games

MDPI

Article

A Model of Trust

Gabriele Bellucci

Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany; gabriele.a.bellucci@gmail.com

Abstract: Trust is central to a large variety of social interactions. Different research fields have empirically and theoretically investigated trust, observing trusting behaviors in different situations and pinpointing their different components and constituents. However, a unifying, computational formalization of those diverse components and constituents of trust is still lacking. Previous work has mainly used computational models borrowed from other fields and developed for other purposes to explain trusting behaviors in empirical paradigms. Here, I computationally formalize verbal models of trust in a simple model (i.e., vulnerability model) that combines current and prospective action values with beliefs and expectancies about a partner's behavior. By using the classic investment game (IG)—an economic game thought to capture some important features of trusting behaviors in social interactions—I show how variations of a single parameter of the vulnerability model generates behaviors that can be interpreted as different "trust attitudes". I then show how these behavioral patterns change as a function of an individual's loss aversion and expectations of the partner's behavior. I finally show how the vulnerability model can be easily extended in a novel IG paradigm to investigate inferences on different traits of a partner. In particular, I will focus on benevolence and competence—two character traits that have previously been described as determinants of trustworthiness impressions central to trust. The vulnerability model can be employed as is or as a utility function within more complex Bayesian frameworks to fit participants' behavior in different social environments where actions are associated with subjective values and weighted by individual beliefs about others' behaviors. Hence, the vulnerability model provides an important building block for future theoretical and empirical work across a variety of research fields.

Keywords: trust; investment game; trust game; loss aversion; benevolence; competence; trustworthiness; computational modeling

Citation: Bellucci, G. A Model of Trust. *Games* **2022**, *13*, 39. https://doi.org/10.3390/g13030039

Academic Editors: Cristina Bicchieri and Ulrich Berger

Received: 26 March 2022
Accepted: 7 May 2022
Published: 17 May 2022

1. Introduction

Trust is a ubiquitous experience throughout human life. Trust is required when you ask someone to look over your laptop in a café while you go to the restroom. Trust is required when you talk with your colleague about a bad experience you had with your boss. Trust is required when you describe your health issues to your doctor. Trust has also pivotal consequences on economics and health. For example, trust positively correlates with GDP growth and trust in a doctor correlates with better health outcomes and longer life expectancy [1–3]. The social phenomenon of trust has been dissected by different scientific fields. Philosophers have described the conditions under which trust occurs. Psychologists have captured varying dynamics of trusting behaviors. Economists have studied features of trust in game-theoretic terms. Cognitive neuroscientists have unearthed its neural underpinnings. Thanks to all these efforts, we have gone a long way in understanding the psychological, behavioral, and neural mechanisms underlying trusting behaviors and experiences. Previous research has further distinguished different types of trust based on its target and the mechanisms involved, such as institutional trust, group trust, and interpersonal trust [4–6].

In the following, I will focus on the psychological and computational mechanisms of interpersonal trust. Interpersonal trust refers to the trust people put in another person

and is distinguished from trust in an indistinct group of people (such as football teams) or abstract social entities (such as political parties and institutions). Philosophers conceive of interpersonal trust as an *attitude* people have toward those who they think are trustworthy. In particular, the concept of *reliance* is central to trust for most philosophers, as trust describes those phenomena of social lives in which someone relies on (expected-to-be trustworthy) others for different purposes [7]. Importantly, as trusting individuals make themselves dependent on others, trust implies a form of vulnerability to those trusted (such as vulnerability to betrayal). Finally, the characteristics that make the trustee trustworthy vary as a function of the context and the purposes for which individuals trust.

Psychologists have defined interpersonal trust in similar ways. A very widespread, cross-disciplinary, psychological model of trust defines trust as follows: trust refers to the willingness to accept vulnerability based on positive expectations of the other person's behavior and intentions [8,9]. Mayer et al. [8] also specified that trust implies refraining from exerting social control. Indeed, previous research has shown that the possibility of exerting social control and the active decision of refraining from using it is central to trust [10,11], which likely is a behavioral signature of the willingness to accept vulnerability and might in general be related to the positive consequences of strategic ignorance [12].

In addition, a recent review has discussed two main experimental paradigms that have been implemented across disciplines and have provided us with many insights into the behavioral and neural underpinnings of trusting behaviors [13]. On the one hand, trust has been studied by investigating people's tendency to take advice from different sources. For instance, participants might be required to make inferences on some quantities and can use the opinions of others (advisers) to improve their estimations [14]. These are forms of sequential, interacting paradigms, in which a participant, in the role of adviser, provides pieces of advice that another participant decides to take or discard for solving a monetary task. Here, trust and reciprocity are operationalized by participants' statistics in advice utilization, and they are in general modulated by manipulations of different social characteristics of the advisers, such as their competence, confidence, and kindness [15–18].

On the other hand, trust has been investigated by using another sequential, interacting paradigm, the investment game (IG) [19,20]. In the IG (Figure 1A), two players interact as investor and trustee. The investor receives an initial endowment and makes an investment decision about whether to pass any, some, or all of the received endowment to the partner (i.e., the trustee). The amount, if any, invested is multiple by a multiplier that varies across studies, but is typically three and stationary within a study (except in some special cases; see [21]). After the investor's decision, the multiple amount is given to the trustee, who can decide to return any portion of what they receive on that trial. The investor's decision is thought to capture trust, while the trustee's decision is supposed to capture reciprocity [20,22–25]. At least two general versions of the IG have been used beyond specific modifications of the details of the interacting dynamics or interacting partners. These two IG versions concern the operationalization of trusting behavior and the evolution of the trusting interaction. In particular, trust decisions have been studied either as binary or continuous (on a pre-specified range). In the first case, participants are asked to make an all-or-nothing decision, that is either trust their partner or not [20,24,26–31]. In the second case, participants are asked to share a proportion with different resolutions [19,32–34]. Further, the interaction can be limited to a single round/trial (one-shot IG) or unfold over multiple rounds/trials (multi-round IG). In the last case, participants might know how many rounds/trials they will interact with each other or not. All these differences have been shown to lead to interestingly varying behavioral and neural dynamics [35,36], and recent simulation studies have more formally shown under which conditions trust and trustworthiness in the IG evolve [37,38]. In the following, I will focus on the trust dynamics that occur in the IG—mostly within the context of multiple interactions between the investor and the trustee (e.g., multi-round IG). However, some of the things discussed will also apply paradigmatically to the one-shot IG.

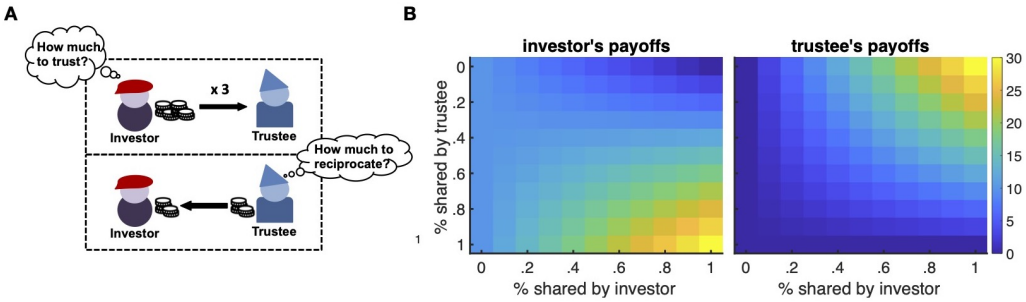

Figure 1. Investment game and payoff structure. The investment game (IG) is a sequential game where the investor receives an initial sum (endowment) and can decide whether to share any of this with the trustee (trust decision). The shared sum is then multiplied (e.g., tripled) and passed on to the trustee, who can decide to share back some of the received amount (reciprocity decision). In multiple one-shot or multi-round IGs, the same trial depicted is repeated multiple times with either a different trustee or the same trustee, respectively (**A**). Investor's and trustee's payoffs at the end of each trial for each possible action (in % of the available sum). The payoff value grid clearly shows the dilemma of the investor as a function of economic interests (payoff values). While the investor is incentivized by the payoff structure to share higher amounts, the trustee is incentivized to keep more of the received amounts (**B**).

Despite many decades of empirical studies on trust in the IG, very little work exists that tries to more formally define the behavior observed in these trusting interactions. This is a general issue in social neuroscience and psychology, where the few computational models that have been implemented and tested have been borrowed from other fields and developed to study other cognitive and learning processes than social mechanisms (such as nonsocial, solo learning processes like in the case of reinforcement learning models) [39]. This is also the case for trust. There exists computational attempts to model people's behaviors in the IG. However, the applied models are borrowed from other fields that developed them for other purposes, such as reinforcement learning models and their Bayesian versions [40,41] or inequality models [42,43] developed for other economic games (e.g., the ultimatum game) [44]. Here, I attempt a new approach recently suggested as a solution to move the field of social neuroscience and psychology forward [39]. That is, I aim to generate a computational model of trust in reciprocity inspired by philosophy, social psychology, behavioral economics, and social neuroscience by mapping verbal models to mathematically tractable quantities that will enable psychological assumptions and mechanisms to be clearly quantified and explicitly tested. Thereby, I will leverage the knowledge of years of observations in social psychology and combine it with a mathematical formalism. For this purpose, I will focus on the investor's dilemma to trust or not to trust their partner in the exemplar interactive environment of the multi-round IG. I will not only show that the proposed model is able to describe some very interesting, paradigmatic behaviors previously observed in the IG, but also make testable predictions about possible behaviors in new versions of the IG that could be used to investigate other trust-relevant social variables.

2. Components of Trust

One essential aspect of trust is the recognition of the other partner's trustworthiness. Despite a willingness to be vulnerable to those who trust being central for trust to occur, trusting partners do in general trust on the basis of positively-valenced expectations of their interacting partner. These expectations are nurtured by direct and indirect knowledge about the other partner, which justifies beliefs about the partner's trustworthiness [45]. Previous research has identified different aspects of a person's behavior and character that evoke impressions of trustworthiness. An influential work on the determinants of trusting behavior lists three main characteristics of a partner for the formation of trustworthiness impressions: (1) ability; (2) benevolence; and (3) integrity [8].

The ability dimension is associated with the skills and competencies of a partner that might be helpful to the trusting individual in reaching her/his goals. These skills might be very general and experience-based, like when asking a friend about the work culture in a different work field, or domain- or situation-specific, like when people rely on a doctor for medical issues, but ask a lawyer for legal counseling. The benevolence domain, on the contrary, is related to the intentionality aspect of the interacting partner. In particular, a trusting partner cares about whether the trustee has positive intentions and attitudes toward his/her. This is probably due to the fact that the vulnerable status a trusting partner accepts to be in is a highly risky and dangerous situation that might jeopardize the trusting partner's social status, reputation, physical integrity, and even personal life. Finally, integrity refers to a moral dimension that describes the degree to which the partner adheres to a set of principles that the trusting partner finds acceptable. This factor is closely associated with the moral character of the trustee and refers to the trustee's behavioral consistency in his/her congruence to a determined set of values [13].

At least two of these determinants of trust have previously been successfully studied in the IG. The trustee's decision of sending back a proportion of the received amount (i.e., reciprocity) is interpreted as a sign of benevolence through which the trustee signals to the investor that he/she has good intentions. This compels trustees to back-transfer money to the investor even when it is not strictly economically advantageous for them [32]. For instance, decreases in reciprocal behavior, especially after increases in trust levels, disrupt cooperation [46]. Further, unconditional kindness, but neither positive nor negative reciprocity, has strongly been associated with trusting behaviors [47]. Similarly, impressions of a partner's moral and immoral character from previous social interactions have long-lasting effects on individuals' decisions to trust in subsequent encounters [48]. Finally, individuals with strong moral characteristics are not only perceived to be more trustworthy but also more likely to be trusted, which is likely due to the fact that they are also believed to be more likely to reciprocate [49].

Despite a wide array of empirical investigations on the effects of the intentional and moral dimensions on trust in the IG, little is known about how the ability dimension influences people's trust decisions. One of the difficulties is definitely related to the fact that the IG does not bend itself that easily to investigations of the effects of the trustee's competence on the investor's sharing decisions. In the following, I will show how the *vulnerability model* formalizes trusting behavior in the IG and captures the effects of the expectations of trustworthiness on trust. In particular, I will show how such expectations are formed based on the impressions of the trustee's benevolence arising from reciprocal behavior, which signals the trustee's good intentions and willingness to cooperate. Moreover, I will show how the vulnerability model parameterizes trust attitudes as the theorized, weighting trade-off between the acceptance of vulnerability and positive expectations of the partner's behavior [8,9]. I will then describe a novel experimental setting by employing a modification to the classic IG in order to study the effects of competence on trust. Thereby, I will demonstrate how the vulnerability model can be easily extended to capture the effects of the ability dimension on trust decisions.

3. Formalism

The classic IG is a two-person, sequential task played over a specific horizon T with a predetermined number of trials (Figure 1A). If $T = 1$, the IG is played over a single round (or trial) and is called the one-shot IG. If $T > 1$, the IG is played over multiple rounds (or trials) and is known as the multi-round IG. Additionally, both versions can be played with a single or different partners. Hence, a one-shot IG can be played sequentially with different trustees, in which case we have a multiple one-shot IG. Similarly, participants as investors can play multiple rounds of the IG with different trustees (i.e., a multiple multi-round IG). The following formalism applies in principle to all these cases with adequate modifications to adapt it to the social environment in which the behavior is studied. In my examples, I will work with a single multi-round IG with $T = 10$ (a very often-used number of trials in

experimental research; e.g., [32]), where an investor interacts with the same trustee over the whole length of the experiment.

