or entity is changed. Explicit, or deliberative trust, which is mostly assumed to be a conscious, rational process (McAllister, 1995; Falcone & Castelfranchi, 2001), can be described as a process of reducing prediction errors towards the goal of maximising utility (although, a prediction-based view of trust also allows for situations where outcomes other than the maximisation of utility are pursued by a trustor). Thus, active search for evidence to inform a rational agent's decision whether or not to trust (i.e., engaging in a causal attribution process) is captured in the aim to reduce errors in prediction.

Importantly, CPTT also accounts for affective and intuitive trust. In fact, predictive coding (one of the theories upon which CPTT was built) has been applied to affective states by considering the perception of emotion as arising through a process of interoception (internal sensing of body states) (Critchley & Garfinkel, 2017). Thus, assessing one's own implicit affect-driven trust can be considered as a perception driven by the same predictive mechanisms as are involved in deliberative trust.

Whilst Fell et al. (2020) provide a general view of trust in terms of prediction, they do not focus on the elements of intuitive or deliberative trust. Proto-trust is proposed as an extension of CPTT specifically dealing with the initial intuitive perception of predictive errors and the attribution process that brings trust into the realm of conscious, deliberative cognition.

Definition of Proto-trust

Proto-trust is thus defined as an intuitive, predictive appraisal of the degree to which one's immediate environment accords with implicit standards of safety and security, actioned via a predictive mechanism and perceived via emotional affect.

Proto-trust can project to deliberative trust decisions, attitudes and behaviours via an attribution process, whereby the affective experience of proto-trust is consciously attributed to particular elements within one's environment, e.g., a person or entity with whom one perceives the need to trust/not trust.

Examining Proto-trust

Proto-trust, like many experiential and intuitive concepts, is by its very nature difficult to study directly. In order to empirically evaluate proto-trust, therefore, one must turn to its signals. One such signature, included as part of the theory, is the effect of attribution. Just as raw emotion can be studied through various appraisal and attribution processes, it would follow that so too can proto-trust.

A fundamental feature of proto-trust is its holistic nature. This primitive modality of trust is proposed to draw more on associations than rule-based reasoning (Stoltz & Lizardo, 2018). It makes sense, then, for the first investigation into this type of trust to be focused on trust in faces. Faces are known to be processed holistically, and intuitive trust judgements of faces have been found to occur extremely rapidly, making face stimuli good candidates for studying this primitive form of trust.

Imagine that, whilst viewing a face, an individual's general sense of trust, conveyed via a feeling, is closer to trust than distrust. Suppose, then, that this person was prompted by some situational need to decide on the trustworthiness of the face in their visual field. The theory of proto-trust would predict that the person's general intuitive trust (based on the whole of their surroundings) influences their decision about the trustworthiness of a particular element of their environment (the face) by means of attribution. They attribute their general feeling of trust to the object about which they must make a trust decision. Likewise, if further need arises to attribute that trust to any part or feature of the face (e.g., kind eyes, smiling mouth, etc.), their initial assessment would also inform this.

One consequence of this process may be that mis-attribution occurs, where, for example, a particularly untrustworthy face is mis-attributed to be trustworthy due to a person's general high trust state. However, mis-attribution will not be the focus of this study. Instead, this study will focus on a possible amplification effect, whereby the act of reasoning and attributing one's feeling of trust/distrust may strengthen one's confidence in that feeling and thus result in an amplification of an intuitive trust assessment when converted to an attributed decision. This type of effect has been found in studies investigating the effect of certain emotional appraisals on moral judgements (Horberg, Oveis, & Keltner, 2011), as well as general emotional states reinforced by attention and appraisal mechanisms (Lowe, Herrera, Morse, & Ziemke, 2007).

In light of the preceding, two research questions will be addressed in this paper: 1) What is the nature of attributions of trust in faces, and 2) how do attributions alter trust in faces. Whilst the first question is largely exploratory, the hypothesis for the second question is as follows: proto-trust assessments of faces will be amplified when individuals are prompted to attribute reasons for their trust assessments.

