HOW DO WE EVEN TRUST? A CRITIQUE OF ECONOMIC APPROACHES ON TRUST FORMATION

Maria Banu
University of Bucharest, Romania

Received: April 29, 2024 Accepted: May 22, 2024 Online Published: May 25, 2024

Abstract

The aim of this paper is to provide a review of trust formation models in early economic accounts and behavioral accounts of trust. Drawing on foundational theories of trust across several disciplines in the social sciences and empirical studies, I critically examine their assumptions and implications to identify the theoretical and conceptual gaps within each approach. I argue that neither approach explains successfully trust formation in interactions between individuals, due to a lack of clarity and robustness in defining trustworthiness and inadequately accounting for the role of trustworthiness expectations. Beliefs about others' trustworthiness are central to trust. A robust account of trust formation must explain how we form such beliefs. To this end, this paper outlines an epistemological account of trust formation, meant to provide a more cohesive understanding of trust dynamics at the individual level.

Keywords: Trust; Trust formation; Rationality; Trustworthiness; Epistemology of trust; Warranted trust.

1. Introduction

Friends or strangers – we trust others every day. Trust is beneficial for our well-being, our relationships and society in general. Some people we trust blindly, many others – we ponder whether to trust them or not. Trust is risky, so we must invest it wisely, particularly when things that are important to us depend on others. The question is: what makes us trust others and how do we know our trust is warranted?

Early accounts of trust stemming from sociology, political science, and social psychology, use the standard, neoclassical rational choice model to answer this question. Assuming rational actors are utility-maximizing, self-interested individuals, early "economic" accounts of trust investigate if trust is rational. They lock together the concepts of trust, interests, and cooperation via the notion of individual rationality used in the

standard rational choice model. Trust reflects a strictly cognitive expectation about the trustworthiness of others, which results from the subjective probability calculation that others would be interested in reciprocating our trust.

Behavioral economists argue against the notion of individual rationality used in early economic accounts of trust. They show that ordinary people rarely behave like rational self-interested individuals. They often trust without being able to form expectations about others' trustworthiness. In fact, trust is emotionally wired. Does that make us all irrational? If trust does not necessarily rely on expectations about trustworthiness, then what determines it? Behavioral economists hypothesize that trust might have to do with social preferences, the expressive function of trust, or merely with the fact that our rationality is bounded. They took an entirely different path in explaining trust compared to early trust scholars, but in the process, they conflated trust with cooperative behaviors.

In this paper, I question the notions of trust and trustworthiness used in early economic accounts and behavioral accounts of trust. I argue that neither approach explains trust formation successfully. Trust formation refers to the process through which a truster forms an expectation about the trustworthiness of the trustee. This expectation reflects the truster's belief that the trustee is trustworthy. Based on this belief, the truster may eventually decide whether to cooperate with the trustee or not. Early economic accounts of trust provide an incomplete explanation of how individuals form such expectations whereas behavioral accounts tend to elude the role of trustworthiness expectations in trust-taking.

Expectations about others' trustworthiness are inherent in trust. An adequate explanation of trust formation must investigate how people form justified beliefs about the trustworthiness of others. To this end, I provide an epistemological account of trust formation where expectations about trustworthiness form based on assessments of the trustees' competence, predictability, and responsiveness.

The paper is structured as follows. Section 2 briefly presents the methodology I used to build my research. Section 3 presents a critique of early economic accounts and behavioral accounts of trust, as well as an attempt at integrating the two approaches. In Section 4, I outline my epistemological account of trust formation. In Section 5, I conclude and briefly discuss the implications of my account, its limitations, and further research avenues.

2. Methodology

Rather than using a conventional literature review, I organized my paper around a thematic critique of early economic accounts and behavioral accounts of trust. This method enables a targeted examination of the theoretical and conceptual gaps within both approaches and is suitable to highlight the competing intuitions on the definitions of trust. I systematically compare the assumptions and implications within both approaches. This way, I prepare the stage for a more cohesive understanding of trust dynamics at the individual level. Anchoring my account in methodological individualism, I focus on the micro-level processes that lead to trust formation from an epistemological point of view. I pinpoint the elements of an epistemological explanation on how individuals form expectations about the trustworthiness of those they interact with, outside the rational choice framework. I develop my account based on the definition of trustworthiness I provide in Banu (2023).

The selection of studies I incorporate in the analysis covers several disciplines in the social sciences. It includes foundational theories of trust by scholars such as sociologist

James Coleman (1990), political scientists Diego Gambetta (1988) and Russell Hardin (2002), and social psychologist Toshio Yamagishi (1998). The common thread in their works is the use of the standard rational choice model to explain trust formation. The critical examination naturally extends to behavioral economics. I review key experimental studies to illustrate how the understanding of trust determinants has shifted in behavioral economics. I start with Berg *et al.* (1995), who designed the experimental research into trust. I then identify clusters of empirical studies which advance various determinants of trust (for instance, social preferences or expressive trust) and I critically discuss their implications on the definition of trust.

3. The economics of trust

3.1 Early economic accounts of trust

Two hunters, *A* and *B*, must decide whether to hunt a stag together or catch a hare on their own. The stag is more rewarding, but hunting it requires the effort of two individuals. The issue is if they can trust each other. If *A* trusts *B* to hunt the stag but *B* goes rogue and catches the hare, then *A* gets nothing. The same goes for *B*. Rousseau first discussed this dilemma of the "stag hunt" in *Discourse on inequality* (1993 [1755]). It later became known as the "trust game" or the "assurance game."