Investors receive an initial endowment e_t and invest a proportion of it (i_t) in trial t. The amount, if any, shared by the investor (i.e., s_t, which is $s_t = e_t i_t$) is multiplied by a factor $\mu > 1$. This multiplier varies across studies, but is stationary within a study (except in some special cases; see [21]) and is generally $\mu = 3$. After the investor's decision, the multiple amount is given to the trustee, who can decide to return a proportion r_t of the endowment $\eta_t = \mu s_t$ that they receive on that trial. As often done in experimental research, I will imagine that the currency of the endowment is in monetary units (MUs), which can be translated to different currencies with any mapping function. Moreover, I will assume e_t to be $e_t = 10$ MUs as the investor's endowment at the beginning of each new trial, again as often done in experimental research. Since the investor does not know the trustee's return at the time of choice, the proportion returned by the trustee at this time will be denoted as the random variable R_t.

Moreover, I will assume a finite discretization of the action space, following empirical research that allows investors and trustees to make only a subset of predefined sharing actions. For instance, binary IGs allow participants to make one of only two possible actions. Some other studies allow participants more actions corresponding to some submultiples of the available amount (a very common version is one with multiples of four from 0 until the total amount of 12). In studies with $e_t = 10$, investors are allowed to share any available unit, which is equivalent to letting them share any portion of the received amount (from 0 to the whole amount) in steps of 10%. In such studies, the trustees are endowed with the same action rule, so that they also can share any unit of the received triple amount. However, previous computational work has considered more discrete action spaces for trustees who can have a vaster action space. A convenient practice is to consider similar action spaces for both investors and trustees [50]. This implies that trustees will be allowed to share only predefined proportions of the available amount. Given the results in previous research highlighting the importance of specific proportions such as 30%, 40%, and 50% in economic games [51–54] and the effects of reciprocal behavior in social interactions [16], I will consider the same proportional action space for both trustees and investors. Hence, if A is the total number of possible actions for investors and trustees, $A = 11$ represents the range of available actions for the two players from sharing nothing to the complete available endowment in steps of 10%.

Figure 1B provides a snapshot of the dilemma of the investor as a function of the economic interests involved (payoff values). While the investor is incentivized by the payoff structure to share higher amounts and hopes for higher returns by the trustee, the trustee hopes for higher shares, but is incentivized to keep higher portions of the received amounts. How does the investor solve this problem? In particular, as the trustee only loses by returning money to the investor, the economic solution is for the investor not to return any money. However, plenty of empirical studies have shown that both in one-shot and multi-round IGs, people deviate from this optimal solution and do share some amounts. The general explanation of these behavioral patterns is to think of utility as subjective and, hence, modulating different components of the utility computation. For instance, the investor's payoffs might be reduced by considerations of a possible betrayal, leading to lower trust levels, while the trustee's payoffs might be reduced by the social disutility of a defection (implying higher back-transfers as a consequence). Further, in multiple interactions, individuals might learn and make inferences about others' intentions and behavior to reduce the uncertainty associated with their sequential decisions.

3.1. Inequity Aversion Model

A well-known formalization of action utilities for social behaviors in simple two-person interactions is provided by the inequity aversion model proposed by [44]. Although the inequity aversion model was originally developed to model participants' fair behavior in the ultimatum game, it has been used in other social experimental paradigms in which

it is reasonable to assume fairness concerns, such as the IG [34,42,50]. For this and for completeness reasons, I will here briefly discuss the inequity aversion model and its predictions. Later, I will compare them to the proposed vulnerability model.

The inequity aversion model aims to capture the well-known observation that individuals dislike outcomes that are perceived as inequitable and that they do so more when such inequity is at their disadvantage [55]. Hence, inequity aversion can be *self-centered* when individuals care about their own material payoff relative to that of others and increases with increasing disadvantageous inequity; or *other-centered* when individuals care about others' payoffs relative to their own and increases with increasing advantageous inequity [13,44]. Based on this definition, let x_t^n be the payoff of person n at time t and x_t^m be the payoff of person m at time t, which both depend on both players' shares i_t, and r_t. The utility of person n at time t is then

$$u_t^n(i_t, r_t) = x_t^n - \zeta^n max(x_t^n - x_t^m, 0) - \iota^n max(x_t^m - x_t^n, 0) \qquad m \neq n, \tag{1}$$

where ζ^n and ι^n are subject-specific parameters (here of person n) that quantify the degree of disutility derived from advantageous (guilt) and disadvantageous (envy) inequality aversion, respectively. Thus, Equation (1) describes the utility of an individual n as the sum of a utility derived by their current payoff x_t^n and two potential disutilities—one when person n ends up having more than his/her partner (i.e., advantageous inequality when $x_t^n - x_t^m > 0$) and one when person n ends up having less than his/her partner (i.e., disadvantageous inequality when $x_t^m - x_t^n > 0$). When person n has less than person m (i.e., $x_t^n < x_t^m$), person n is considered to be envious of person m, while when person n has more than person m (i.e., $x_t^n > x_t^m$), person n is regarded as feeling guilty of having more than person m. In both cases, the current material utility x_t^n is uniquely reduced by one of these two disutilities and the parameters ζ and ι capture individual differences in people's sensitivity to these two disutilities. The optimal scenario is represented by the case when the payoff difference between person n and person m is zero, which captures people's universal tendency to fairness [56,57]. Moreover, since people have been shown to be more averse to disadvantageous than advantageous inequity [55], the model parameters have to satisfy the constraint $\iota \geq \zeta$.

In the context of the IG, if we denote by b_t the absolute amount back-transferred by the trustee from the received multiple amount (i.e., $b_t = \mu s_t r_t$) at time t, x_t^n stands for the investor's payoff and is equal to

$$x_t^n = e_t - s_t + b_t.$$

Consequently, x_t^m represents the trustee's payoff and is equal to

$$x_t^m = \mu s_t - b_t,$$

which is the amount the trustee keeps from the multiple amounts received after reciprocity.

Finally, results from empirical works suggest that investors make their decisions based on a stochastic policy. If $u_t(i_j, r_t)$ is the utility of the investor for making investment i_j and receiving return r_t at time t, the probability of i_j is

$$p(i_j) = \frac{\exp(\chi u_t(i_j, r_t))}{\sum_{a=1}^{A} \exp(\chi u_t(i_a, r_t))}, \tag{2}$$

where χ is the inverse temperature of the softmax function that captures randomness in participants' decision policy. Specifically, lower χ values reflect more diffuse and variable choices.

After the definition of a utility function and policy for the investor, we can examine the action values and probabilities across the combinations of i and r (investor's and trustee's available actions). From Figure 2, we can easily observe that the model predicts that guilt-

free investors ($\zeta = 0$) prefer a cooperative interaction with high investments and returns that would maximize their payoffs. Similarly, when paired with uncooperative trustees, guilt-free investors prefer sharing nothing. However, when investors are more and more averse to advantageous inequality (higher ζ values), the model predicts that they should prefer action combinations that yield the most equitable outcomes, which would reflect their increasing fairness concerns. Hence, for interactions with uncooperative trustees, investor's policy shifts toward higher share proportions (e.g., 20–30%), which should even up the initial disparity caused by the endowment $e_t^n = 10$ provided to the investors at the start of the trial t (as the trustee does not receive any initial endowment, so $e_t^m = 0$). This pattern of policy shift reaches the extremes for very high ζ values, where we see that investors prefer cooperative trustees that share around half of what they received, with advantageous payoffs increasingly disliked.

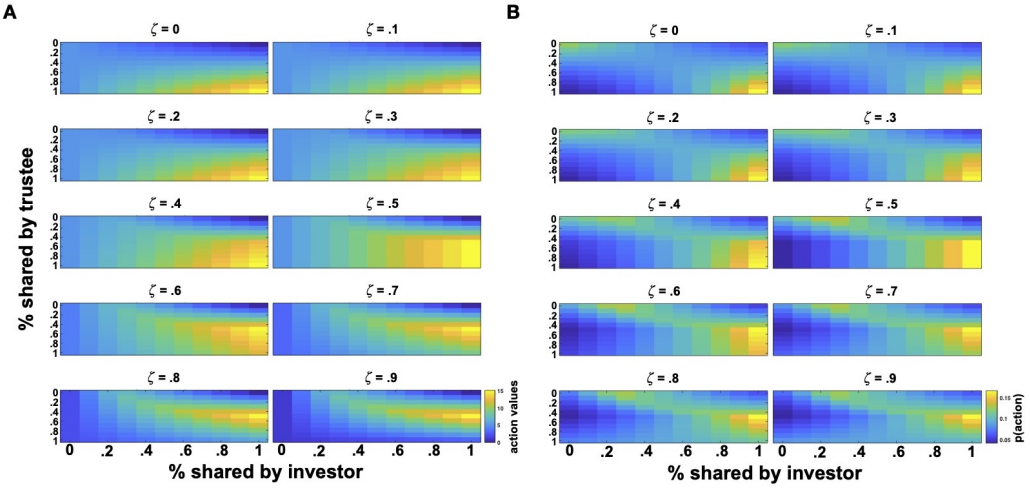

Figure 2. Action values and probabilities with varying ζ. Action values (**A**) and action probabilities (**B**) for investors with different ζ parameters (i.e., guilt or advantageous inequality) as computed according to the inequity aversion model. In all plots, ι is set to 0. For action probability estimation, the χ parameter of investors' policy is set to 0.2.

Variations in ι values generate a very similar policy space with some subtle differences (Figure 3). For low ι levels, action policies are similar to those generated by low ζ values with the difference that aversion to disadvantageous inequality more strongly favors lower or no investments with uncooperative trustees (higher action probabilities). For higher ι values, the policy does not change much, but sharpens over the two self-interested extremes with action combinations that maximize the investor's payoffs. That is, higher investments for high returns in cooperative interactions and increasingly lower investments for small returns in uncooperative interactions are enforced by high ι values (almost a tit-for-tat or copying strategy). This implies that higher ι values yield policies most divergent from those generated by ζ values, which, on the contrary, shift the policy toward a slightly different action space.

Given that $\iota \geq \zeta$, the action policy of the inequity aversion model persists on self-interested, payoff-maximizing strategies with very little explorations of other policy regions. This makes it hard to see how inequity-averse investors differ from pure money-maximizers. This is in accord with a previous criticism of the inequity aversion model (see for instance [58]). Indeed, previous work has used a reduced version of the model that only takes into account advantageous inequity [50]. However, also in this reduced version, only a small subset of ζ parameters have been found to generate distinguishable behavioral predictions [50]. Interestingly, these parameters fall within the same region of the parameter space from which [44] drew their own parameter values in their original paper.

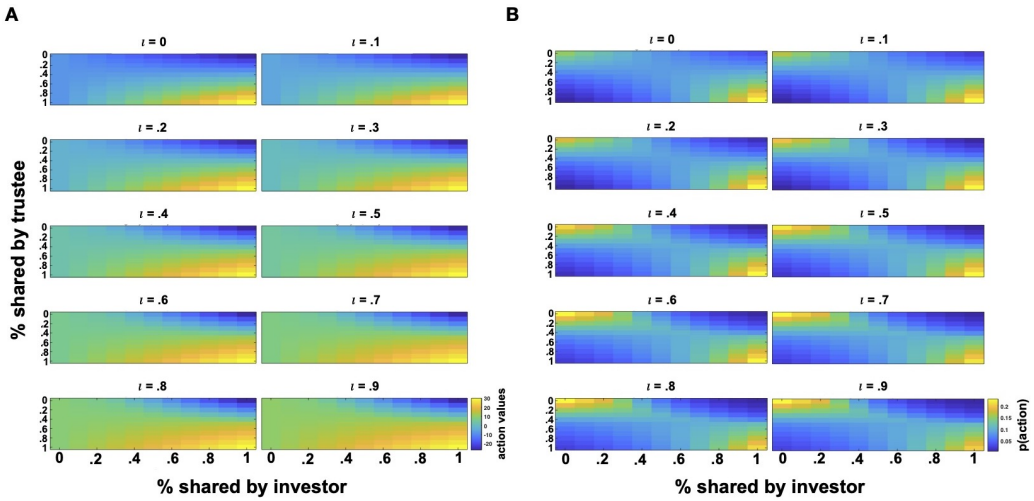

Figure 3. Action values and probabilities with varying ι. Action values (**A**) and action probabilities (**B**) for investors with different ι parameters (i.e., envy or disadvantageous inequality) as computed according to the inequity aversion model. In all plots, ζ is set to 0. For action probability estimation, the χ parameter of investors' policy is set to 0.2.

Moreover, even though fairness concerns are reasonable in the context of the IG, the inequity aversion model is in its essence an outcome-based model that might be more suitable in contexts where a person's decision outcomes are more salient. On the contrary, in the IG (especially in versions with non-binary action spaces), the actions of the interacting partner become salient. This explains, for example, why even for low investments, trustees still share back some money and, in particular, try to send back at least as much as shared by the investor (which amounts to approximately 30% of their available endowment for $\mu = 3$), even though this has the effect of increasing the payoff difference at their disadvantage. Different reasons can be provided to explain such a behavior. For instance, they might be acting so to remunerate the investor for their trust and/or to signal willingness to cooperate (especially in multi-round settings). Indeed, when investors do not see such willingness from the trustee (even if this means he/she needs to put himself/herself in a disadvantageous situation), he/she refrains from trusting any longer. This is probably because the investor is not focusing on minimizing the payoff difference between himself/herself and the trustee, but rather intends to try the trustee for informative inferences on how his/her partner would behave in the future. Among other possible reasons, it is clear that an outcome-based model falls short of adequate explanations for the behavioral patterns observed in the IG.