Methodology

Participants Participants consisted of 270 members of the crowdsourcing platform Prolific, 148 of which identified as female, 120 male, and 2 who identified as non-binary/third gender. Participants were over 18 years, with a mean age of 27, and participated from a variety of countries. Remuneration was in the form of a small payment (£0.5), as per Prolific convention, and an informed consent page was presented to participants prior to commencement.

Materials Stimuli consisted of three computer generated faces adapted from Oosterhof and Todorov (2008). These faces were modified by Oosterhof and Todorov (2008) to vary on perceived trustworthiness, and the three faces used depicted high trustworthiness, low trustworthiness and neutral trustworthiness (see Fig. 1). Participants were asked a set of questions pertaining to their intuitive and deliberative trust judgements of the face shown to them (one per condition), their reasons for their trust decision, as well as demographic data.

The intuitive trust question asked "Whilst viewing the face, indicate on the below scale the level of trust you felt.", and provided a scale from 0 (lowest trust) to 100 (highest trust). This question was similar to standard questions used in previous face literature, however, was adapted to provide a more general picture of participants' overall sense of trust, as proposed by the proto-trust theory. The deliberative trust question asked "Thinking carefully, rate again the level of trust you feel when viewing the face.", again with a scale from 0 (lowest trust) to 100 (highest trust). This was again adapted from standard trust questions in face literature, but this time was adjusted to prompt participants to consider their answer more carefully. Finally, the reason-attribution question consisted of three versions: trustworthy prompt, untrustworthy prompt and neutral prompt. The question asked for the neutral prompt condition was "For the next 45 seconds, please

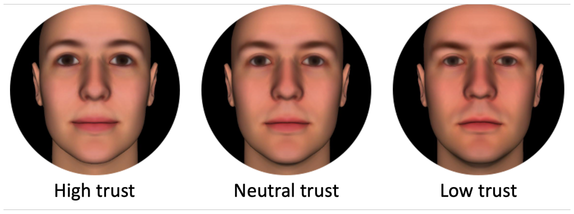

Figure 1: Three face conditions displayed to participants

write reasons why you believe this face appears *trustworthy/untrustworthy (whichever you decided in your previous answer)*. Write as many reasons as you can.". For the trustworthy prompt and untrustworthy prompt conditions, the italicised text was replaced by *trustworthy* and *untrustworthy*, respectively.

Procedure The mixed design involved two within subjects conditions across nine between subjects conditions. Within subjects conditions consisted of the two trust measures (proto- and deliberative). The between subjects conditions were represented by a 3 x 3 design, with 3 image conditions (high trustworthy face, neutral trustworthy face and low trustworthy face) across the three attribution conditions (trustworthy, neutral, untrustworthy).

Participants begun by reading the informed consent sheet, agreeing to participate and reading instructions. Following this, one face (i.e., either the high trustworthiness, low trustworthiness or neutral trustworthiness face) was presented for a total of one second. The proto-trust question was then presented, followed by the reason-attribution question (i.e., either trust prompt, untrust prompt or neutral prompt, depending on the condition). Participants were prevented from moving on from this question until 45 seconds had passed, in order to encourage an adequate reflection on the reasons for their initial trust assessment and prompt attributions. The analytic trust question followed, along with the face being again presented without time limit, and the survey ended with the collection of demographic information.

Results

Paired samples t-tests revealed that trust significantly decreased from the proto- (intuitive) trust assessment to the subsequent deliberative trust assessment for the low trustworthy face in the untrustworthy-prompted attribution condition, $t(29) = 2.39$, $p = .012$. Additionally, a difference approaching statistical significance between proto- and deliberative trust for the high trustworthy face in the trust-prompted attribution condition, $t(29) = -1.57$, $p = .064$) was found. In this condition, the trust scores increased in the post-attribution deliberative trust decision. No significant differences were found between proto- and intuitive trust scores for the remaining 7 conditions.