There are two Nash equilibria in the game (Fig. 1): the optimal equilibrium where *A* and *B* cooperate and hunt the stag, and the sub-optimal equilibrium where each goes for the hare. In the standard rational choice model, rational individuals seek to maximize their utility by pursuing self-interest. On this assumption, both players in the trust game would prefer cooperation, but it can only work if they trust each other. Otherwise, the rational choice is to catch the hare. Early trust scholars in sociology (e.g., Coleman, 1990), political science (e.g., Gambetta, 1988; Hardin, 2002), and social psychology (e.g., Yamagishi, 1998) draw on the insights of the trust game and its assumptions to study trust. They conclude that trust is rational if: (i) cooperation is the strategy that brings the best possible results, and (ii) the trustee is trustworthy.

Figure 1 – Payoff matrix in the stag hunt game

		B	
		Hunts the stag	Hunts the hare
A	Hunts the stag	4, 4	0, 3
	Hunts the hare	3, 0	3, 3

This conclusion bears implications on the definition of trust. First, trust is intrinsically linked to rational choice. Early accounts ground it in self-interest via the notion of individual rationality used in the standard rational choice model. As Gambetta (1988, p. 222) put it, trust "is a matter, … also of interest." Interests "govern action independently of a given level of trust, but it can also act on trust itself by making behavior more predictable." To Coleman (1990, p. 99), trust is rational if the ratio between the probability that the trustee will cooperate and the probability that she will not is higher than the ratio

of potential losses and gains. The second implication is that expectations about others' trustworthiness are central to trust. Early trust scholars take trust to reflect the belief that the trustee is trustworthy. Player *A* must figure out if trusting *B* allows her to pursue her interest. In early accounts, a trustworthy trustee is simply one that will answer favorably to our trust. She "will perform an action that is beneficial or at least not detrimental to us" (Gambetta, 1988, p. 217). The third implication is that trust is purely cognitive. To assess the trustee's trustworthiness, a truster relies on the assumption that the trustee is a rational agent. A rational agent will choose the strategy with the highest payoff. In the trust game, the truster has complete information about the trustee's payoffs and possible strategies. She can calculate and anticipate the probability that the trustee will cooperate.

There are three issues here. First, real people rarely act in conditions of full certainty, have limited cognitive skills, and their decisions are heavily context dependent. The claim that they make perfectly rational decisions is unfounded and discounts the context and uncertainty in real deliberative processes. Kahneman & Tversky (1979) show that human decision-making is biased and follows simplistic cognitive shortcuts. We are not good at estimating probabilities or making predictions. Because we are "boundedly rational," we look for our decisions to satisfy our preferences, not to maximize them (Simon, 1997). The point is to reach a "good enough" decision based on the available information. Early trust scholars recognize the uncertainty involved in trust (Gambetta, 1988; Yamagishi, 1998). Yet, they still hold trust and trustworthiness to the high standard of individual rationality and decision-making in the standard rational choice model.

Second, if trustworthiness is contingent on payoffs, then the concept expresses a simplistic notion of merely being reliable when it pays off. But one can be reliable without necessarily being trustworthy. In his theory of trust as encapsulated interest, Hardin (2002, p. 28) argues that a robust definition of trust must account for the motivations of the trustee to fulfill the trust. He argues that interests are the central motivation for trustworthiness because interests motivate most of our encounters. Other motivations cannot enable "systematic accounts of trust" (2002, p. 52). However, Hardin admits that one is not trustworthy simply because they have an interest to pursue in their relationship with us. To him, "I trust you" means that I think that it is in your interest to consider my interests, that is, to encapsulate them into yours, because you value the relationship with me, and you wish for its continuation.

The issue in Hardin's theory is that his notion of interest is compatible with various reasons people may have for trustworthiness, like "love or friendship and the caring for another" (2002, p. 24). "At a minimum, you may want our relationship to continue because it is economically beneficial to you … In richer cases, you may want our relationship to continue and not to be damaged by your failure to fulfil my trust because you value the relationship for many reasons, including nonmaterial reasons" (2002, p. 4). Allowing interests to match this wide spectrum of reasons for trustworthiness, the notion itself becomes too loose. One could easily interpret almost any kind of reason for trustworthiness as belonging to interests. Like this, we will not be able to differentiate between different motivations for trustworthiness.

A robust theory of trust must allow us to do that if we want to understand trust in all its shapes and forms. Trust is dynamic, it is complex and versatile. The trust I have in my friend is different from the trust I have in my sibling or a work colleague. In each of these relationships I trust people with things that are specific to that relationship. I will not confide in my work colleague as I confide in my friend. We must account for the context-

dependency of trust and its determinants from one context to another. This includes the different reasons the trustees may have for trustworthiness, if the truster finds those reasons relevant enough to trust, and how she assesses them. People have multiple reasons to fulfill others' trust, beyond interests. Sometimes, interests go hand in hand with moral commitments or sympathy. Other times, pursuing our interests might mean not being able to keep a promise or fulfill a commitment.

The final issue is that early economic accounts of trust cannot explain social or generalized trust. Social trust refers to trust we have in general others, including strangers. In the standard rational choice model, trust in strangers is irrational because we know little about them and we cannot form expectations about their trustworthiness. Generalized trust can only happen in small, close-knit communities (Williams, 1988, p. 12), with tight social relations, repeated interactions, and easily accessible information about others, even if from third parties (Hardin, 2002, p. 21). Now, the social sciences took an interest in trust precisely because social trust is assumed to foster economic prosperity (Knack & Keefer, 1997; Zak & Knack, 2001) and democracy (Putnam, 1993; Warren, 2018). Trust in strangers can solve social dilemmas.