3.2. Vulnerability Model

The vulnerability model more closely captures the theorized psychological and computational components involved in a decision to trust, both in general and specifically in the IG. In the context of a single exchange in the IG, vulnerability is associated with the investment s_t, since the investor stands to lose this. Such a possibility is however contingent on what the trustee actually returns. However, when the investor makes his/her decision, he/she only knows i_t, but does not know what the trustee is going to return. Here, expectations of the trustee's return function as a proxy for the trustee's future behavior. Hence, the investor needs to trade-off the possibility of losing the investment against his/her expectations of back-transfers (i.e., the trustee's future reciprocal behavior). Psychologically speaking, the investment s_t is likely framed as a potential "loss" with an immediate negative utility for the investor. On the contrary, the expectation of the trustee's reciprocity r_t is likely framed as a potential positive "gain", as it represents what the investor could receive from the trustee (i.e., the return b_t). The investor thus ponders which proportion of the multiple received endowment (i.e., $\eta_t = \mu s_t$) the trustee will choose to return. This computation

enters into the investor's estimation of his/her utility $u_t(i_t, r_t)$ at time t. At the time of the investment, the investor does not know the proportion preferred by the trustee and has to consider the distribution over all possible proportional returns $p(R_t = r_t|\mathcal{D}_{t-1})$, where $\mathcal{D}_{t-1}$ includes the history of exchanges (investor's and trustee's actions) up to time t. As proportions naturally range between 0 and 1 so that $r_t = [0, 1]$, I' will assume that the probability over $p(R_t = r_t|\mathcal{D}_{t-1})$ has the form of a beta distribution. Hence, the investor's utility of choosing a particular proportion i at time t is

$$u_t(i_t, \mathcal{D}_{t-1}) = \mathbb{E}[u_t(i_t, r_t)|\mathcal{D}_{t-1}] \tag{3}$$

$$= e_t - s_t + \eta_t \, \mathbb{E}[r_t|\mathcal{D}_{t-1}] \tag{4}$$

$$= e_t - s_t + \eta_t \int_{r_t=0}^{r_t=1} dr_t \, r_t \, Beta(r_t; \alpha_t^r(\mathcal{D}_{t-1}), \beta_t^r(\mathcal{D}_{t-1})). \tag{5}$$

$Beta(r_t; \alpha_t^r(\mathcal{D}_{t-1}), \beta_t^r(\mathcal{D}_{t-1}))$ is the beta distribution over all possible proportional returns r_t and represents the probability of the trustee's return at time t based on the observed history of returns until time t ($\mathcal{D}_{t-1}$). This beta distribution is described by two parameters α^r and β^r as follows:

$$Beta(r; \alpha^r, \beta^r) = p(r|\alpha^r, \beta^r) \propto r^{\alpha^r - 1}(1 - r)^{\beta^r - 1}, \tag{6}$$

with r being the trustee's proportional returns. In multi-round settings, expectations of returns are updated via Bayesian updating on every trial after observing a new proportional return. Hence, the investor's posterior probability over r at the next time step $t + 1$ becomes $Beta(r_{t+1}; \alpha_{t+1}^r(\mathcal{D}_t), \beta_{t+1}^r(\mathcal{D}_t))$, where $\alpha_{t+1}^r = \alpha_t^r + Nr_t$ and $\beta_{t+1}^r = \beta_t^r + N(1 - r_t)$. N determines the update rate for new observations and is assumed to be $N > 0$ with the special case of $N = 0$ when no updating of the partner's behavior takes place.

Apart from being likely closer to the psychological definition of trust, this additive temporal decomposition is also computationally more convenient than a formalization that works with payoffs (like in the inequity aversion model), since it allows considering a variety of psychological factors such as loss aversion [59,60] or savoring and dread [61–63] associated with the duration of vulnerability, which have been previously found to affect the computational processes underlying decisions in different decision-making settings.

Now, let us follow the philosophical and psychological traditions that formalize trust as a general *attitude* to make oneself vulnerable to others [9,64] and operationalize this attitude as a stationary weighting preference (note, however, that trust attitudes might also change over time). In the context of our formalism, this translates into a competing weighting between the negative consequences of an act of trust (i.e., the loss of the investment generated by the investor's action i_t) and the expectations of the partner's trustworthiness (utility generated by the proportional return chosen by the trustee, i.e., $\eta_t \, \mathbb{E}[r_t|\mathcal{D}_{t-1}]$). This implies that, beyond all the reasons a trustee might provide to the investor to favor future expectations of returns over the avoidance of vulnerability, investors further differ in their weighting preferences of these two components. For example, facing the same amount of evidence about a partner's trustworthiness, some individuals might be ready to trust the partner, while others might still remain reluctant and prefer distrusting. Hence, the investor's utility for a decision to trust is envisioned as a competition between a preference for loss aversion and a preference for trustworthiness expectations as follows:

$$u_t(i_t, \mathcal{D}_{t-1}) = e_t - (1 - \tau)s_t + \tau(\eta_t \, \mathbb{E}[r_t|\mathcal{D}_{t-1}]). \tag{7}$$

where τ is an individual-specific free parameter that weights the positive expectations of trustworthiness against the negative utility of vulnerability, and hence represents a propensity to trust. This parameter τ takes values between 0 and 1 ($0 \leq \tau \leq 1$), and in the current implementation, I assume it to be fixed for each investor over the short time period of a game (to implement a stationary weighting preference for the investor). In this respect, τ can be thought of as a stable personality trait of the investor, similar to

trust conceptualizations in psychology [65,66]. However, extensions to include a variable parameter are feasible and can capture other differences in trusting behavior observed in empirical work such as those depending on the social context (e.g., in-group vs. out-group), changing statistics of the environment (e.g., volatile vs. stable environment), or a person's age [4,5,40,43].

Like for the inequity aversion model, investors are assumed to make their decisions based on a stochastic policy as in Equation (2), which can capture idiosyncrasies in the decision-making behaviors of human agents. Figure 4 showcases the action values and probabilities generated by the vulnerability model for different propensities to trust (i.e., τ parameter). For very low τ parameters, the investor's policy is centered on low proportions of sharing and barely changes as a function of the trustee's actions. For the extreme case of $\tau = 0$, investors do not put any weight on their expectations of the partner's trustworthiness and the model makes predictions as if those expectations were 0 (as if the investor knew the trustee will not share anything). For this and similar cases of very low τ parameters (e.g., $\tau = 0.1$), investors give little weight on the potential benefits derived from the partner's behavior and focus on the immediate loss implied by the investment, leading to very similar action probabilities over the trustee's action space and showing very distrustful behavior, as sharing nothing or very low proportions turns out to be the preferred strategy. Consequently, we can think of these low τ values as generating behavioral solutions similar to those predicted by economic theories, as distrustful investors with these τ values will not make themselves vulnerable to the trustee and prefer keeping the received endowment e_t.

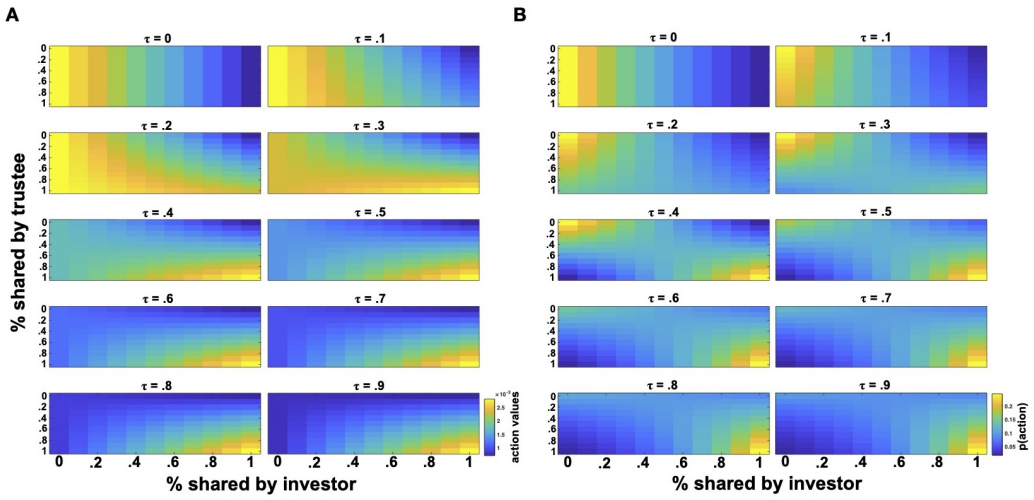

Figure 4. Action values and probabilities with varying τ. Action values (**A**) and action probabilities (**B**) for investors with different τ parameters (i.e., propensity to trust or accept vulnerability) as computed according to the vulnerability model. For action probability estimation, the χ parameter of investors' policy is set to 0.2.

On the contrary, for increasingly higher τ values, we see a shift of the policy toward higher sharing proportions, leading to higher trust levels and more cooperative behaviors. In particular, higher τ values shift the investor's preferences toward a more cooperative action space with investors preferring higher investments when interacting with cooperative trustees or no investments if their trustee is uncooperative (e.g., for $\tau = 0.4$ or $\tau = 0.5$ in Figure 4). Comparing the action probabilities of the inequity aversion model with those generated by the vulnerability model, we observe that action policies for $\tau \approx 0.4$ are very similar to those of envious investors in Figure 3, while action policies for $\tau > 0.5$ are similar to those of mildly guilty or guilt-free investors in Figure 2.

These descriptive observations indicate that the vulnerability model is able not only to explore different regions of action policies, but also to cover parts of the action policy space generated by the inequity aversion model. Hence, investors with a weak trust attitude generate behaviors that, in the context of the inequity aversion model, can be interpreted

as signaling "envy", while investors with a strong trust attitude will manifest behaviors that, according to the inequity aversion model, will be more likely categorized as signaling "guilt". Moreover, when $\tau = 0.5$, the investor maximizes his/her expected payoff by maximizing the expected utility function as given by Equation (7). This is a special case, as the same does not occur for other values of τ. In other words, the corresponding level of trust can be interpreted as being economically rational given the information an investor has. The interesting aspect of the vulnerability model is that it can further capture forms of gullibility, for instance in the case of $\tau = 0.9$, where the investor continues to transfer high shares even in the face of low proportional returns, making his/her subject to exploitation by the trustee.

Hence, the vulnerability model is able to cover a wide range of trusting behaviors, from suspicion (low τ values) through rational and trustful behaviors (τ values in the middle range) to unreasonably inflated/exaggerated trust or gullibility (high τ values). Further, as the vulnerability model describes the investor's utility function as incorporating expectations of the partner, it entails a learning mechanism that naturally allows the investor to adapt the realized trusting behavior to the observed behavior of the trustee. The updating of such expectations over time instantiates a learning mechanism, which can be implemented in different ways. In this paper, I will consider an unsophisticated learning mechanism based on the history of observations of returns that updates the beta distribution representing the investor's expectations of the trustee's trustworthiness. However, more sophisticated learning mechanisms can be used where the probability distribution over the trustee's behavior does not only depend on the history of observations, but also on a generative model of the partner's behavior for planning (see the Discussion).

4. Results

In the following, I more closely investigate which behaviors the vulnerability model generates in a multi-round IG and what predictions it makes. In particular, I consider two paradigmatic types of interactions (i.e., breaches of trust and repairs of trust) with two investor types (trustful and distrustful) and two different psychological traits of the investor (i.e., his/her a priori social expectations and loss aversion). After having considered these dynamics that have been previously described in the experimental findings using the classic multi-round IG, I will turn to consider a novel social environment with a modification of the multi-round IG that allows disentangling expectations of benevolence (belonging to the intentionality dimension of the ABI model discussed above) from expectations of competence (belonging to the ability dimension of the ABI model). I will hence show how the vulnerability model can be extended to mimic learning of these two features of the trustees' trustworthiness (i.e., benevolence and competence) and how different levels of benevolence and competence lead to different predictions of behavior as formalized by the vulnerability model.

4.1. Breach of Trust

I will first consider cases of breaches of trust. Let us suppose that the trustee starts off cooperatively and changes his/her strategy shortly after to attempt to reap the benefits derived from the investor's trust. As shown in Figure 5, the initial cooperativeness by the trustees induces the investor to increase his/her investments, because the investor would benefit greatly from high investments with cooperative trustees. On the contrary, a myopic, unsophisticated trustee would of course gain the most out the interaction at a given time t by denying reciprocation, especially after a high commitment from the investor (i.e., high shares). Assuming uninformative priors over the trustee's possible reciprocal behaviors (i.e., $\alpha^r = 1.2$ and $\beta^r = 1.2$), the investor starts with a relatively costly, but socially acceptable strategy, that is sharing half of the initial endowment. The first trials allow the investor to try the trustee's cooperativeness. Hence, in the first two trials, the investor applies a default policy of sharing 50% of the available amount. Blind application of a default policy is consistent with empirical work and with what the previous literature has presumed about

the underlying cognitive mechanisms of inexperienced subjects [54,58]. In the paradigmatic example here considered, the trustee's cooperativeness was operationalized by back-shares of 50% of the received amount. This strategy is not consistent with economic principles of inequity minimization, as in most cases, the trustee consistently ends up with a smaller amount than the one owned by the investor. Hence, from an economic prospective, the trustee should be "envious" of the investor's higher amount and avoid the realization of such an outcome. However, not only has this strategy been often observed in empirical studies, it also guarantees that the final payoff of the investor is higher than his/her initial endowment. Hence, through this strategy, the trustee can reward the investor for his/her share, and the investor realizes that sharing is advantageous and desirable. On the contrary, when the investor shares 50% of his/her initial endowment and the trustee realizes equal outcomes for both players by reciprocating with a back-transfer of one third of the received amount, the interaction turns out unfavorable for the investor, who has risked the investment for no extra gain, while the trustee signals "ingratitude" to the investor's risk. Hence, the trustee should start off with a strategy that might even be slightly unfavorable to them at the beginning and adjust it according to how the investor reacts to their reciprocity.