A difference score for proto- and deliberative trust was calculated by subtracting deliberative trust scores from proto-trust scores. Independent samples t-tests conducted on trustworthy-prompted and untrustworthy-prompted attribution conditions within each of the face conditions showed a significant difference between the two types of prompted attributions for the difference score in the low trustworthy face condition, $t(58) = -2.22$, $p = .015$, and a difference approaching significance in the high trustworthy face condition, $t(58) = -1.48$, $p = .072$. A Pearson correlation was conducted to determine if initial proto-trust scores predicted the amount of change to trust scores after the attribution process was prompted. Negative correlations were found for the low trustworthy face in the untrust-prompted and neutrally prompted attribution conditions, $r(28) = -0.60$, $p = <.001$, $r(28) = -0.42$, $p = .02$), and for the neutral face in the untrust prompted attribution condition, $r(28) = -0.48$, $p = .007$. No other conditions showed significant correlations. Graphs for these effects are shown in Fig. 2.

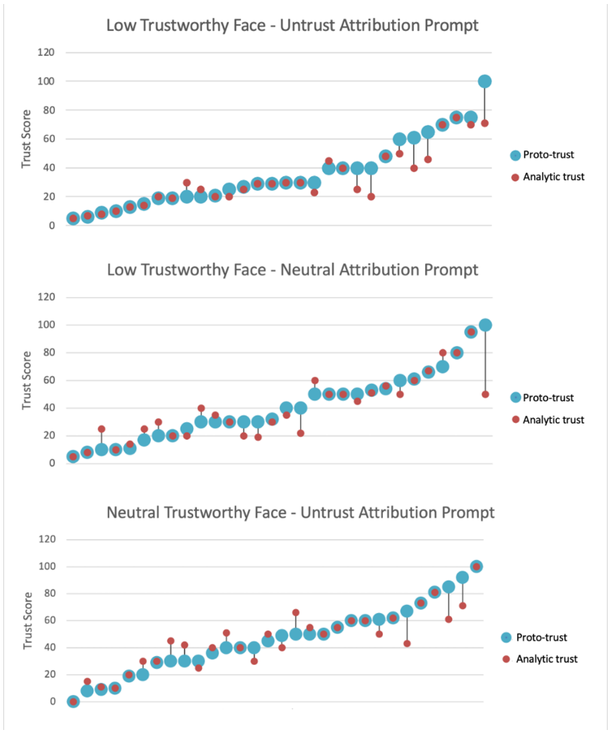

Figure 2: Changes per participant shown against initial proto-trust scores

Thematic analysis was conducted on qualitative data using an open coding method. Keywords were categorised into themes and coded for each participant response to the reason attributions question. Based on key words and themes identified by this manual inductive coding process, 7 categories emerged for the reasons reported by participants to explain their initial trust assessments. Many participants' answers fell into more than one category, so each category was coded separately. The categories for attributed reasoning were: an emotional reaction to the face (n = 46, 17%), feature(/s) of the face (n = 130, 48.1%), the fact that the face was artificial (n = 62, 23%), the perceived expression on the face (n = 142, 52%), similarity of the face to a stereotype (n = 28, 10.4%), the familiarity of the face (n = 14, 5.2%), and a general (unexplainable) sense of the face (n = 48, 17.8%).

Reasoning categories were tested against computed difference scores to test the effect of reported attributed reasons for initial proto-trust assessment on subsequent changes from proto-trust to deliberative trust assessments.

For the high trustworthy face in the trustworthy-prompted attribution condition, the mean trust score decreased significantly from proto- to deliberative trust scores when participants reported their emotional reaction as a reason for their initial proto-trust assessment, compared with those who did not, $t(28)$ = 1.78, p = .043). The opposite occurred for emotion-based reasoning in the low trustworthy face with untrustworthy-prompted attribution condition, $t(27)$ = -2.34, p = .013, for feature-based reasoning in the neutral face with neutral-prompt attribution condition, $t(28)$ = -1.87, p = .036) and reasoning based on facial expression in the high trustworthy face with trustworthy-prompted attribution, $t(28)$ = -2.93, p = .003.