The issue is, are we all irrational to trust strangers? In his emancipation theory of trust, Yamagishi (1998) manages to explain trust in strangers within the rational choice framework. He points to trustfulness rather than trustworthiness as an antecedent of trust. Trustfulness is a disposition to trust in general, different from generalized expectations about trustworthiness (1998, p. 46). It is relevant in situations of high social uncertainty because it drives us to engage in new interactions with people whom we have little information about. In Yamagishi's view, this is rational because it allows us to pursue our interests (1998, p. 67-68). We may subjectively think that interests do not motivate our trusting others, yet we benefit from trust, directly or indirectly. In this sense, Yamagishi maintains trust in the standard rational choice framework.

To Yamagishi, expectations about trustworthiness still play a role in trust formation. High trusters are not simply naïve. They are "good at understanding the minds and characters of other people" (1998, p. 132), and can "read" cues of untrustworthiness better than low trusters. There are two issues with Yamagishi's theory, though. First, he does not explain what cues we use to assess others' trustworthiness, particularly when we deal with strangers. Second, the notion that trust is rational because it *eventually* secures our interests is not a satisfying explanation. If we want to understand trust drivers, we should be able to explain the subjective deliberative processes of individuals and how expectations about others' trustworthiness affect them.

3.2 Behavioral accounts of trust

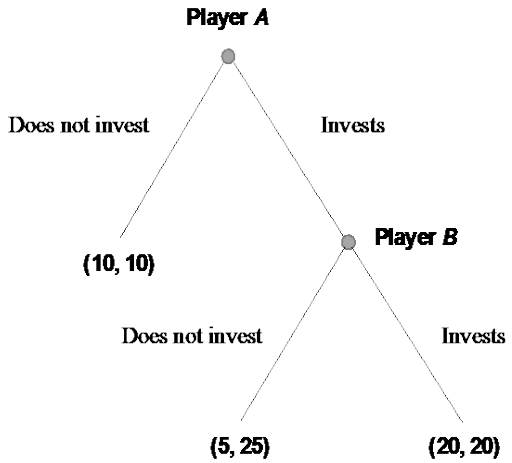
Behavioral economists use laboratory experiments to show that the standard rational choice model does not adequately predict trust. Berg *et al.* (1995) are among the first to test the trust game in experimental settings. They designed the so-called "investment game" to measure differences in trust and trustworthiness among individuals. The game reflects an economic exchange where trust is assumed to be the main driver of investment decisions. The game is played once, under full anonymity. This way, researchers control for other variables that could influence investment decisions, like reputation or sanctions.

In the game, two subjects, *A* and *B*, receive an equal sum of money. *A*, the investor, must decide whether to keep the money or transfer it all or part of it to *B*. If *A* decides to

invest, *B* must make a similar choice. The experimenter triples each transfer or "investment." Both players are aware of the outcomes of their possible choices. If economic agents are interested only in their own gains, there is no rational reason for *A* and *B* to transfer money to each other (that is, cooperate). *B* gains nothing by transferring money back to *A*. *A* anticipates this and makes no initial investment. Although both players could increase their earnings if they invested, the rational choice for each is to go home with the endowment initially received. Figure 2 illustrates this decision-making situation in sequential form. Both players initially hold 10 dollars. If each invested 5 dollars, each would ultimately gain 20 dollars.

The experiment shows that the standard rational choice model used in the early economic accounts cannot predict people's trust-taking behaviors. Over 90% of "investors" chose to invest, and many of trustees in *B*'s role reciprocated (Berg *et al.* 1995). The hypothesis is that *B* transfers money back because she interprets *A*'s behavior as trust. Replications of the game in various countries and varying experimental settings produced similar results (e.g., Ortmann *et al.*, 2000; Koford, 1998; Willinger *et al.*, 1999; Snijders & Keren, 1998).

Figure 2 – The investment game



The implications of behavioral studies on the notion of trust oppose those of early economic accounts. First, trust does not necessarily reflect expectations about others' trustworthiness. People trust in the absence of such expectations. The investment game is usually played in full anonymity, so player *A* cannot assess and predict *B*'s trustworthy behavior. In early economic accounts, *A*'s decision to invest is irrational. Yet, a high percentage of people do it. Behavioral economists conclude that the concept of rationality in the standard rational choice model is wrong. Second, if trust does not reflect trustworthiness expectations, it belongs to the realm of decisions, actions, or behaviors. Some behavioral economists even separate cognitive trust from behavioral trust. Fetchenhauer *et al.* (2017, p. 140) argue, "trust at the cognitive level is not identical to trust at the behavioral level. The two are dissociated, and that can lead to patterns of thought and behavior suggesting, somewhat paradoxically, that people trust both too little and too much at the same time, depending on the level of trust one is focusing on."

Finally, trust-taking relies on emotions "in addition to economic considerations concerning monetary outcomes" (Engelmann & Fehr, 2017, p. 34). Failure to reciprocate

trust triggers an emotional reaction called betrayal or exploitation aversion. This is a form of social anxiety that inhibits cooperative behaviors (Fehr *et al.*, 2005; Bohnet & Zeckhauser, 2004; Fehr, 2009). It emerges when betrayal seems to be intentional (Bohnet *et al.*, 2008). Pharmacological studies further explore how emotions mediate trust-taking by administering oxytocin to subjects in the investment game (e.g., Kosfeld *et al.*, 2005; Baumgartner *et al.*, 2008; Mikolajczak *et al.*, 2010; De Dreu *et al.*, 2010; Van IJzendoorn and Bakermans-Kranenburg, 2012; Zhong *et al.*, 2012). Oxytocin regulates bonding with offspring and sexual partners (Churchland 2011). It influences social behavior (Skuse & Gallagher, 2009; Meyer-Lindenberg *et al.*, 2011) and facilitates social cognition (Kirsch *et al.*, 2005; Domes *et al.*, 2007a; 2007b; Guastella *et al.*, 2008; 2010). Kosfeld *et al.* (2005), for instance, show that oxytocin enhances trust-taking behaviors. Investors in the treatment group transferred money 45% of the time compared to 21% in the control group. The effect disappeared when subjects played the game with a computer.