In the paradigmatic example considered here, the investor receives positive evidence from the trustee in the first two trials and updates his/her expectations of the trustee's returns accordingly (dashed grey lines in Figure 5 represent changes in the investor's expectations). As a consequence, the vulnerability model in $t = 3$ predicts an increase of the investor's shares. Importantly, this increase in shares is predicted to be of different magnitudes depending on the investor's attitude to trust (i.e., τ parameter). In Figure 5A, we see that investors with a relatively low τ value (i.e., "distrustful investors" with $\tau = 0.4$) increase their investment levels significantly less than investors with a relatively high τ value (i.e., "trustful investors" with $\tau = 0.6$).

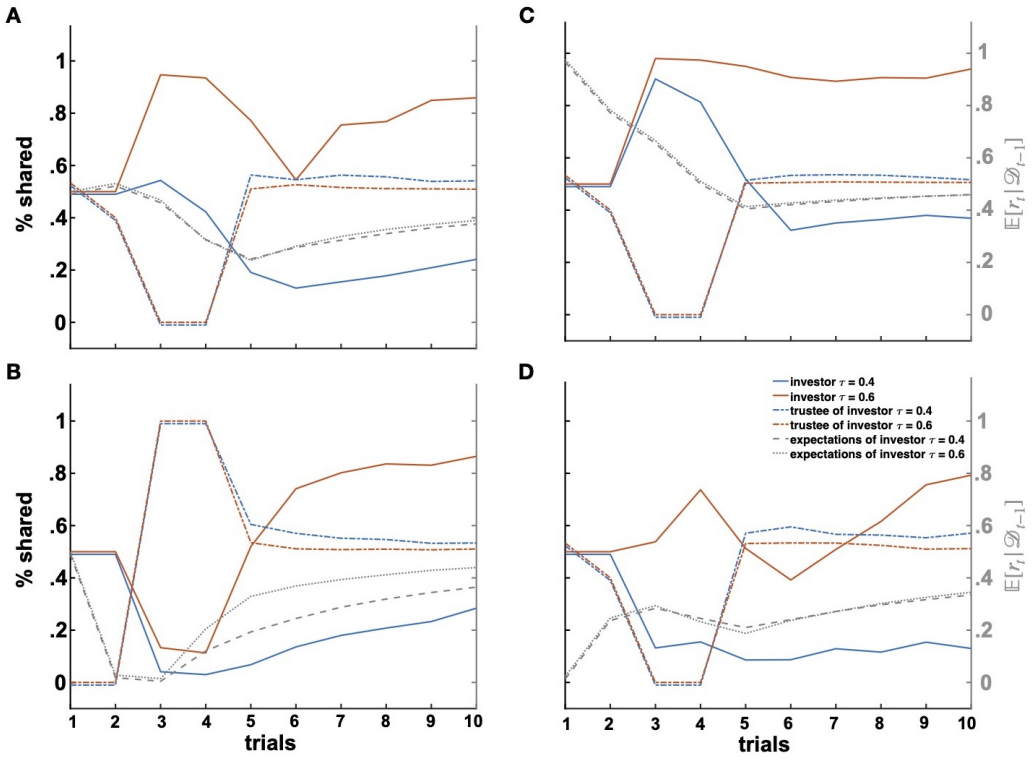

Figure 5. Breach and repair of trust. Effects of breach of trust (Trials 3–4) with trustful (red line and $\tau = 0.6$) and distrustful (blue line and $\tau = 0.4$) investors (**A**). Repair of trust (**B**) after initial uncooperativeness (Trials 1–2). Effects of positive (**C**) and negative (**D**) expectations on repair of trust.

Next, in trials $t = 3$ and $t = 4$, the trustee breaches the investor's trust by keeping the entire amount he/she received (i.e., η_t). Thereby, the trustee violates the investor's expectations of returns. As a result, the investors reduce their investments in the subsequent trials. Already in $t = 4$, we observe a reduction in shares proportional to the investors' attitude to trust, with stronger investment drops for distrustful investors. Such decreases in investment levels have previously been observed after breaches of trust and have been interpreted as an attempt by the investor to signal disapproval or irritation [32,50]. Importantly, variations in the reactions to breaches of trust have been associated with the investor's personality. For instance, while some investors ("conditionally" trustful or strategic individuals) react by immediately decreasing their investment, other investors ("unconditionally" trustful or benevolent individuals) keep investing the same amount if not more, at least for a short period after the initial breach of trust, as if they wanted to give their partners further chances to amend the relationship [24,32]. The different τ parameter values mimic these different reactions to breaches of trust, as trustful investors seem to behave in a more benevolent and forgiving fashion as compared to distrustful investors. A second breach of trust in $t = 4$ induces a further decrease in investments in trial $t = 5$, this time larger than the previous one for both investors' types, as the posterior expectations of proportional returns more strongly deviate from cooperative shares (at around 25% for both investors). Again, the degree of this second reduction in investments was modulated by the investor's trust attitude, leaving trustful investors more exposed to exploitation. Taken together, the model predicts that breaches of trust occur when a distrustful investor meets an uncooperative trustee, while an uncooperative trustee might reap the most benefit when breaching the trust of a trustful investor who is the most exposed to exploitation.

4.2. Repair of Trust

Another paradigmatic scenario is represented by cases of trust repair. In the example depicted in Figure 5A, the trustee attempts to repair trust by following a 50% share strategy from trial $t = 5$ onward, after uncooperative behavior. As can be seen in Figure 5A, neither the trustful nor the distrustful investor are at first responsive to such a signal of cooperation by the trustee, despite the fact that they have definitely learned this change of the trustee's behavioral and their expectations of returns in trial $t = 6$ are more positive as compared to trial $t = 5$. It takes two trials for both trustful and distrustful investors to reverse their investment behaviors and start sharing increasingly higher amounts. Again, the degree to which the trustee is able to restore the investor's trust depends on the investor's trust attitude. We can see in trial $t = 7$ that despite similar expectations of returns, the trustful investor increases his/her proportional shares by more than 10%, while the distrustful investor manifests a much smaller increase in shares by a few percentage points. Importantly, trust fully recovers only for interactions with trustful investors who share amounts comparable to those at the beginning of the interaction (e.g., trial $t = 3$). A slight increase in investments can be observed for distrustful investors as well, but the investment levels remain low, especially when compared to their investment amounts at trial $t = 3$ when the distrustful investors had positive expectations of returns. Hence, the vulnerability model predicts that repair of trust is harder and cannot be fully restored when interacting with distrustful partners who manifested a "resentful" behavior for a longer period of time.

Similar results are observed when supposing that the trustee starts the interaction by being uncooperative (Figure 5B). Given that now, there are no positive expectations alleviating the negative consequences of uncooperative behavior, we observe an almost complete refusal of further investments by both the distrustful and the trustful investors (Figure 5B, $t = 3$). In the next two trials ($t = 3$ and $t = 4$), the trustee repairs trust by sending back the whole amount received and later on by enforcing an equal-split rule (from $t = 5$ onward). As trustful investors are slightly more likely to send some higher amounts to the trustee, they more promptly capture these cooperation signals by the trustee, update their expectations, and increase their investments, so that already in trial

$t = 5$, they are willing to share around 50% of their initial endowment. Since the trustee maintains a reliably cooperative behavior, such an increase proves to be steady until the end of the interaction. On the contrary and similar to the previous case, investments of distrustful investors increase more slowly and remain at very low levels, with investors sharing only around 20% of their initial endowment by the end of the interaction (Figure 5B, $t = 10$). Again, this occurred despite expectations of returns very similar to those of trustful investors.

Two observations can be made here related to two important predictions of the vulnerability model. On the one hand, distrustful investors need more behavioral evidence from the trustee to allow for a repair of trust and cooperation and for the establishment of a new phase of the social interaction. For instance, distrustful investors maintain their trust levels practically unchanged in trial $t = 4$ after having seen the trustee's increase in reciprocity in the previous trial, while trustful investors are more receptive and responsive (Figure 5B). This is because distrustful investors put more weight on the potential losses implied by a trust decision and need a bigger change in their expectations of returns to counterbalance the disutility they might derive from trust.

Second, as distrustful investors keep their investment levels low to avoid losses, they give little opportunities to trustees to demonstrate their repentance and behavior change. That is, distrustful investors end up gathering less evidence on the trustee's behavior, leading to expectations of returns that greatly diverge from the actual return strategy adopted by their trustee. Indeed, investors can "see" the trustee's behavior (and thus, make inferences on his/her return intentions) only if they give their partners the opportunity to make a back-transfer. This phenomenon is also reflected by the evolution of the investor's expectations, where it can be seen that expectations of returns by distrustful investors were more negative than those held by trustful investors. This is equivalent to real-life situations in which people can get to know someone else only if they allow for interactions with that person to occur (which does not happen if that person is intentionally avoided). Hence, the vulnerability model predicts that distrustful people more strongly avoid interactions with partners who manifest signs of uncooperative behaviors. If such a strategy better protects distrustful people from exploitation, it also predicts that their inferences on other people's intentions and personality should be less accurate. Taken together, the vulnerability model predicts that repair of cooperation is only possible if the investor is trustful enough to allow the trustee to demonstrate his/her willingness to cooperate again.

4.3. Expectation Change

In the previous paradigmatic cases, we investigated the effects of investors' trust attitudes on their trust (investment) levels. Thereby, we have not only seen that a stronger trust attitude might be fraught with the danger of exploitation by the trustee, we also observed that it allows for higher levels of cooperation in cases of repair of trust. Further, we also had a glimpse at how restoration of cooperation was partly driven by the investor's expectations of returns. Now, I will more closely examine the role of these expectations in the evolution of the interactive dynamics between investor and trustee.

First, instead of assuming uninformative priors over the trustee's reciprocity, let us provide the investors with different expectations of returns (Figure 5C,D). I will consider the effects of positive expectations by setting the α^r and β^r parameters of the beta distribution over the trustee's reciprocal behavior to $\alpha^r = 50$ and $\beta^r = 1.2$, leading to a prior expected proportional return of $\mathbb{E}[r_t] = 0.98$. These are quite unrealistic expectations of returns, but will help us better see the differences in behavior between trustful and distrustful investors endowed with such positive expectations. In Figure 5C, we see that extremely positive expectations lead to high levels of optimism (if not even gullibility). Importantly, breaches of trust in $t = 3$ and $t = 4$ only marginally impact subsequent investment levels, especially for trustful investors, and investment levels manifest little variability until the end of the interaction for trustful investors. This is because, despite a constant decrease in expectations of returns after breaches of trust, expectations do not go below a critical threshold that

could have driven investors' shares down. On the contrary, distrustful investors keep decreasing their investment amounts throughout the entire set of interactions—a very common behavioral finding that is not only restricted to the IG, but has also been observed in other cooperative games such as the Public Good Games [67–70].

Second, let us take a closer look at the opposite case of very negative expectations of proportional returns. Now, I will consider the effects of negative expectations by setting the α^r and β^r parameters to $\alpha^r = 1.2$ and $\beta^r = 50$, leading to a prior expected proportional return of $\mathbb{E}[r_t] = 0.02$. Figure 5D demonstrates a very detrimental effect of negative expectations on the investors' behavior. The effect is so strong that it overshadows the impact of breaches of trust. Again, the investor's trust attitude slightly mitigates the effects of these negative expectations. As before, this is due to a higher number of opportunities for the trustee to show reciprocal behavioral to trustful investors. As a consequence, trustful investors can form more accurate expectations of their partners and adapt their investment levels accordingly. Due to the initially very negative expectations, however, such an increase in investments by trustful investors takes a while to take off and can be observed only in the last trials of the interactions. On the contrary, cooperation cannot be recovered with distrustful investors.

Hence, the vulnerability model shows interesting dynamics that occur as a function of expectation change. Importantly, while psychological models describe trust only as a function of positive expectations [8,9], the proposed computational model more generally captures how trust changes as a function of varying expectations of the partner and differences in his/her behaviors. Moreover, expectations that strongly diverge from reality may help capture degenerate behaviors indicative of varying clinical symptoms.

4.4. Loss Aversion

Until now, the vulnerability model has assumed a linear utility function. However, potential losses and expected returns can be also combined in non-linear ways. Previous psychological and economic research on people's subjective utilities has demonstrated that losses and gains are not equally weighted in risky and value-based decisions [60,71–73]. In particular, losses are better described by a concave function, while gains are better described by a convex function. This corresponds to an overweighting of losses and an underweighting of gains, which can be interpreted as evidence for loss-aversion behaviors requiring much higher potential gains to counterbalance the potential losses implied by a choice with uncertain outcomes.

Given that investors' decisions are a form of decisions under uncertainty, namely social uncertainty (similar to ambiguous and risky choices), it is reasonable to assume that similar non-linearities observed in non-social choices under uncertainty, underlie the investor's decisions in the IG. In particular, I will consider cases in which losses are overweighted (i.e., loss aversion) and underweighted (i.e., loss insensitivity) by modifying the individual susceptibility to the potential loss implied by an act of trust as follows:

$$u_t(i_t|\mathcal{D}_{t-1}) = e_t - (1-\tau)s_t^{\Lambda} + \tau(\eta_t\,\mathbb{E}[r_t|\mathcal{D}_{t-1}]), \tag{8}$$

with $\Lambda > 1$ representing loss-aversion and $\Lambda < 1$ loss-insensitivity. Hence, the formalization of the vulnerability model considered until now (which tacitly assumed $\Lambda = 1$) can be thought of as the instantiation of a loss-neutral investor model. Similar to previous analyses, I will next investigate predictions of this extension of the vulnerability model by considering loss-averse and loss-insensitive investors reacting to attempts of trust repair after an initially uncooperative behavior by the trustee (Figure 6). Loss-averse investors are modeled with a $\Lambda = 1.2$, while loss-insensitive investors with $\Lambda = 0.4$.