Discussion

As expected, the way attribution was prompted influenced the difference between the two trust assessments. The amplification hypothesis was supported for one of the conditions, marginally for another but not for the remaining conditions. The fact that a significant difference between proto-trust and deliberative trust scores did not occur for the majority of cases is not surprising given the theory presented. In the absence of any salient reasons to the contrary that might arise through careful consideration, most participants may have simply been able to form consistent attributions for their initial intuitive assessment. It may be that differences are only seen when there are disruptions to this coherency. For the low trustworthy face when untrustworthy-centric attributions were prompted, and (although only marginally significant) for the high trustworthy face when trustworthy-centric attributions were prompted, there was likely more to draw from in these faces (e.g., trustworthy/untrustworthy facial features, positive/negative valanced perceived emotional expressions, etc.) that aligned with the suggested attribution prompts. This may be particularly true if an individual's initial intuitive assessment went against the conventional assessment of the face (i.e., by the way the face was originally developed to be high/low on trustworthiness). Indeed, in light of the significant correlation between this difference and the initial proto-trust score for the former condition, it is clear that it was generally only the higher proto-trust scores that were subsequently lowered by untrustworthy-centric attribution prompts.

A curious effect that emerged from the qualitative data was that amplification appeared to be dampened or even reversed in these conditions when participants attributed their initial trust assessment to emotion. Consistent with the affect-driven component of proto-trust, this result may point to this subset of participants' lack of true attribution of their feeling of trust to a concrete feature, thus simply *recognising* their intuitive feeling of trust rather than *rationalising* it. Their own emotional reaction to the face would therefore be internally focused, rather than externally attributed.

The results of qualitative analysis provide preliminary insight into the reasoning and attribution process individuals may undergo when explaining their intuitive trust assessments of strangers under conditions involving limited information. For example, the fact that many participants attributed their initial trust assessments to the similarity of a face to an abstract stereotype or a familiar person may suggest that, where similarities are available, people tend to refer to these associations in the appearance of faces to form opinions (or, at least, explain their opinions). In addition, the fact that many participants referred to features of the faces and the perceived expression on the faces as reasons for their decisions of trustworthiness is unsurprising given that the features of the faces themselves were modified in order to vary on trustworthiness (Oosterhof & Todorov, 2008), as well as the documented overlap between impressions of trustworthiness and perceived expressions (Oosterhof & Todorov, 2009).

Conclusion and future work

This study represents an initial investigation into the new concept of proto-trust. Although intuitive trust is not necessarily a new concept, its nature beyond relational or affect based trust had not yet been defined in the literature. In addition, there has traditionally been little attempt to describe or demonstrate the interaction between intuitive, primitive forms of trust and more reason-based decisions of trust. This first attempt to access the trust termed in this paper as proto-trust, as well as investigate its relationship with more deliberative trust decisions via an attribution process provides some promising avenues for future research.

By acknowledging a more primitive form of trust in terms of emotional and intuitive elements, the proposed theory of proto-trust, and its interaction with deliberative trust via an attribution process, stands to provide a much richer understanding of the variety of trust types attached to conscious deliberations of risk, vulnerability, and expectation. Affect and intuition based theories of trust have been drawn upon to inform an initial conceptualisation of proto-trust in this paper. The intent of exploring a theory of proto-trust is to provide common ground upon which trust is understood and to help explain why such a variety of attitudes, decisions and behaviours all fall under the common umbrella of *trust*.

Future directions for this research will be to further investigate the effect of proto-trust assessments on deliberative trust decisions by manipulating proto-trust itself (e.g., by altering the perceived trustworthiness of a number of elements in an individual's experience) to determine the effect on the attribution of trust to a specific object (e.g., a face).

Acknowledgments

This research was supported by the Asian Office of Aerospace Research and Development (AOARD) grant: FA2386-17-1-4016

References

Alrajih, S., & Ward, J. (2014). Increased facial width-to-height ratio and perceived dominance in the faces of the uk's leading business leaders. *British Journal of Psychology*, *105*(2), 153-161.

Amaral, G., Sales, T. P., Guizzardi, G., & Porello, D. (2019). Towards a reference ontology of trust. In *Otm confederated international conferences" on the move to meaningful internet systems"* (pp. 3–21). Springer.

Ben-Ner, A., & Halldorsson, F. (2010). Trusting and trustworthiness: What are they, how to measure them, and what affects them. *Journal of Economic Psychology*, *31*(1), 64–79.