Behavioral economists investigated what drives trust in the absence of expectations about trustworthiness. Some argue that trust reflects the agents' social preferences, like aversion to inequity (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000), concern for social welfare (Charness & Rabin, 2002; Andreoni & Miller, 2002), or reciprocity (Rabin, 1993; Chatterjee *et al.*, 2019). The issue with these models is that choices are driven by anticipated outcomes. There is no investigation of the intentions or decision-making processes underlying trust-taking (McCabe *et al.*, 2003; Krueger *et al.*, 2017). The models are more suitable to explain trustworthiness rather than trust. Other studies focus on the expressive function of trust. The act itself and its meaning seem to motivate trust-taking. Trust involves a complex association of beliefs, norms, and emotions (Fetchenhauer *et al.*, 2017). Dunning *et al.* (2012; 2014) argue that trust is about compliance with an injunctive norm. Individuals feel like they have a social obligation to trust, associated with fulfilling a duty or social responsibility, whereas distrust is associated with guilt or anxiety. Yamagishi *et al.* (2015) argue that trust may be motivated by a preference related to one's own identity – we wish to perceive ourselves as trustful individuals. Newer studies show that trust is neither fully rational nor fully expressive. The decision to trust reflect the agents' "bounded rationality" (Evans & Krueger, 2016). Krueger *et al.* (2017, p. 122) explain that the truster simply compares possible gains and losses. She estimates the probability that the trustee will reciprocate by projecting her own prosocial orientation onto the trustee. This assessment is limited because it is difficult for the truster to assess the trustee's motives and work with probabilities optimally. This is a quick, egocentric, and heuristic strategy that allows the truster to make a "good enough" decision.

There are two issues with the definition of trust and trust formation in behavioral economics. First, behavioral economists discount the role of expectations about trustworthiness in trust-taking. This leads to the paradoxical situation where one can trust at a behavioral level without believing that the other is trustworthy. Trust must involve at least a minimal degree of belief about others' trustworthiness. Otherwise, what we call "trust" may simply be a cooperative behavior. Behavioral economists assume that, if *A* trusts *B*, then *A* will cooperate with *B* and, if *A* cooperates with *B*, then it means that *A* trusts *B*. Trust is not a necessary condition for cooperation, though.

Second, there is no insight on the trusters' subjective experience in trust-taking. We do not know what eventually determines trust and the decision-making processes associated with it. The accounts above do not separate the payoff matrix from the agents' subjective experience, that is, how they perceive and interpret the objective structure of the

trust problem (Rompf, 2015). In real life, people assess trustworthiness based on certain cues, and these assessments are intrinsically linked to the context of the trust situation and the object of trust, that is, the things we trust others with. Explanations of trust formation should focus how people form expectations about trustworthiness, beyond material outcomes. Subjective probabilities are a consequence of the actors' beliefs and desires, their worldview, and their perceptions of those around them (Castelfranchi & Falcone, 2000, p. 6). Trust influences actors' perceptions on the possible gains from cooperation. To understand how, we must first explain how beliefs about trustworthiness form.

3.3 An integrative account of economic approaches on trust

Rompf (2015) integrates the competing explanations of trust formation in early economic accounts and behavioral accounts. Considering that rationality is flexible and adaptive, he uses the "dual process paradigm" to explain that "human cognition may occur in either a rational or an automatic mode" (2015, p. 157). The automatic mode involves minimal cognitive effort, is unconscious, and relies on intuition, readily available associations, and heuristics. The rational mode uses complex, explicit, conscious, inferential, and controlled reasoning. The former is fast, emotional, associative, and learns slowly (Kahneman, 2011). The latter is slow, sequential, consciously monitored, voluntarily controlled, and requires cognitive effort (Kahneman, 2011).

To explain trust formation, Rompf argues that subjects interpret the trust situation via mental trust-related schemata. These are "(1) frames, that is, mental models of typical situations, and (2) scripts, broadly understood as 'programs of behavior'" (Rompf, 2015, p. 198). Frames largely belong to common knowledge. They are socially and culturally shared. Activation of a frame triggers certain scripts or behavioral patterns. The "sufficiency principle" and "effort-accuracy tradeoffs" govern the selection between the automatic and rational mode of information processing via frames and scripts. Unconditional trust, that is, "trust without doubtful and conscious elaboration" (Rompf, 2015, p. 216) uses the automatic mode of information processing. Conditional trust, where "the trustor subjectively faces the trust problem and elaborates on his future course of action" (Rompf, 2015, p. 216), activates the rational mode. Expectations about trustworthiness form only in conditional trust (Rompf, 2015, p. 224). Situational cues derived from communication, identity signaling, and impression management help the formation of such expectations.

Rompf provides a more coherent account of the link between trust and rationality. The shortcoming in his model is the lack of a clear and complete definition of trustworthiness and how we form such expectations about others. A robust explanation of trust formation should map the criteria and the information based on which we form trustworthiness expectations, as well as the factors that influence the belief formation processes. It should also cover unconditional trust. While the cognitive processes involved in it are quick, thin, unconscious, and prone to errors, it does not mean that trustworthiness expectations are absent. If they were, we would not be talking about trust. The fact that they form spontaneously does not absolve us from the task of understanding how they form. Rompf's model cannot explain, for instance, why a smile is enough to trust in certain contexts, while in others we need more solid evidence.