As compared to the interactions with loss-neutral investors (Figure 5B), repair of trust cannot be achieved with loss-averse, distrustful investors with a low trust attitude (low τ parameter values) (Figure 6A). On the contrary, for loss-averse, trustful investors, we observe a very slow recovery of trust. Very different behavioral patterns can be observed in loss-insensitive investors (Figure 6C). First, the initial decrease in investments in trials

$t = 3, 4$ is more moderate than investment drops observed with loss-neutral and loss-averse investors. Second, trust is repaired more readily in loss-insensitive investors (for both trustful and distrustful types) who strongly increase investments from trial $t = 5$ onward, after evidence of steady, cooperative behavior by the trustee. Again, these behavioral patterns are further modulated by the investor's attitude to trust, with distrustful investors increasing their investments at a lower rate as compared to trustful investors. Overall, trust repair seems to be most successful in interactions with loss-insensitive investors.

As previously observed, one way trustees can remedy a fatal breach of cooperation with loss-averse investors is to prove a willingness to behave cooperatively, thereby changing the investors' expectations of returns. Previous work has indeed shown that trustees demonstrate extraordinary levels of reciprocity from time to time to coax investors to trust [46]. Hence, when interacting with a loss-averse investor, trustees might incur high, momentary costs to persuade the investor into higher shares. I simulated this case by having the trustee intend to share all the received amount in the three trials subsequent to the initial breach of trust (i.e., $t = 3, 4, 5$).

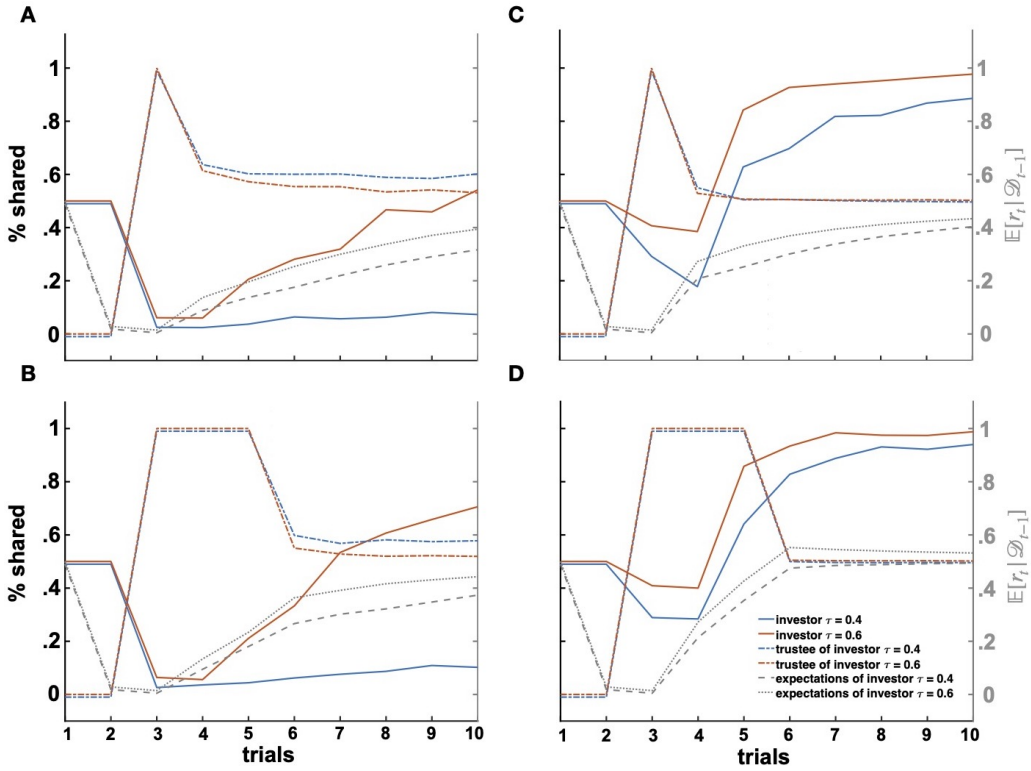

Figure 6. Effects of loss aversion on repair of trust. Repair of trust after initial uncooperativeness (trials 1–2) with loss-averse investors without (**A**) and with (**B**) a coaxing trustee and with loss-insensitive investors without (**C**) and with (**D**) a coaxing trustee.

Figure 6B shows that such extraordinary levels of reciprocity lead to a strong increase in investment behavior in loss-averse, trustful investors with higher shares than the ones observed in the previous case without such extraordinary levels of reciprocity (Figure 6A). However, they do not seem to ameliorate the development of trust with loss-averse, distrustful investors who keep sharing nothing or very low amounts until the end of the considered interaction. Importantly, despite these high levels of reciprocity, investment levels in loss-averse, trustful investors remain lower then those achieved by loss-neutral trustful investors, despite a steady increase in shares until the end of the interaction. Finally, extraordinary levels of reciprocity appear to narrow the τ-dependent differences in investments of loss-insensitive, trustful, and distrustful investors (Figure 6D) who manifest the over-time most similar investment levels among the examples until now considered. These results suggest that the interaction between trust attitudes, loss aversion, and expectations

of returns yields interesting statistics with peculiar behavioral dynamics. Taken together, the model predicts that potential losses loom longer over trusting interactions with loss-averse partners and that such effects can be partly counteracted by eliciting extremely positive expectations, especially when the partner has a strong attitude to trust and is hence more willing to accept vulnerability.

4.5. Benevolence and Competence

Until now, expectations of returns ($\mathbb{E}[r_t|\mathcal{D}_{t-1}]$) were taken as the only criterion for the trustee's trustworthiness. As previously mentioned, the reciprocal behavior of the trustee is considered as signaling a willingness to cooperate or, generally, good intentions. Hence, the trustee's reciprocity is a proxy for his/her intended benevolence, and the expectations of trustworthiness considered so far were limited to this intentionality dimension of trustworthiness. However, as discussed in the Introduction, impressions of trustworthiness arise from varying features of another person's personality. For instance, the investor's expectations of the trustee's trustworthiness might be further modulated by the trustee's competence.

In the classic IG, however, the behavioral information that can be gathered from the trustee is too poor to allow the investor to disambiguate different dimensions of trustworthiness. As the observed reciprocity is assumed to be indicative of the trustee's intentions to cooperate in the game, it follows that investors might at best make inferences on the trustee's intentions and state of mind based on the observed reciprocal behavior—such as theory of mind inferences [74]. However, little room is left for very complicated inferences, and recent work [75] has questioned that participants engage in the kind of theory of mind inferences suggested by cognitive hierarchy or "depth-of-thoughts" approaches [34,76]. Needless to say, inferences on other trust-relevant traits of the trustee, such as the trustee's competence, are close to impossible given the impoverished interaction. Indeed, the only ability that a trustee might at best exhibit would be for Machiavellian reasoning, which is likely to deter the investor.

As previously mentioned, the additive temporal decomposition of the vulnerability model is convenient for considerations of different factors that play a role in a decision to trust such as savoring and dread [61–63] associated with the duration of vulnerability or loss aversion and loss insensitivity, as considered in this paper. Hence, a simple extension of the vulnerability model in a novel, richer environment will allow us to study how the trustee's competence affects the investor's trust.

Let us first consider a modified version of the IG that disentangles the intentional dimension represented by the trustee's decision to reciprocate from a dimension of competence. Importantly, the competent behavior of the trustee should be clearly distinct from his/her intention-revealing reciprocal behavior, but at the same time, be relevant for the evolution of the social interaction. Hence, let us suppose that the realized multiplier (μ) on every trial is contingent on some behavioral performance of the trustee. For instance, the trustee might be required to solve a task, and his/her performance on this task (i.e., w, which will be assumed to be a binary variable) determines the current state of the multiplier in the IG as follows:

$$\mu_t = \begin{cases} 0 & w_t = 0 \\ 3 & w_t = 1 \end{cases}$$

where w_t is the trustee's performance in the task (winning outcome) in the current trial t. If the trustee is successful ($w_t = 1$), the game reduces to the classic IG. On the contrary, if the trustee cannot solve the task ($w_t = 0$), the trustee does not receive any amount from the investor, the investor loses the shared amount, and the current trial ends. Now, even though the trustee's performance is on a different task than the IG, the behavior on that task impacts the ongoing interaction in the IG by indirectly affecting the potentially attainable payoffs. In fact, the investor's payoffs still hinge on the trustee's reciprocity (benevolence), but the degree to which the trustee can reciprocate depends on how well he/she performs on the performance task (competence).

The extension of the vulnerability model to account for competence effects is straightforward. Now, the investor's utility is augmented by another expectation about the trustee's success, which can be represented as a probability distribution over the trustee's competence. Hence, by using the formalization of a loss-neutral investor, the investor's utility becomes:

$$u_t(i_t, \mathcal{W}_{t-1}, \mathcal{D}_{t-1}) = e_t - (1-\tau)s_t + \tau(\eta_t \, \mathbb{E}[w_t|\mathcal{W}_{t-1}]\mathbb{E}[r_t|\mathcal{D}_{t-1}]). \quad (9)$$

As before, $\mathbb{E}[r_t|\mathcal{D}_{t-1}]$ represents the investor's expectations of proportional returns (benevolence) at time t after having observed the set of exchanges of investments and returns $\mathcal{D}_{t-1} = (i_{1:t-1}, r_{1:t-1})$, while $\mathbb{E}[w_t|\mathcal{W}_{t-1}]$ represents the investor's expectations of the trustee's success (competence) at time t after having observed the set of the trustee's winnings and losses in the performance task $\mathcal{W}_{t-1} = (w_{1:t-1})$. Like $\mathbb{E}[r_t|\mathcal{D}_{t-1}]$, $\mathbb{E}[w_t|\mathcal{W}_{t-1}]$ takes the form of a beta distribution with parameters α^w and β^w:

$$Beta(w; \alpha^w, \beta^w) = p(w|\alpha^w, \beta^w) \propto w^{\alpha^w - 1}(1-w)^{\beta^w - 1}, \quad (10)$$

with w being the trustee's successes. Expectations of the trustee's success are updated via Bayesian updating after observing a new performance outcome at time t (i.e., a winning w_t or a loss $1 - w_t$). Hence, the posterior probability of $p(w|\alpha^w, \beta^w)$ for the investor becomes $Beta(w_{t+1}; \alpha^w_{t+1}, \beta^w_{t+1})$, where $\alpha^w_{t+1} = \alpha^w_t + w_t$ and $\beta^w_{t+1} = \beta^w_t + (1 - w_t)$.

Different predictions about the investor's behavior can be made based on varying expectations of the trustee's benevolence and competence. First, let' us consider how these expectations affect the investor's behavior in the case of interactions with both competent and benevolent trustees (Figure 7A). For the simulations, let us assume that the trustee has a good performance, so that the investor's expectation of the trustee's success is $\mathbb{E}[w] = 0.9$, and medium benevolence, so that the investor's initial expectation of the trustee's proportional returns is $\mathbb{E}[r_0] = 0.3$. This parameterization leads to a high number of situations indistinguishable from the classic IG. Hence, the vulnerability model predicts that investors should increase their investments. The magnitude and increase of investment remain proportional to the investors' trust attitude. Unlike the previous cases, now, the expectation of success $\mathbb{E}[w]$ introduces an additional source of uncertainty in the investor's utility that increases the contribution of the immediate losses associated with an act of trust i_t. As a consequence, the realized investments, especially by distrustful investors, are predicted to be lower than, for instance, the investments observed in Figure 5A. Hence, with a perfect trustee's performance $\mathbb{E}[w] = 1$, no additional uncertainty is added and the predicted behavior is predicted to be like the one observed in the classic IG.

Another straightforward and rather uninteresting case that I add for completeness is given by interactions with incompetent and malevolent trustees. For the simulations, the investor is taken to have an expectation of the trustee's success of $\mathbb{E}[w] = 0.6$ (that is, an overall trustee's performance close to chance level) and an initial expectation of proportional returns of $\mathbb{E}[r_0] = 0.1$. As can be seen in Figure 7B, both trustful and distrustful investors stop cooperating with these trustees to avoid losses. Again, the investor's trust attitude slightly modulates the speed and degree of how cooperation ceases.

More interesting are cases in which the benevolence and competence dimensions of trustworthiness are incongruent. Let us first consider cases of competent, but malevolent trustees (Figure 7C). For the simulations, the investor will have an expectation of the trustee's success of $\mathbb{E}[w] = 0.9$ and an initial expectation of proportional returns of $\mathbb{E}[r_0] = 0.1$. This is a scenario very similar to the one previously considered with an uncooperative trustee in the classic IG. Here, investments should decrease as correctly predicted by the vulnerability model. However, despite the malevolence of the trustee, investors should earn more from interactions with competent than incompetent partners and should hence try to take advantage of the partner's competence by sharing more. That is, investors playing with malevolent, but competent trustees should manifest higher investment levels (translating into higher payoffs) than investors playing with malevolent and incompetent trustees from whom they can receive no benefits. Importantly, such a difference should be minimal

(given the malevolence of both trustee types) and again proportional to the investor's trust attitude. This is indeed predicted by the vulnerability model that generated slightly higher investment ratios for investors interacting with malevolent, but competent trustees than malevolent and incompetent trustees (Figures 7B,C and 8A). Such a preference is further reflected by higher payoffs for investors interacting with malevolent, but competent trustees than malevolent and incompetent trustees (Figure 8B). This of course holds as long as the trustee shows some reciprocal behavior. In cases of extreme malevolence, where the investor's expectations of proportional returns are $\mathbb{E}[r] = 0$, the incentivizing contribution of competence vanishes and no investments are made.