Campbell, L., & Stanton, S. C. E. (2019). Adult attachment and trust in romantic relationships. *Current opinion in psychology*, *25*, 148–151.

Castelfranchi, C. (1999). Affective appraisal versus cognitive evaluation in social emotions and interactions. In *International workshop on affective interactions* (pp. 76–106). Springer.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model* (Vol. 18). John Wiley Sons.

Castle, E., Eisenberger, N. I., Seeman, T. E., Moons, W. G., Boggero, I. A., Grinblatt, M. S., & Taylor, S. E. (2012). Neural and behavioral bases of age differences in perceptions of trust. *Proceedings of the National Academy of Sciences*, *109*(51), 20848-20852.

Chowdhury, S. (2005). The role of affect-and cognition-based trust in complex knowledge sharing. *Journal of Managerial issues*, 310–326.

Chua, R. Y. J., Ingram, P., & Morris, M. W. (2008). From the head and the heart: Locating cognition-and affect-based trust in managers' professional networks. *Academy of Management journal*, *51*(3), 436–452.

Clark, A. (2015). Radical predictive processing. *The Southern Journal of Philosophy*, *53*, 3–27.

Critchley, H. D., & Garfinkel, S. N. (2017). Interoception and emotion. *Current opinion in psychology*, *17*, 7–14.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427.

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., ... Dalgleish, T. (2010). Listening to your heart: How interoception shapes emotion experience and intuitive decision making. *Psychological science*, *21*(12), 1835–1844.

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. *Journal of personality and social psychology*, *88*(5), 736.

Ellsworth, P. C., & Scherer, K. R. (2003). *Appraisal processes in emotion.* Oxford University Press.

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.

Evans, A. M., Dillon, K. D., Goldin, G., & Krueger, J. I. (2011). Trust and self-control: The moderating role of the default. *Judgment Decision Making*, *6*(7).

Evans, J. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, *21*(4), 313–326.

Falcone, R., & Castelfranchi, C. (2001). Social trust: A cognitive approach. In *Trust and deception in virtual societies* (pp. 55–90). Springer.

Fell, L., Gibson, A., Bruza, P. D., & Hoyte, P. (2020). Human information interaction and the cognitive predicting theory of trust. In *Proceedings of the 2020 conference on human information interaction and retrieval* (pp. 145–152).

Freitag, M., & Traunmüller, R. (2009). Spheres of trust: An empirical analysis of the foundations of particularised and generalised trust. *European journal of political research*, *48*(6), 782–803.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.

Gompei, T., & Umemuro, H. (2018). Factors and development of cognitive and affective trust on social robots. In *International conference on social robotics* (pp. 45–54). Springer.

Günaydin, G., Zayas, V., Selcuk, E., & Hazan, C. (2012). I like you but i don't know why: Objective facial resemblance to significant others influences snap judgments. *Journal of Experimental Social Psychology*, *48*(1), 350-353.

Hardin, R. (2002). *Trust and trustworthiness.* Russell Sage Foundation.

Hermes, J., Behne, T., & Rakoczy, H. (2018). The development of selective trust: Prospects for a dual-process account. *Child Development Perspectives*, *12*(2), 134–138.

Hohenstein, J., & Jung, M. (2020). Ai as a moral crumple zone: The effects of ai-mediated communication on attribution and trust. *Computers in Human Behavior*, *106*, 106190.

Hohwy, J. (2013). *The predictive mind.* Oxford University Press.

Horberg, E. J., Oveis, C., & Keltner, D. (2011). Emotions as moral amplifiers: An appraisal tendency approach to the influences of distinct emotions upon moral judgment. *Emotion Review*, *3*(3), 237–244.

Hou, C., & Liu, Z. (2019). The survival processing advantage of face: The memorization of the (un) trustworthy face contributes more to survival adaptation. *Evolutionary Psychology*, *17*(2), 1474704919839726.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, *58*(9), 697.

Kiousis, S. (2001). Public trust or mistrust? perceptions of media credibility in the information age. *Mass communication society*, *4*(4), 381–403.