The analysis of early economic and behavioral accounts of trust, as well as Rompf's attempt at integrating the two, highlights a significant gap in how trust and trust formation have been conceptualized. That is the failure to adequately define trustworthiness and

account for the expectations of trustworthiness that are integral to trust. Early economic accounts reduce trust to the calculation of risks and outcomes based on rational choice theory and present a simplistic notion of trustworthiness. Behavioral models introduce psychological avenues to explain trust at the behavioral level but overlook the belief forming processes underlying trust-taking. Rompf provides a notion of individual rationality that is more suitable to explain individual decision-making in trust situations. But his account still falls short to explain how individuals form and justify their beliefs about the trustworthiness of others. These shortcomings point to the need of separating trust from rational choice and exploring trust formation from an epistemological perspective. In this perspective, an account of trust formation should focus on the reasons that justify the belief that one is trustworthy. This belief can indeed determine trust-taking, but it need not lead there. Reasons for cooperation, like the anticipation of certain outcomes, should not be conflated with reasons for believing one is trustworthy.

An epistemological account of trust formation addresses these shortcomings by focusing on the nature and justification of beliefs about trustworthiness. Such an account should integrate cognitive, emotional, and contextual factors that influence how beliefs are formed and adjusted considering new evidence or experiences. By examining the mechanisms through which individuals assess trustworthiness, an epistemological approach can provide a more comprehensive framework for understanding trust. This framework not only considers the logical and empirical grounds upon which trust-related beliefs are based but also incorporates the socio-cultural dimensions that influence these processes. Thus, moving towards an epistemological model of trust formation not only bridges the conceptual gaps identified in existing theories. It also enriches our understanding by highlighting the complex interplay of factors that underpin trust in dynamic social interactions.

4. The epistemology of trust

An adequate explanation of trust formation must account for the information that feeds into the belief formation processes about the trustworthiness of others. It must define the criteria based on which trusters assess available information, the psychological and contextual factors that influence beliefs about trustworthiness, as well as when they are well-grounded or justified.

My account of trust formation relies on the definition of trustworthiness which I propose in Banu (2023). I argue there that one is trustworthy if she is competent, predictable, and responsive with respect to the thing that we trust her with (the object of trust). Competence means possessing the necessary abilities to fulfill one's trust with respect to something, as well as the capacity to assess and choose the optimal way to apply those abilities in fulfilling the trust. Predictability means both being reliable and having the relevant reasons to fulfill one's trust with respect to the object of trust. Responsiveness captures the trustee's intentionality about or directed at the truster, her willingness to meet the truster's needs with respect to the object of trust. These criteria provide the necessary and sufficient conditions for trustworthiness assessments. While they apply to any trust situation, my account acknowledges that specific information that feeds into the assessment of each criterion will vary with context. Competence, predictability, and responsiveness mean different things in different trust situations.

My account of trust formation rests on the following prerequisites. First, uncertainty is inherent in trust formation. The available information about others is always incomplete and imperfect. Second, trust is a three-place predicate of the form "*A* trusts *B* with *X*." That is, trust is specific to a certain object, it does not simply describe *A*'s general trust in *B*. *A*'s assessment of *B*'s trustworthiness is particularized to the object of her trust in *B*. Third, trust is highly context-dependent, and it varies in time. The three-place formula above must be restated as: "*A* trusts *B* with *X* in context *C*, at time *t*" (Bauer & Freitag, 2018, p. 16). Finally, trust reflects a degree of belief about another's trustworthiness. My account allows for instant or spontaneous trust, but not complete or "blind" trust. The latter amounts to faith, which is evidently different from trust.

The information we gather about others from different sources constitutes the input of our assessments about their trustworthiness. We can gather information from (i) direct observations of their actions, behaviors, reactions; (ii) second-person reporting, that is, statements they make about themselves; (iii) the history of interactions and the relationship we have with them; (iv) third-party sources, including reputation; (v) the context of interaction. We ground assessments about others' trustworthiness in cues we collect from all these sources. We may not have access to all sources at the same time. When trusting a stranger, the information we can gather is limited and it is mainly based on the context of interaction. Plus, not all the information we can gather is relevant to assess trustworthiness. The available information is relevant only the extent to which it allows us to assess the trustee's competence, predictability, and responsiveness.

Various psychological factors influence our assessments about trustworthiness, such as our own high or low propensity to trust in general, affects (moods and emotions), heuristics and biases. As high trusters are more optimistic than low trusters (Carver & Scheier, 2018), they will focus their attention on signals of trustworthiness rather than untrustworthiness (Aspinwall *et al.*, 2001). Moods and emotions feed our beliefs about others by "telling" us how we feel about them (Schwarts & Clore, 1988). They provide interpretation patterns based on which we assess the relevance of available information (Jones, 1996, p. 12). Heuristics and biases affect the accuracy of our judgments (Kahneman, 2011). For example, we overestimate intentions or personality traits in observing others' behaviors and neglect situational constraints (Ross, 1977). Contextual factors such as the environment, norms of interaction, and culture can also influence our perceptions of others. Context means the entire set of contingent circumstances that make up a particular situation. It can both inform and influence perception. I may be more inclined to perceive you as trustworthy if I meet you at the university library in the morning rather than on a deserted street at night (Altman Klein *et al.*, 2019, p. 12). Cultural norms and practices provide a framework for understanding and interpreting others' behaviors (Hinton, 2016, p. 148). It distorts to some extent our view of the world (Wilson, 2007).

As imperfect as they may be, perceptions and judgements about others must meet the three criteria of trustworthiness to be able to say that we trust someone. Based on them, we must be able to assess if the trustee is competent, predictable, and responsive with respect to the thing we want to trust her with. The three criteria are necessary and sufficient conditions for trustworthiness. However, depending on the context of trust, one criterion may be more relevant than the others, so it will weigh heavier in our assessments. My trust in my doctor is grounded mainly in her competence. Competence is such an important indicator in this case that I might unconsciously overlook the other criteria. Yet, no matter how competent she might be, I will not trust her if I notice that she treats her patients badly.