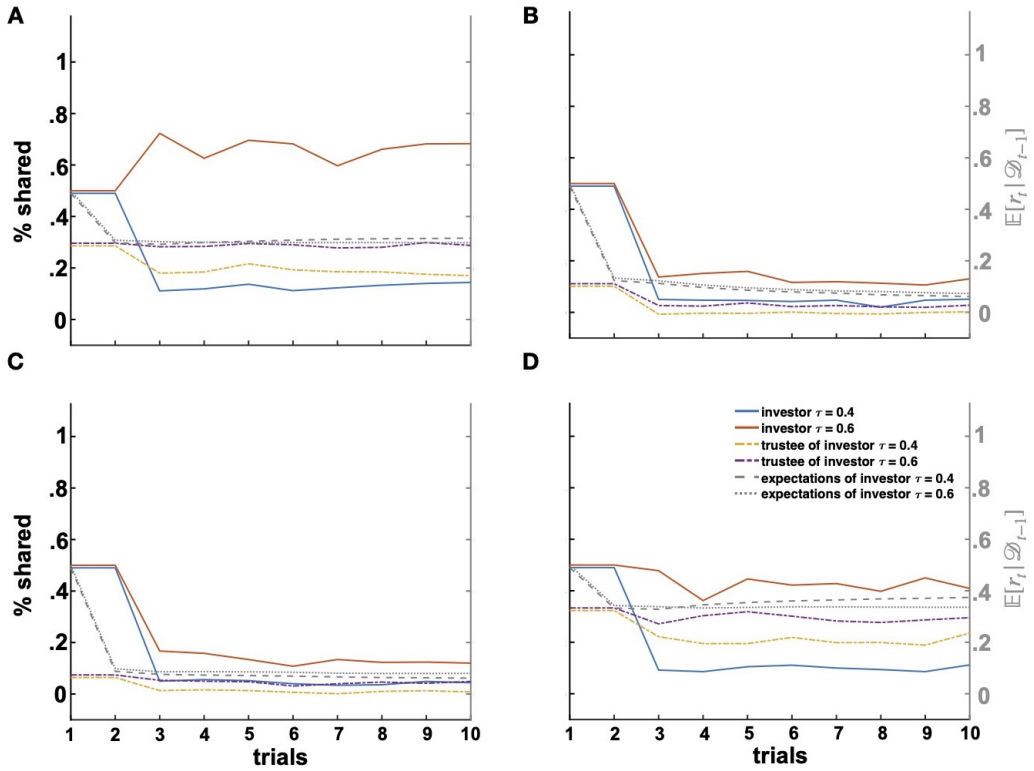

Figure 7. Competence and benevolence. Investments over time (trials) shared by investors interacting with different trustee types, that is: benevolent and competent (**A**); malevolent and incompetent (**B**); malevolent and competent (**C**); and benevolent and incompetent (**D**).

Let us now consider the case of an interaction with an incompetent, but benevolent trustee. For this case, the investor will have an expectation of the trustee's success of $\mathbb{E}[w] = 0.6$ and an initial expectation of proportional returns of $\mathbb{E}[r_0] = 0.3$. Of course, the incompetence of the trustee reduces the expected benevolence, and as a result, the weights on the different model components are slanted in favor of the immediate loss implied by the act of trust. However, the trustee's benevolence still guarantees a potential return as the quantity $\eta_t \; \mathbb{E}[w_t|\mathcal{W}_{t-1}]\mathbb{E}[r_t|\mathcal{D}_{t-1}] > 0$ at a given time t. Hence, while expectations of a poor performance by benevolent, but incompetent trustees should lead to lower investment levels as compared to interactions with benevolent, but competent trustees, higher expectations of returns by benevolent and incompetent trustees should induce higher investments than the ones observed in interactions with malevolent trustees (irrespective of their competence). Figure 8C showcases these behavioral patterns for both trustful and distrustful investors. In particular, Figure 8C depicts the differences of percentage shares between the default initial strategy of $i_t = 0.5$ at $t = 1, 2$ and the first expectation-based investment at $t = 3$ for the interactions with the four possible trustee types. As predicted, investments with benevolent trustees (blue) are consistently higher than investments with malevolent trustees (red). For interactions with benevolent trustees, investments are higher when trustees are also competent (dark blue) than when they are incompetent (light blue).

Moreover, investments with incompetent, but benevolent trustees (light blue) are higher than investments with both competent and incompetent, but malevolent trustees (dark and light red). In addition, investment differences are dependent on the investor's trust attitude with overall higher investments by trustful investors. Finally, the model also predicts that in all but one of the above scenarios, trust should degenerate. The only exception in which investments increase is given by the optimal case of interactions with trustees who are both benevolent and competent (Figure 8C).

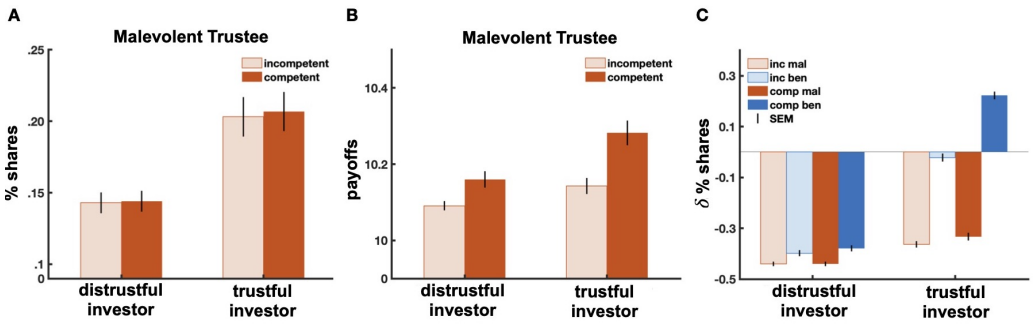

Figure 8. Investment and payoff comparison across trustee types. Comparison of percentage of investments (**A**) and obtained payoffs (**B**) of distrustful (left) and trustful (right) investors for interactions with malevolent trustees who were additionally incompetent (light red) or competent (red). Differences of percentage shares (**C**) between the default initial strategy of $i_t = 0.5$ at $t = 1, 2$ and the first expectation-based investment i_t at $t = 3$ across all trustee types for distrustful (left) and trustful (right) investors. Inc: incompetent; Comp: competent; Mal: malevolent; Ben: benevolent.

5. Discussion

In this paper, inspired by philosophy, social psychology, behavioral economics, and social neuroscience, I provided a formalization of psychological and philosophical verbal models of trust as the *attitude* to rely on expectations of a social partner despite the vulnerability to the trusted partner that the act of trust implies. I have shown behavioral predictions of the model as a function of the investor's loss aversion and expectations. I have further shown how according to the model, an individual reacts to different interacting dynamics such as attempts to breach and repair trust from the trustee. Finally, I have demonstrated how the model can be extended to incorporate other sources of uncertainty relevant to a decision to trust in a novel environment in which expectations of the trustee's competence need to be incorporated into the utility function before a trust decision is made.

The considered model makes predictions consistent with previous behavioral results and provides a mechanistic understanding of previous observations in empirical work. For instance, the model predicts a gradual decrease of investment after breaches of trust, which is proportional to the change rate of the investor's expectations of the trustee's reciprocity. Such a gradual decrease has been interpreted as the investor's willingness to, on the one hand, keep open the possibility of future cooperation after an adequate trust repair and, on the other, signal disapproval or irritation [24,32,50].

Further, model simulations have shown how implausibly high expectations impair behavioral adaptation to learning. In particular, trustful investors with highly positive expectations were more likely to be exploited by an uncooperative trustee despite the fact that they did learn about the trustee's exploitation. These behavioral patterns have been observed in different clinical populations such as in individuals with generalized anxiety disorder and individuals with attention deficit hyperactivity disorder (ADHD) [40,77,78]. Importantly, a computational modeling of behaviors in these populations has previously suggested a "learning impairment" with lower rates of learning [40]. However, this previous model did not incorporate individuals' expectations of a partner's behavior, and it is still an open question whether those participants had an actually worse learning performance or whether their behavioral responses were biased by their unrealistic expectations despite accurate learning. Fitting a model that does not account for expectations to the

behavioral patterns generated by the vulnerability model in Figure 5C might most likely suggest an impairment in learning about the trustee's behavior. However, as the changes in the investor's expectations show, an individual with unrealistically positive expectations might still be able to correctly learn about a partner's behavior without him/her being actually reflected in the observed behavior.

Similarly, Figure 5D shows how highly negative expectations disrupt the ability to establish or recover a trusting interaction. Previous work has suggested that highly negative expectations in some clinical populations might impede healthy social interactions [79]. For instance, individuals with borderline personality disorders playing as trustee do not seem to be able to appropriately increase their returns as a response to investment increases, leading to a catastrophic cascade of the social interaction, which is consistent with evidence on impaired interpersonal functioning in borderline personality disorder [34,80,81]. Similarly, individuals with psychosis playing as investor manifest lower baseline trust (especially at the beginning of the interaction) and are less likely to increase investment after trust-honoring reciprocity than healthy individuals, suggesting a reduced sensitivity to social feedback that has been interpreted as reflecting their impairment in sensitivity to interpersonal cues [82,83]. Based on these findings, it has been suggested that individuals with borderline personality disorder and psychosis do not accurately learn or respond inadequately to cooperation signals by their social partners [46,84–88]. However, the impact of negative expectations in these populations has not been investigated yet, and in line with a negativity bias in social cue evaluations in individuals with these clinical disorders [89,90], we have seen in this paper that highly negative expectations might indeed explain part of the behavioral patterns observed in previous research.

Importantly, even though the following model was inspired by psychological and philosophical conceptualizations of trust, it is consistent with other formalizations in other fields such as finance. For instance, [91] formalized distrust as a multiplier of a negative disutility that reflects investment uncertainty and reduces expectation of positive investment returns. This formalization is formally equivalent to the one in the present work, but bears some important differences in interpretation. In particular, while [91] formalized trust only as an anxiety-reduction mechanism associated with the uncertainty over investment-related costs, the following work thinks of trust as both a multiplicative boost to the net expected return and a complementary multiplicative reduction of the possible costs incurred by investors when trusting. The current work's formalization is more consistent with empirical findings and psychological theories supporting the idea that both one's sensitivity to potential losses and positive expectations of potential gains should be taken into account in a model of trust [92].

Further, in the current work, I have shown cases of what we could call "the self-fulfilling prophecy of a lack of trust". In particular, as forming accurate expectations about the trustee's behavior is possible only if the investor gives him/her a chance to reciprocate, a lack of trust implies poorer behavioral information about the actual reciprocity of the trustee and, hence, less accurate expectations of the partner. This occurs both when the investor decides to share nothing (with no information whatsoever received from the partner) and when the investor decides to share small amounts (allowing less variability in the possible amounts to share back and, hence, contributing to noisier behavioral information about the partner). This creates a spiral of self-fulfilling prophecies for which the distrustful investor who does not trust much in the first place receives negative reciprocating feedback from the trustee, thereby being confirmed in his/her negative expectations of the partner. In turn, such a confirmation of his/her negative expectations makes distrustful investors behave even more distrustfully in subsequent interactions. As stated above, such dynamics are very close to a variety of real-life interactions, suggesting that the paradigmatic examples considered here might be generalizable. However, not all interactions (in real life and experimental settings) have to abide by this self-fulfilling prophecy of a lack of trust. For instance, when interactions are not asymmetric and participants receive feedback about

their partner's behavior at every time step, such as in many economic games, such as the Prisoner's Dilemma, this trap of self-fulfilling prophecy can be avoided.

In addition, in the current work, inferences on two different psychological traits of another person and their impact on trust decisions during social interactions have been for the first time explored. Importantly, in the current implementation, the investor was supposed to receive information directly related to both character traits of his/her partner, that is his/her competence and benevolence. This greatly simplifies the inference problem. For instance, knowing how well the trustee does in the performance task likely protects against the breakdown of trust, since it is easier to forgive a mistake than malevolence. However, in many real-life situations, there exists partial observability over psychological features of an interacting partner. For example, when a particular behavior cannot so easily be traced back to one of these psychological traits, some further strategic thinking and planning are necessary to figure out the "true nature" of the interaction and avoid exploitation. Such a partial observability is likely to trigger interesting dynamics worth exploring in future work, for example in games in which the investor does not know the realized multiplier [21,93,94].

Moreover, the vulnerability model is also able to describe economic predictions of rational behavior. In the one-shot IG, a rational, economic agent playing as investor would assume that the trustee does not have any incentives to reciprocate and should hence not trust. According to the vulnerability model, this is equivalent to saying that the positive contribution of the expectations of returns $\mathbb{E}[r_t|\mathcal{D}_{t-1}]$ in Equation (5) is marginal or non-existent. This implies that the utility function is overly affected by the disutility of sharing s_t, making the model predict that the investor should not share anything (which would yield the highest utility for the investor). In this case, the model predictions are consistent with a rational agent. The same applies in repeated games. The model currently only mimics open-end interactions (e.g., when participants do not how many rounds they are going to play together), but does not account for the temporal effects of a social exchange. Such temporal effects have been proposed to be accounted for when individuals know how long the interaction with the partner is going to last and are supposed to explain end-game effects [52]. For instance, the expectations of returns $\mathbb{E}[r_t|\mathcal{D}_{t-1}]$ might be contingent on a utility function of the trustee that takes into account a future investment. In the last round, when this future investment is not possible, the utility of the trustee will be mainly affected by the loss implied by reciprocation, which predicts that the trustee should not share anything. Now, if the expectations of returns $\mathbb{E}[r_t|\mathcal{D}_{t-1}]$ depend on the trustee's utility function, when the trustee's utility predicts a defection, the investor should think that reciprocity is less likely (i.e., should have more negative expectations of the trustee's returns). Hence, the vulnerability model predicts that the investor decreases his/her shares or stops sharing altogether. As the current work does not propose a model of reciprocity (with a utility function for the trustee), future research is needed to address these scenarios.