Lambert, B., Declerck, C. H., & Boone, C. (2014). Oxytocin does not make a face appear more trustworthy but improves the accuracy of trustworthiness judgments. *Psychoneuroendocrinology*, *40*, 60-68.

La Porta, R., Lopez-de Silanes, F., Shleifer, A., & Vishny, R. W. (1997). Trust in large organizations. *The American economic review*, 333–338.

Levi, M., & Stoker, L. (2000). Political trust and trustworthiness. *Annual review of political science*, *3*(1), 475–507.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social forces*, *63*(4), 967–985.

Little, A. C., Roberts, S. C., Jones, B. C., & DeBruine, L. M. (2012). The perception of attractiveness and trustworthiness in male faces affects hypothetical voting decisions differently in wartime and peacetime scenarios. *Quarterly Journal of Experimental Psychology*, *65*(10), 2018-2032.

Lowe, R., Herrera, C., Morse, A., & Ziemke, T. (2007). The embodied dynamics of emotion, appraisal and attention. In *International workshop on attention in cognitive systems* (pp. 1–20). Springer.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, *38*(1), 24–59.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, *2*(2), 1–25.

McKnight, D. H., & Chervany, N. L. (2000). What is trust? a conceptual analysis and an interdisciplinary model. *AMCIS 2000 proceedings*, 382.

McKnight, D. H., & Chervany, N. L. (2001). Trust and distrust definitions: One bite at a time. In *Trust in cybersocieties* (pp. 27–54). Springer.

Oh, D., Dotsch, R., Porter, J., & Todorov, A. (2019). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General*.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092.

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, *9*(1), 128.

Punyatoya, P. (2019). Effects of cognitive and affective trust on online customer behavior. *Marketing Intelligence Planning*.

PytlikZillig, L. M., & Kimbrough, C. D. (2016). Consensus on conceptualizations and definitions of trust: Are we there yet? *Interdisciplinary perspectives on trust*, 17–47.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of personality and social psychology*, *49*(1), 95.

Sbaffi, L., & Rowley, J. (2017). Trust and credibility in web-based health information: a review and agenda for future research. *Journal of medical Internet research*, *19*(6), e218.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review*, *69*(5), 379.

Shaked, A., & Clore, G. L. (2017). Breaking the world to make it whole again: Attribution in the construction of emotion. *Emotion Review*, *9*(1), 27–35.

Shapiro, S. P. (1987). The social control of impersonal trust. *American journal of Sociology*, *93*(3), 623–658.

Slepian, M. L., Young, S. G., & Harmon-Jones, E. (2017). An approach-avoidance motivational model of trustworthiness judgments. *Motivation Science*, *3*(1), 91.

Stoltz, D. S., & Lizardo, O. (2018). Deliberate trust and intuitive faith: A dual-process model of reliance. *Journal for the Theory of Social Behaviour*, *48*(2), 230–250.

Sutherland, C. A., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology*, *108*(2), 397-415.

Swe, D. C., Palermo, R., Gwinn, O. S., Rhodes, G., Neumann, M., Payart, S., & Sutherland, C. A. M. (2020). An objective and reliable electrophysiological marker for implicit trustworthiness perception. *Social cognitive and affective neuroscience*, *15*(3), 337–346.

Tan, H. H., & Lim, A. K. H. (2009). Trust in coworkers and trust in organizations. *the Journal of Psychology*, *143*(1), 45–66.

Tingley, D. (2014). Face-off: Facial features and strategic choice. *Political Psychology*, *35*(1), 35-55.

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences*, *1124*(1), 208-224.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social cognitive and affective neuroscience*, *3*(2), 119-127.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813-833.

Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication*, *20*(5), 520–535.

Uslaner, E. M. (2002). *The moral foundations of trust*. Cambridge: Cambridge University Press.

Webber, S. S. (2008). Development of cognitive and affective trust in teams: A longitudinal study. *Small group research*, *39*(6), 746–769.

Wesson, D. E., Lucey, C. R., & Cooper, L. A. (2019). Building trust in health systems to eliminate health disparities. *Jama*, *322*(2), 111–112.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the united states and japan. *Motivation and emotion*, *18*(2), 129–166.