That means she is not responsive to the needs of those that rely on her. My account of trust formation does not presume that in real trust situations people consciously and thoroughly assess the three criteria as such. Yet, assessments of others' trustworthiness must meet the three criteria to consider them trustworthy.

Based on the three criteria, the output of *A*'s assessment about *B*'s trustworthiness is the degree to which *A* trusts *B*. This is a function of *A*'s degree of belief about *B*'s trustworthiness (with respect to *X* in context *C*, at time *t*). This reflects how certain *A* is about her belief, that is, about the sufficiency, relevance, and accuracy of the evidence that feeds her belief, and the reliability of the cognitive processes via which she formed that belief.

How do we know our expectations about others' trustworthiness are warranted? What is the acceptable degree of belief above which we are justified to say we trust someone? In the context of trust, warranted means that trust "successfully targets a trustworthy person" (McLeod, 2022, para. 2). Taken as such, this does not tell us much. Plus, there are situations where it seems we have every reason to trust and yet we are disappointed. Does that mean that our trust was not justified? What does justification mean in the context of trust and how do we know we have good grounds for believing that a person is trustworthy?

There is no foolproof answer to these questions. Trust does not reflect a belief in the sense in which we discuss about beliefs in philosophical epistemology. Trust relies rather on a "street-level epistemology" (Hardin, 2002, p. 115). The knowledge involved in it is highly subjective. While we may wish to form true beliefs about others, we do not seek to find out if we can rely on them for the sake of truth, but because we aim to achieve some goals. Trust helps us navigate uncertain social situations and achieve our objectives, without it being instrumental. When we assess trustworthiness, the point is to gather enough information to satisfy, not fulfill, the three criteria for trustworthiness to an extent that we find acceptable. The stakes of the trust situation (how important the object of trust is to us) and the specific circumstances of it define "acceptable." I will not hand over the keys to my apartment to some stranger in the street, but I might trust her to point me correctly to some address that I am looking for.

On this account, warranted beliefs about the trustworthiness of strangers are not impossible to form. They will rely on minimal evidence, drawn from the context of our interactions, and on heuristics. This does not mean that they are completely unreliable. As social beings, we are equipped to observe others and make quick inferences about their intentions or actions. We rely on various cues to form beliefs when we lack solid information. We automatically judge trustworthiness based on appearance or status (Thielmann & Hilbig, 2015), facial expressions (Todorov *et al.* 2005; 2008), or emotions (Dunn & Schweitzer, 2005). Surely, we should not rely too much on the accuracy of such evidence (Uddenburg *et al.,* 2020). However, we form such beliefs and we act based on them all the time, without being aware. As Kahneman (2011) explains, the automatic cognitive mode, which is biased and prone to systematic errors, can produce quite accurate representations in social situations. Plus, it can be trained to solve swiftly and correctly various problems. Trust in strangers may warranted if beliefs about their trustworthiness are calibrated to the stakes and circumstances of the specific trust situation.

5. Concluding remarks

In this paper, I critically examine early economic and behavioral models of trust formation. I highlight their failure to address trustworthiness expectations adequately. Early models restrict trust formation to rational choice. They focus on narrow self-interest and overlook the broader complexity of trustworthiness expectations. Behavioral models equate trust with cooperative behavior, neglecting the central role of trustworthiness expectations. To address these gaps, my paper introduces an epistemological account of trust formation that emphasizes competence, predictability, and responsiveness as key criteria for trustworthiness assessments. My account goes beyond economic or behavioral incentives for trust while acknowledging the psychological and contextual factors that could influence the formation of trustworthiness expectations. Trust is warranted when evidence about others' trustworthiness satisfies these criteria given the stakes and specific context of the trust situation.

This paper challenges previous models of trust formation and proposes a different outlook on trust and trust research. It builds on the assumption that human decision-making is even more sophisticated than current traditional and behavioral models of rational choice. It pleads for an acknowledgement on how people's beliefs guide choice. Decision-making models should integrate belief forming processes in explaining and predicting people's choices, as complex and imperfect as these processes may be. Theoretical and empirical accounts of trust should expand to include the multifaceted ways in which trust is shaped by individual history, psychological states, and cultural background. Methodologically, the account I propose calls for a mix of quantitative and qualitative methods to understand trust formation across different contexts and cultures. Case studies and interviews could substantiate quantitative research and offer an integrative narrative for the currently contradictory results produced by attitudinal measurements and behavioral studies on trust. From a practical point of view, the proposed criteria for trustworthiness serve as valuable tools for assessing and building trust in interpersonal relationships and organizations. They are particularly relevant for professionals such as managers and therapists, whose effectiveness heavily relies on establishing trust. This understanding could inform trust-building strategies tailored to specific contexts like business negotiations or team collaboration.

The primary limitation of my account is the lack of empirical research to support it. The three criteria are stated as logical conditions for trustworthiness assessments. While I use empirical research to support the build-up of my account, further empirical research is required to validate and refined the proposed framework. Future research should examine what cues people look at when assessing trustworthiness in others and how trustworthiness perceptions evolve over time and under varying circumstances. It should explore the cognitive processes behind trustworthiness assessments and their dependency on factors like culture and personal experiences.

References

1. Altman Klein, H., Miller, N.L., Militello, L.G., Lyons, J.B., & Finkeldey, J.G. (2019). Trust Across Culture and Context. *Journal of Cognitive Engineering and Decision Making, 13*(1), 10-29.