Another interesting avenue for future research is to investigate how different trust attitudes (for example, the prevalence of different τ parameters) might emerge and evolve in populations with different network structures. For instance, previous work has investigated how trust and trustworthiness evolve in different networks by using a binary version of the IG where players have a binary decision to make (trust vs. distrust for the investor and reciprocate vs. betray for the trustee) [38]. The authors found that trust and trustworthiness evolve only under specific network structures and only for a subset of investment and back-transfer combinations, suggesting how fragile the establishment of prosocial behaviors such as trust and reciprocity is. These results echo empirical evidence showing that a single norm-violating agent reduces compliance with a social norm by a way larger extent than the extent to which a single norm-compliant agent is able to increase it [95]. However, the work by Kumar et al. [38] focused only on the prevalence of different behavioral strategies and not on the evolution and diffusion of different social preferences and attitudes. One interesting research question concerns investigations on how robust individual, social

preferences, and attitudes are across different populations and under which conditions they thrive and die out.

Despite these advances, the considered model makes a set of simplifying assumptions. First, I assumed that participants are computing their current utility $u_t(i_t, r_t)$ only with respect to the current endowment e_t available to them at time t. However, in repeated interactions, it is conceivable that participants actually keep a cumulative sum of the total gains they have been gathering that modulates the contribution of the negative utility associated with the investment i_t. This might have the effect of making participants increasingly less loss averse or more loss averse as time passes. Participants might do this especially when they know the horizon length of the game and such a length is short enough or when the total earned sum is shown to them as a feedback at the end of every trial (e.g., in experiments in which participants have information about their earning balance). In multi-round games with the same trustee, it is of course difficult to disentangle these effects from the learning about the trustee's cooperativeness, as a good balance in such IG versions highly correlates with cooperative trustees that have shared fairly in previous trials (or vice versa). Of course, such effects could be explained by other factors as well, such as acquired familiarity with the game or generalization of previous positive and negative interactions. In particular, it has been shown that experienced and inexperienced individuals implement different investment strategies and even the same individual changes his/her behavior as a function of experience with the game [52,54,58,96].

Importantly, these effects can also hinge on prospective calculations about the amount of money individuals are likely to earn. For instance, not only participants have been seen to decrease their investments in the last trials of a finite IG with a known number of trials [34,52], they have also been seen to share significantly less when they are told that their monetary compensation will only consist of the amount of money they have earned in a small subset of trials (e.g., one or slightly more) randomly drawn at the end of the experiment [35]. If cumulative earnings reduce the contribution of the negative utility associated with investment i_t, participants might indeed be more likely to share higher proportions of monetary amounts when they know that the sum of what they earn in each single trial of the experiment is what counts for their final payoffs. As a side note, these two different compensation schemes might frame the experiment either as a reward-poor environment or as a reward-rich environment, thereby creating different preferences that shift participants' investment behaviors [97].

Further, in the presented simulations, I showed how the vulnerability model behaves based on different investor's parameters and on learning about the average behavior of different trustees who implemented fixed behavioral strategies. However, more complicated frameworks might be used to capture more sophisticated cognitive processes during such trusting interactions. For instance, in the current work, the investor's expectations is based on a non-sophisticated model according to which the investor only maintains a running average of the trustee's behavior. This is likely the most common strategy employed by investors in the impoverished classic IG. For instance, [34] reported that at least 50% of their participants playing as investor behaved this way. However, investors might be sensitive to other statistics of the partner's behavior or employ different inference processes to support their decisions. For example, as the game is repeated more and more often, the beta distribution, which represents the investor's expectations of the partner's behavior, becomes more and more concentrated on the average return. This has two consequences: First, highly peaky distributions (e.g., after having seen much evidence) tend to be less sensitive to observations far from the mean. This predicts that investors who know a great deal about their trustee should more strongly dismiss evidence incongruent with their expectations. Such a prediction is consistent with previous work [40,41,98], but in the current implementation, the beta distribution might require an unreasonable amount of observations before a behavioral change can be promoted. Second, investors might also be sensitive not just to the partner's average behavior, but also to his/her behavioral variability. This behavioral variability of the partner is likely associated with a form of

uncertainty of both a psychological (epistemic uncertainty) and mathematical (variance) nature. Previous work has provided some evidence that behavioral variability has a negative impact on people's trustworthiness impressions of others [98], indicating that social agents are sensitive to other people's behavioral consistency when learning about their personality.

Further, some investors might engage in prospective inferences about their partner's intentions and future behavior, by incorporating beliefs they think they might trigger in their partner by behaving in a particular fashion. These are known as second- (and higher-) order beliefs of and about the interacting partner. For instance, in the current implementation, when trustees exhibit extraordinary levels of reciprocity in one of the considered paradigmatic cases, investors increase their investments without strategic considerations stopping them. However, real participants might think that the trustee wants them to think she/he has started being cooperative to make him/her share more money that the trustee will ultimately keep. Hence, a strategic, far-sighted investor might avoid falling into such a trap by refusing to share altogether. The current myopic model, however, is not able to capture such strategic dynamics. Nonetheless, the utility function of the vulnerability model can be incorporated into more general frameworks that allow such strategic inferences (e.g., into Bayesian inference frameworks such as the partially observable Markov decision processes) [67].

Finally, the desideratum for every model is its generalizability, that is its ability to extend to other trusting interactions and behaviors. Intuitively, the model is in principle able to capture all interactions that presuppose a weighting trade-off between potential losses and attainable gains made possible by an act of trust. However, different situations are likely to require the modeling of other aspects of a social interaction or other social inferences on a partner's quality (like in the exemplary case of competence investigated above), which all require extensions and modifications. However, the current model is not only flexible enough to incorporate further expectations and inferences on a partner's behavior, it also allows capturing a wide range of trust attitudes with a very simple and easily understandable mechanism. Further, beyond explanations of behavior in healthy subjects, the vulnerability model has also proven to generate behavioral patterns akin to behavioral findings previously observed in different clinical populations, suggesting that it has the potential to be employed for predictions relevant for clinical investigations. Hence, the proposed formalization lays the groundwork for new theoretical and empirical investigations for refinements and testing of further behavioral predictions in the multifaceted dynamics of trusting interactions.

Funding: This work was supported by the Max Planck Society.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Baker, R.; Honeyford, K.; Levene, L.S.; Mainous, A.G.; Jones, D.R.; Bankart, M.J.; Stokes, T. Population characteristics, mechanisms of primary care and premature mortality in England: A cross-sectional study. *BMJ Open* **2016**, *6*, e009981. [CrossRef] [PubMed]
2. Gray, D.J.P.; Sidaway-Lee, K.; White, E.; Thorne, A.; Evans, P.H. Continuity of care with doctors—A matter of life and death? A systematic review of continuity of care and mortality. *BMJ Open* **2018**, *8*, e021161. [CrossRef]
3. Horváth, R. Does trust promote growth? *J. Comp. Econ.* **2013**, *41*, 777–788. [CrossRef]
4. De Jong, B.A.; Dirks, K.T. Beyond shared perceptions of trust and monitoring in teams: Implications of asymmetry and dissensus. *J. Appl. Psychol.* **2012**, *97*, 391. [CrossRef] [PubMed]
5. Etang, A.; Fielding, D.; Knowles, S. Does trust extend beyond the village? Experimental trust and social distance in Cameroon. *Exp. Econ.* **2011**, *14*, 15–35. [CrossRef]
6. Krueger, F.; Meyer-Lindenberg, A. Toward a model of interpersonal trust drawn from neuroscience, psychology, and economics. *Trends Neurosci.* **2019**, *42*, 92–101. [CrossRef] [PubMed]

7. Goldberg, S.C. Trust and Reliance 1. In *The Routledge Handbook of Trust and Philosophy*; Routledge: London, UK, 2020; pp. 97–108.
8. Mayer, R.C.; Davis, J.H.; Schoorman, F.D. An integrative model of organizational trust. *Acad. Manag. Rev.* **1995**, *20*, 709–734. [CrossRef]
9. Rousseau, D.M.; Sitkin, S.B.; Burt, R.S.; Camerer, C. Not so different after all: A cross-discipline view of trust. *Acad. Manag. Rev.* **1998**, *23*, 393–404. [CrossRef]
10. Strickland, L.H. Surveillance and trust. *J. Personal.* **1958**, *26*, 200–215. [CrossRef]
11. Malhotra, D.; Murnighan, J.K. The effects of contracts on interpersonal trust. *Adm. Sci. Q.* **2002**, *47*, 534–559. [CrossRef]
12. Conrads, J.; Irlenbusch, B. Strategic ignorance in ultimatum bargaining. *J. Econ. Behav. Organ.* **2013**, *92*, 104–115. [CrossRef]
13. Bellucci, G.; Dreher, J.C. Trust and Learning. In *The Neurobiology of Trust*; Cambridge University Press: Cambridge, UK, 2021; p. 185.
14. Yaniv, I.; Kleinberger, E. Advice taking in decision making: Egocentric discounting and reputation formation. *Organ. Behav. Hum. Decis. Process.* **2000**, *83*, 260–281. [CrossRef] [PubMed]
15. Hertz, U.; Palminteri, S.; Brunetti, S.; Olesen, C.; Frith, C.D.; Bahrami, B. Neural computations underpinning the strategic management of influence in advice giving. *Nat. Commun.* **2017**, *8*, 2191. [CrossRef] [PubMed]
16. Mahmoodi, A.; Bahrami, B.; Mehring, C. Reciprocity of social influence. *Nat. Commun.* **2018**, *9*, 2474. [CrossRef]
17. Biele, G.; Rieskamp, J.; Krugel, L.K.; Heekeren, H.R. The neural basis of following advice. *PLoS Biol.* **2011**, *9*, e1001089. [CrossRef] [PubMed]
18. Biele, G.; Rieskamp, J.; Gonzalez, R. Computational models for the combination of advice and individual learning. *Cogn. Sci.* **2009**, *33*, 206–242. [CrossRef]
19. Berg, J.; Dickhaut, J.; McCabe, K. Trust, reciprocity, and social history. *Games Econ. Behav.* **1995**, *10*, 122–142. [CrossRef]
20. Krueger, F.; Grafman, J.; McCabe, K. Neural correlates of economic game playing. *Philos. Trans. R. Soc. B Biol. Sci.* **2008**, *363*, 3859–3874. [CrossRef]
21. van Baar, J.M.; Chang, L.J.; Sanfey, A.G. The computational and neural substrates of moral strategies in social decision-making. *Nat. Commun.* **2019**, *10*, 1483. [CrossRef]
22. Chaudhuri, A.; Gangadharan, L. An experimental analysis of trust and trustworthiness. *South. Econ. J.* **2007**, *73*, 959–985. [CrossRef]
23. Csukás, C.; Fracalanza, P.; Kovács, T.; Willinger, M. The determinants of trusting and reciprocal behaviour: Evidence from an intercultural experiment. *J. Econ. Dev.* **2008**, *33*, 71–95. [CrossRef]
24. Krueger, F.; McCabe, K.; Moll, J.; Kriegeskorte, N.; Zahn, R.; Strenziok, M.; Heinecke, A.; Grafman, J. Neural correlates of trust. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 20084–20089. [CrossRef]
25. Chaudhuri, A.; Sopher, B.; Strand, P. Cooperation in social dilemmas, trust and reciprocity. *J. Econ. Psychol.* **2002**, *23*, 231–249. [CrossRef]
26. McCabe, K.A.; Rigdon, M.L.; Smith, V.L. Positive reciprocity and intentions in trust games. *J. Econ. Behav. Organ.* **2003**, *52*, 267–275. [CrossRef]
27. Fetchenhauer, D.; Dunning, D. Do people trust too much or too little? *J. Econ. Psychol.* **2009**, *30*, 263–276. [CrossRef]
28. Burnham, T.; McCabe, K.; Smith, V.L. Friend-or-foe intentionality priming in an extensive form trust game. *J. Econ. Behav. Organ.* **2000**, *43*, 57–73. [CrossRef]
29. Dunning, D.; Anderson, J.E.; Schlösser, T.; Ehlebracht, D.; Fetchenhauer, D. Trust at zero acquaintance: more a matter of respect than expectation of reward. *J. Personal. Soc. Psychol.* **2014**, *107*, 122. [CrossRef]
30. Baumgartner, T.; Fischbacher, U.; Feierabend, A.; Lutz, K.; Fehr, E. The neural circuitry of a broken promise. *Neuron* **2009**, *64*, 756–770. [CrossRef]
31. Phan, K.L.; Sripada, C.S.; Angstadt, M.; McCabe, K. Reputation for reciprocity engages the brain reward center. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 13099–13104. [CrossRef]
32. King-Casas, B.; Tomlin, D.; Anen, C.; Camerer, C.F.; Quartz, S.R.; Montague, P.R. Getting to know you: reputation and trust in a two-person economic exchange. *Science* **2005**, *308*, 78–83. [CrossRef]
33. Fairley, K.; Sanfey, A.; Vyrastekova, J.; Weitzel, U. Trust and risk revisited. *J. Econ. Psychol.* **2016**, *57*, 74–85. [CrossRef]
34. Xiang, T.; Ray, D.; Lohrenz, T.; Dayan, P.; Montague, P.R. Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Comput. Biol.* **2012**, *8*, e1002841. [CrossRef]
35. Johnson, N.D.; Mislin, A.A. Trust games: A meta-analysis. *J. Econ. Psychol.* **2011**, *32*, 865–889. [CrossRef]
36. Bellucci, G.; Chernyak, S.V.; Goodyear, K.; Eickhoff, S.B.; Krueger, F. Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. *Hum. Brain Mapp.* **2017**, *38*, 1233–1248. [CrossRef] [PubMed]
37. Capraro, V.; Perc, M. Mathematical foundations of moral preferences. *J. R. Soc. Interface* **2021**, *18*, 20200880. [CrossRef] [PubMed]
38. Kumar, A.; Capraro, V.; Perc, M. The evolution of trust and trustworthiness. *J. R. Soc. Interface* **2020**, *17*, 20200491. [CrossRef]
39. FeldmanHall, O.; Nassar, M.R. The computational challenge of social learning. *Trends Cogn. Sci.* **2021**, *25*, 1045–1057. [CrossRef]
40. Lamba, A.; Frank, M.J.; FeldmanHall, O. Anxiety impedes adaptive social learning under uncertainty. *Psychol. Sci.* **2020**, *31*, 592–603. [CrossRef]
41. Chang, L.J.; Doll, B.B.; van't Wout, M.; Frank, M.J.; Sanfey, A.G. Seeing is believing: Trustworthiness as a dynamic belief. *Cogn. Psychol.* **2010**, *61*, 87–105. [CrossRef]

42. Luo, Y.; Hétu, S.; Lohrenz, T.; Hula, A.; Dayan, P.; Ramey, S.L.; Sonnier-Netto, L.; Lisinski, J.; LaConte, S.; Nolte, T.; et al. Early childhood investment impacts social decision-making four decades later. *Nat. Commun.* **2018**, *9*, 4705. [CrossRef]
43. Hula, A.; Moutoussis, M.; Will, G.J.; Kokorikou, D.; Reiter, A.M.; Ziegler, G.; Bullmore, E.; Jones, P.B.; Goodyer, I.; Fonagy, P.; et al. Multi-Round Trust Game quantifies inter-individual differences in Social Exchange from Adolescence to Adulthood. *Comput. Psychiatry* **2021**, *5*, 102–118. [CrossRef]
44. Fehr, E.; Schmidt, K.M. A theory of fairness, competition, and cooperation. *Q. J. Econ.* **1999**, *114*, 817–868. [CrossRef]
45. Terenzi, D.; Liu, L.; Bellucci, G.; Park, S.Q. Determinants and modulators of human social decisions. *Neurosci. Biobehav. Rev.* **2021**, *128*, 383–393. [CrossRef]
46. King-Casas, B.; Sharp, C.; Lomax-Bream, L.; Lohrenz, T.; Fonagy, P.; Montague, P.R. The rupture and repair of cooperation in borderline personality disorder. *Science* **2008**, *321*, 806–810. [CrossRef]
47. Thielmann, I.; Hilbig, B.E. The traits one can trust: Dissecting reciprocity and kindness as determinants of trustworthy behavior. *Personal. Soc. Psychol. Bull.* **2015**, *41*, 1523–1536. [CrossRef]
48. Delgado, M.R.; Frank, R.H.; Phelps, E.A. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat. Neurosci.* **2005**, *8*, 1611–1618. [CrossRef] [PubMed]
49. Everett, J.A.; Pizarro, D.A.; Crockett, M.J. Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* **2016**, *145*, 772. [CrossRef] [PubMed]
50. Hula, A.; Vilares, I.; Lohrenz, T.; Dayan, P.; Montague, P.R. A model of risk and mental state shifts during social interaction. *PLoS Comput. Biol.* **2018**, *14*, e1005935. [CrossRef]
51. Johansson-Stenman, O.; Mahmud, M.; Martinsson, P. Does stake size matter in trust games? *Econ. Lett.* **2005**, *88*, 365–369. [CrossRef]
52. Engle-Warnick, J.; Slonim, R.L. The evolution of strategies in a repeated trust game. *J. Econ. Behav. Organ.* **2004**, *55*, 553–573. [CrossRef]
53. Tisserand, J.C.; Cochard, F.; Le Gallo, J. *Altruistic or Strategic Considerations: A Meta-Analysis on the Ultimatum and Dictator Games*; CRESE-Université de Franche-Comté: Besançon, France, 2015.
54. Henrich, J.; Boyd, R.; Bowles, S.; Camerer, C.; Fehr, E.; Gintis, H.; McElreath, R.; Alvard, M.; Barr, A.; Ensminger, J.; et al. Economic man in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behav. Brain Sci.* **2005**, *28*, 795–815. [CrossRef] [PubMed]
55. Loewenstein, G.F.; Thompson, L.; Bazerman, M.H. Social utility and decision making in interpersonal contexts. *J. Personal. Soc. Psychol.* **1989**, *57*, 426. [CrossRef]
56. Fehr, E.; Schmidt, K.M. The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handb. Econ. Giv. Altruism Reciprocity* **2006**, *1*, 615–691.
57. Cochard, F.; Le Gallo, J.; Georgantzis, N.; Tisserand, J.C. Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game. *J. Behav. Exp. Econ.* **2021**, *90*, 101613. [CrossRef]
58. Binmore, K.; Shaked, A. Experimental economics: Where next? *J. Econ. Behav. Organ.* **2010**, *73*, 87–100. [CrossRef]
59. Tversky, A.; Kahneman, D. Loss aversion in riskless choice: A reference-dependent model. *Q. J. Econ.* **1991**, *106*, 1039–1061. [CrossRef]
60. Kahneman, D.; Tversky, A. Prospect theory: An analysis of decisions under risk. *Econometrica* **1979**, *47*, 263–291. [CrossRef]
61. Iigaya, K.; Story, G.W.; Kurth-Nelson, Z.; Dolan, R.J.; Dayan, P. The modulation of savouring by prediction error and its effects on choice. *eLife* **2016**, *5*, e13747. [CrossRef]
62. Iigaya, K.; Hauser, T.U.; Kurth-Nelson, Z.; O'Doherty, J.P.; Dayan, P.; Dolan, R.J. The value of what's to come: Neural mechanisms coupling prediction error and the utility of anticipation. *Sci. Adv.* **2020**, *6*, eaba3828. [CrossRef]
63. Loewenstein, G. Anticipation and the valuation of delayed consumption. *Econ. J.* **1987**, *97*, 666–684. [CrossRef]
64. Lahno, B. Trust and Emotion. In *The Routledge Handbook of Trust and Philosophy*; Routledge: London, UK, 2020; pp. 147–159.
65. Ferguson, A.J.; Peterson, R.S. Sinking slowly: Diversity in propensity to trust predicts downward trust spirals in small groups. *J. Appl. Psychol.* **2015**, *100*, 1012. [CrossRef] [PubMed]
66. Rotter, J.B. Generalized expectancies for interpersonal trust. *Am. Psychol.* **1971**, *26*, 443. [CrossRef]
67. Khalvati, K.; Park, S.A.; Mirbagheri, S.; Philippe, R.; Sestito, M.; Dreher, J.C.; Rao, R.P. Modeling other minds: Bayesian inference explains human choices in group decision-making. *Sci. Adv.* **2019**, *5*, eaax8783. [CrossRef] [PubMed]
68. Choi, J.K.; Ahn, T. Strategic reward and altruistic punishment support cooperation in a public goods game experiment. *J. Econ. Psychol.* **2013**, *35*, 17–30. [CrossRef]
69. Fehr, E.; Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **2000**, *90*, 980–994. [CrossRef]
70. Bochet, O.; Page, T.; Putterman, L. Communication and punishment in voluntary contribution experiments. *J. Econ. Behav. Organ.* **2006**, *60*, 11–26. [CrossRef]
71. Benartzi, S.; Thaler, R.H. Myopic loss aversion and the equity premium puzzle. *Q. J. Econ.* **1995**, *110*, 73–92. [CrossRef]
72. Camerer, C.; Babcock, L.; Loewenstein, G.; Thaler, R. Labor supply of New York City cabdrivers: One day at a time. *Q. J. Econ.* **1997**, *112*, 407–441. [CrossRef]
73. Thaler, R.H.; Tversky, A.; Kahneman, D.; Schwartz, A. The effect of myopia and loss aversion on risk taking: An experimental test. *Q. J. Econ.* **1997**, *112*, 647–661. [CrossRef]
74. Frith, C.D.; Frith, U. Interacting minds—A biological basis. *Science* **1999**, *286*, 1692–1695. [CrossRef]

75. Qi, W.; Vul, E. Adaptive behavior in variable games requires theory of mind. *PsyArXiv* **2020**. Available online: http://www.evullab.org/publications.php?key=qi2020adaptive&evullab.v0.bib=evullab.v0.bib (aceesed on 25 March 2022).
76. Camerer, C.F.; Ho, T.H.; Chong, J.K. A cognitive hierarchy model of games. *Q. J. Econ.* **2004**, *119*, 861–898. [CrossRef]
77. Lis, S.; Baer, N.; Franzen, N.; Hagenhoff, M.; Gerlach, M.; Koppe, G.; Sammer, G.; Gallhofer, B.; Kirsch, P. Social interaction behavior in ADHD in adults in a virtual trust game. *J. Atten. Disord.* **2016**, *20*, 335–345. [CrossRef] [PubMed]
78. Sripada, C.S.; Angstadt, M.; Banks, S.; Nathan, P.J.; Liberzon, I.; Phan, K.L. Functional neuroimaging of mentalizing during the trust game in social anxiety disorder. *Neuroreport* **2009**, *20*, 984. [CrossRef] [PubMed]
79. Kube, T.; Rief, W.; Gollwitzer, M.; Gärtner, T.; Glombiewski, J.A. Why dysfunctional expectations in depression persist–Results from two experimental studies investigating cognitive immunization. *Psychol. Med.* **2019**, *49*, 1532–1544. [CrossRef]
80. Liebke, L.; Bungert, M.; Thome, J.; Hauschild, S.; Gescher, D.M.; Schmahl, C.; Bohus, M.; Lis, S. Loneliness, social networks, and social functioning in borderline personality disorder. *Personal. Disord. Theory, Res. Treat.* **2017**, *8*, 349. [CrossRef] [PubMed]
81. Gunderson, J.G. Disturbed relationships as a phenotype for borderline personality disorder. *Am. J. Psychiatry* **2007**, *164*, 1637–1640. [CrossRef] [PubMed]
82. Fett, A.K.J.; Shergill, S.S.; Joyce, D.W.; Riedl, A.; Strobel, M.; Gromann, P.M.; Krabbendam, L. To trust or not to trust: The dynamics of social interaction in psychosis. *Brain* **2012**, *135*, 976–984.
83. Gromann, P.M.; Heslenfeld, D.J.; Fett, A.K.; Joyce, D.W.; Shergill, S.S.; Krabbendam, L. Trust versus paranoia: abnormal response to social reward in psychotic illness. *Brain* **2013**, *136*, 1968–1975.
84. Michael, J.; Chennells, M.; Nolte, T.; Ooi, J.; Griem, J.; Network, M.D.R.; Christensen, W.; Feigenbaum, J.; King-Casas, B.; Fonagy, P.; et al. Probing commitment in individuals with borderline personality disorder. *J. Psychiatr. Res.* **2021**, *137*, 335–341.
85. Gratz, K.L.; Dixon-Gordon, K.L.; Breetz, A.; Tull, M. A laboratory-based examination of responses to social rejection in borderline personality disorder: The mediating role of emotion dysregulation. *J. Personal. Disord.* **2013**, *27*, 157–171.
86. Cavicchioli, M.; Maffei, C. Rejection sensitivity in borderline personality disorder and the cognitive–affective personality system: A meta-analytic review. *Personal. Disord. Theory, Res. Treat.* **2020**, *11*, 1. [CrossRef] [PubMed]
87. Johannsen, W.J. Responsiveness of chronic schizophrenics and normals to social and nonsocial feedback. *J. Abnorm. Soc. Psychol.* **1961**, *62*, 106. [CrossRef] [PubMed]
88. Jansen, I.; Fett, A.K. Trust and Psychotic Disorders-Unraveling the Dynamics of Paranoia and Disturbed Social Interaction. In *The Neurobiology of Trust*; Cambridge University Press: Cambridge, UK, 2021; pp. 389–431.
89. Kleindienst, N.; Hauschild, S.; Liebke, L.; Thome, J.; Bertsch, K.; Hensel, S.; Lis, S. A negative bias in decoding positive social cues characterizes emotion processing in patients with symptom-remitted borderline personality disorder. *Borderline Personal. Disord. Emot. Dysregul.* **2019**, *6*, 17. [CrossRef] [PubMed]
90. Campellone, T.R.; Fisher, A.J.; Kring, A.M. Using social outcomes to inform decision-making in schizophrenia: Relationships with symptoms and functioning. *J. Abnorm. Psychol.* **2016**, *125*, 310. [CrossRef]
91. Gennaioli, N.; Shleifer, A.; Vishny, R. Money doctors. *J. Financ.* **2015**, *70*, 91–114. [CrossRef]
92. Thielmann, I.; Hilbig, B.E. Trust: An integrative review from a person–situation perspective. *Rev. Gen. Psychol.* **2015**, *19*, 249–277. [CrossRef]
93. Motylska-Kuzma, A.; Mercik, J.; Sus, A. Repeatable trust game–preliminary experimental results. In *Asian Conference on Intelligent Information and Database Systems*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 488–498.
94. Markowska-Przybyła, U.; Ramsey, D.M. Norms of reciprocation exhibited by Polish students in the Trust Game: experimental results. *Acta Univ. Lodz. Folia Oecon.* **2016**, *4*, 5–20. [CrossRef]
95. Westhoff, B.; Molleman, L.; Viding, E.; van den Bos, W.; van Duijvenvoorde, A.C. Developmental asymmetries in learning to adjust to cooperative and uncooperative environments. *Sci. Rep.* **2020**, *10*, 21761. [CrossRef]
96. Anderhub, V.; Engelmann, D.; Güth, W. An experimental study of the repeated trust game with incomplete information. *J. Econ. Behav. Organ.* **2002**, *48*, 197–216. [CrossRef]
97. Palminteri, S.; Lebreton, M. Context-dependent outcome encoding in human reinforcement learning. *Curr. Opin. Behav. Sci.* **2021**, *41*, 144–151. [CrossRef]
98. Bellucci, G.; Park, S.Q. Honesty biases trustworthiness impressions. *J. Exp. Psychol. Gen.* **2020**, *149*, 1567. [CrossRef] [PubMed]