2. Andreoni, J. & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica, 70*(2), 737-753.
3. Aspinwall, L.G., Richter, L. & Hoffman III, R.R. (2001). Understading how optimism works: An examination of optimists' adaptive moderation of belief and behavior. In E.C. Chang (Ed.), *Optimism & Pessimism: Implications for Theory, Research, and Practice* (pp. 217-238). American Psychological Association.
4. Banu, M. (2023). Whom do we trust? On how we assess others' trustoworthiness. *Annals of the University of Bucharest – Philosophy Series, 71*(1), 85-106.
5. Bauer, P.C., & Freitag, M. (2018). Measuring Trust. In E.M. Uslaner (Ed.), *The Oxford Handbook of Social and Political Trust* (pp. 15-36). Oxford University Press.
6. Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr. E. (2008). Oxytocin Shapes the Neural Circuitry of Trust and Trust Adaptation in Humans. *Neuron,* 58, 639-650.
7. Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior,* 10, 122-142.
8. Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review, 98*(1), 294-310.
9. Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization,* 55, 467-484.
10. Bolton, G.E., & Ockenfels, A. (2000). ERC: A theory of Equity, Reciprocity, and Cooperation. *The American Economic Review,* 90(1), 166-193.
11. Carver, C.S., & Scheier, M.F. (2018). Generalized Optimism. In G. Oettingen, A.T. Sevincer, & P.M. Gollwitzer (Eds.), *The Psychology of Thinking About the Future* (pp. 214-230). New York, London: The Guilford Press.
12. Castelfranchi, C., & Falcone, R. (2000). Trust is Much More Than Subjective Probability; Mental Components and Sources of Trust. *Proceedings of the 33rd Hawaii International Conference on system sciences* (pp.1-10).
13. Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics, 117*(3), 817-869.
14. Chatterjee, C., Johnson, C.K., Sams, A.B.E. (2019). "Equal or Nothing": Concern for Fairness and Reciprocity in Trust Game. *International Journal of Economic Behavior*, 9, 3-11.
15. Churchland, P.S. (2011). *Braintrust: What Neuroscience Tells Us about Morality.* Princeton, Oxford: Princeton University Press.
16. Coleman, J.S. (1990). *Foundations of Social Theory.* Cambridge (MA), London: The Belknap Press of Harvard University Press.
17. De Dreu, C.K.W., Greer, L.L., Handgraaf, M.J.J., Shalvi, S., Van Kleef, G.A., Baas, M., Ten Velden, F.S., Van Dijk, E., & Feith, S.W.W. (2010). The Neuropeptide Oxytocin Regulates Parochial Altruism in Intergroup Conflict Among Humans. *Science,* 328, 1408-1411.
18. Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S.C. (2007a). Oxytocin Improves "Mind-Reading" in Humans. *Biological Psychiatry,* 61: 731-733.
19. Domes, G., Heinrichs, M., Gläscher, J., Büchel, C., Braus, D.F., & Herpertz, S.C. (2007b). Oxytocin Attenuates Amygdala Responses to Emotional Faces Regardless of Valence. *Biological Psychiatry,* 62, 1187-1190.
20. Dunn, J.R., & Schweitzer, M.E. (2005). Feeling and Believing: The Influence of Emotion on Trust. *Journal of Personality and Social Psychology, 88*(5), 736-748.

21. Dunning, D., Fetchenhaur, D., & Schlösser, T.M. (2012). Trust as a social and emotional act: Noneconomic considerations in trust behavior. *Journal of Economic Psychology, 33*(3), 686-694.
22. Dunning, D., Anderson, J.E., Schlösser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at Zero Acquaintance: More a Matter of Respect Than Expectation of Reward. *Journal of Personality and Social Psychology, 107*(1), 122-141.
23. Engelmann, J.B., & Fehr, E. (2017). The Neurobiology of Trust and Social Decision-Making: The Important Role of Emotions. In P.A.M. Van Lange, B. Rockenbach, & T. Yamagishi (Eds.) *Trust in Social Dilemmas* (pp. 33-56). Oxford: Oxford University Press.
24. Evans, A.M. & Krueger, J.I. (2016). Bounded Prospection in Dilemmas of Trust and Reciprocity. *Review of General Psychology, 20*(1), 17-28.
25. Fehr, E. (2009). On the Economics and Biology of Trust. *Journal of the European Economic Association, 7*(2-3), 235-266.
26. Fehr, E. & Schimdt, K.M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics, 114*(3), 817-868.
27. Fehr, E., Fischbacher, U., & Kosfeld, M. (2005). Neuroeconomic Foundations of Trust and Social Preferences. *IZA Discussion Papers, No. 1641* (pp. 1-13). Institute for the Study of Labor (IZA).
28. Fetchenhauer, D., Dunning, D., & Schlösser, T. (2017). The Mysteries of Trust: Trusting Too Little and Too Much at the Same Time. In P.A.M. Van Lange, B. Rockenbachm & T. Yamagishi (Eds.), *Trust in Social Dilemmas* (pp. 139-154). Oxford University Press.
29. Gambetta, D. (1988). Can We Trust Trust? In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 213-37). Basil Blackwell.
30. Guastella, A.J., Mitchell, P.B., & Mathews, F. (2008). Oxytocin enhances the encoding of positive social memories in humans. *Biological Psychiatry, 64*(3), 256-258.
31. Guastella, A.J., Einfeld, S.L., Gray, K.M., Rinehart, N.J., Tonge, B.J., Lambert, T.J., & Hickie, I.B. (2010). Intranasal Oxytocin Improves Emotion Recognition for Youth with Autism Spectrum Disorders. *Biological Psychiatry,* 67, 692-694.
32. Hardin, R. (2002). *Trust and Trusworthiness.* New York: Russell Sage Foundation.
33. Hinton, P.R. (2016). *The Perception of People: Integrating cognition and culture.* London, New York: Routledge.
34. Jones, K. (1996). Trust as an Affective Attitude. *Ethics, 107*(1), 4-25.
35. Kahneman, D. & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica, 47*(2), 263-291.
36. Kahneman, D. (2011). *Thinking, Fast and Slow.* New York: Farrar, Strauss and Giroux.
37. Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., Gruppe, H., Mattay, V.S., Gallhofer, B., & Meyer-Lindenberg, A. (2005). Oxytocin Modulates Neural Circuitry for Social Cognition and Fear in Humans. *The Journal of Neuroscience, 25*(49), 11489-11493.
38. Knack, S. & Keefer, Ph. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics, 112*(4), 1251-1288.
39. Koford, K. (1998). Trust and reciprocity in Bulgaria: A replication of Berg, Dickhaut and McCabe (1995). Working paper 98-08, *University of Delaware Department of Economics*.

40. Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature,* 435, 673-676.
41. Krueger, J.I., Massey, A.L., & DiDonato, T.E. (2008). A Matter of Trust: From Social Preferences to the Strategic Adherence to Social Norms. *Negotiation and Conflict Management Research, 1*(1), 31-52.
42. Krueger, J.I., Evans, A.M., & Heck, P.R. (2017). Let Me Help You Help Me. In P.A.M. Van Lange, B. Rockenbach, & T. Yamagishi (Eds.), *Trust in Social Dilemmas* (pp. 121-138). Oxford University Press.
43. McCabe, K.A., Rigdon, M.L., & Smith, V.L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization,* 52, 267-275.
44. McLeod, C. (2022). Trust. *Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/archives/fall2021/entries/trust/.
45. Mikolajczak, M., Gross, J.J., Lane, A., Corneille, O., de Timary, P., & Luminet, O. (2010). Oxytocin Makes People Trusting, Not Gullible. *Psychological Science, 21*(8), 1072-1074.
46. Ortmann, A., Fitzgerald, J., & Boeing, C. (2000). Trust, reciprocity, and social history: A re-examination. *Experimental Economics,* 3, 81-100.
47. Putnam, R. D. (1993). *Making Democracy Work: Civic Traditions in Modern Italy.* Princeton University Press.
48. Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review, 83*(5), 1281-1302.
49. Rompf, S.A. (2015). *Trust and Rationality: An Integrative Framework for Trust Research.* Springer VS.
50. Ross, L. (1977). The intuitive psychologist and his shortcomings: distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 173-220). New York: Academic Press.
51. Rousseau, J.-J. (1993 [1755]). *Discourse on the Origins of Inequality (Second Discourse). Polemics, and Political Economy*. Hanover, London: University Press of New England.
52. Simon, H.A. (1997). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations* (4th ed.). New York, London, Toronto, Sydney, Singapore: The Free Press.
53. Schwartz, N., & Clore, G.L. (1988). How do I feel about it? The informative function of affective states. In K. Fiedler & J. Forgas (Eds.), *Affect, cognition and social behavior* (pp. 44-62). Lewinston, NY: Hogrefe.
54. Skuse, D., & Gallagher, L. (2009). Dopaminergic-Neuropeptide Interactions in the Social Brain. *Trends in Cognitive Science 13*(1), 27-35.
55. Thielmann, I., & Hilbig, B.E. (2015). Trust: An Integrative Review From a Person–Situation Perspective. *Review of General Psychology, 19*(3), 249-277.
56. Todorov, A., Mandisodza, A.N., Goren, A., & Hall, C.C. (2005). Inferences of Competence from Faces Predict Election Outcomes. *Science, 308*(5728), 1623-1626.
57. Todorov, A., Baron, S.G., & Oosterhof, N.N. (2008). Evaluating face trustworthiness: a model based approach. *Social Cognitive and Affective Neuroscience,* 3, 119-127.
58. Uddenburg, S., Thompson, B., Vlasceanu, M., Griffiths, T.L., & Todorov, A. (2020). A face you can trust: Iterated learning reveals how stereotypes of facial trustworthiness may propagate in the absence of evidence. *Journal of Vision, 20*(11), 1735.
59. Van IJzendoorn, M.H., & Bakermans-Kranenburg, M.J. (2012). A sniff of trust: Meta-analysis of the effects of intranasal oxytocin administration on face recognition, trust to in-group, and trust to out-group. *Psychoneuroendocrinology,* 37, 438-443.

60. Warren, M.E. (2018). Trust and Democracy. In E.M. Uslaner (Ed.), *The Oxford Handbook of Social and Political Trust* (pp. 75-94). Oxford: Oxford University Press.
61. Williams, B. (1988). Formal Structures and Social Reality. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 3-13). Basil Blackwell.
62. Willinger, M., Lohmann, C., & Usunier, J.-C. (1999). Comparison of trust and reciprocity between France and Germany: Experimental investigation based on the investment game. *University Louis Pasteur*.
63. Wilson, J.P. (2007). The lens of culture: theoretical and conceptual perspectives in the assessment of psychological trauma and PTSD. In J.P. Wilson & C.S. Tang, *Cross-cultural assessment of psychological trauma and PTSD* (pp. 3-30). N.Y.: Springer.
64. Yamagishi, T. (1998). *The Structure of Trust: An Evolutionary Game of Mind and Society.* Tokyo: Tokyo University Press.
65. Yamagishi, T., Akutsu, S., Cho, K., Inoue, Y., Li, Y., & Matsumoto, Y. (2015). Two-Component Model of General Trust: Predicting Behavioral Trust from Attitudinal Trust. *Social Cognition, 33*(5), 436-458.
66. Zak, P.J., & Knack, S. (2001). Trust and Growth. *The Economic Journal,* 111, 295-321.
67. Zhong, S., Monakhov, M., Mok, H.P., Tong, T., Lai, P.S., Chew, S.H., & Ebstein, R.P. (2012). U-Shaped Relation between Plasma Oxytocin Levels and Behavior in the Trust Game. *PLoS ONE, 7*(12), e51095.