# Toward a Unified Theory of Learned Trust in Interpersonal and Human-Machine Interactions

ION JUVINA, Wright State University
MICHAEL G. COLLINS, Wright State University & Air Force Research Laboratory
OTHALIA LARUE, Wright State University
WILLIAM G. KENNEDY and EWART DE VISSER, George Mason University
CELSO DE MELO, University of Southern California

A proposal for a unified theory of learned trust implemented in a cognitive architecture is presented. The theory is instantiated as a computational cognitive model of learned trust that integrates several seemingly unrelated categories of findings from the literature on interpersonal and human-machine interactions and makes unintuitive predictions for future studies. The model relies on a combination of learning mechanisms to explain a variety of phenomena such as trust asymmetry, the higher impact of early trust breaches, the black-hat/white-hat effect, the correlation between trust and cognitive ability, and the higher resilience of interpersonal as compared to human-machine trust. In addition, the model predicts that trust decays in the absence of evidence of trustworthiness or untrustworthiness. The implications of the model for the advancement of the theory on trust are discussed. Specifically, this work suggests two more trust antecedents on the trustor's side: perceived trust necessity and cognitive ability to detect cues of trustworthiness.

CCS Concepts: • **Human-centered computing → Collaborative and social computing theory, concepts and paradigms**;

Additional Key Words and Phrases: Trust, trustworthiness, trust propensity, learned trust, computational cognitive model, unified theories

**ACM Reference format:**
Ion Juvina, Michael G. Collins, Othalia Larue, William G. Kennedy, Ewart de Visser, and Celso de Melo. 2019. Toward a Unified Theory of Learned Trust in Interpersonal and Human-Machine Interactions. *ACM Trans. Interact. Intell. Syst.* 9, 4, Article 24 (October 2019), 33 pages.
https://doi.org/10.1145/3230735



The reviewing of this article was managed by associate editor Xu, Anbang.
This work was supported by the Air Force Office of Scientific Research, grant FA9550-14-1-0206 to IJ.
Authors' addresses: I. Juvina (contact author), M. G. Collins, and O. Larue, Department of Psychology, Wright State University, 3640 Colonel Glenn Hwy., Dayton, OH 45435, USA; emails: {ion.juvina, collins.283, othalia.larue}@wright.edu; W. G. Kennedy, Center for Social Complexity, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA; email: wkennedy@gmu.edu; E. de Visser, Department of Psychology, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA; email: edevisse@gmu.edu; C. de Melo, Institute for Creative Technologies, University of Southern California, 12015 Waterfront Dr., Playa Vista, CA 90094, USA; email: celso.miguel.de.melo@gmail.com.

2160-6455/2019/10-ART24 $15.00
https://doi.org/10.1145/3230735

## 1 INTRODUCTION

The field of trust research affords a productive interplay between theory and practice. Recently, there has been increased interest in research on trust driven by practical problems and applications in areas as diverse as peacekeeping, robotics, healthcare, and education. To meet practical demands, the theory on trust development must bridge conventional gaps between cognition, affect, and meta-cognition, individual and interpersonal, psychology, computer science, and economics, and so on. The recent proliferation of virtual (i.e., geographically distributed and/or technology mediated) teams brings a surge of scientific and practical interest in trust [1, 2].

Allen Newell [3] called for unified theories of cognition specified computationally as cognitive architectures. He mentioned cognitive science was mature enough to start working on unified theories and there should be multiple attempts of such theories. A cognitive architecture is a single system of cognitive mechanisms that operate together to produce the full range of human cognition. Unified theories are the quintessence of scientific progress. They constrain the myriad of possible interpretations of empirical data, facilitate communication among theorists, and motivate new avenues for empirical research. Here, we focus on the field of *learned trust* [4] and attempt to integrate it in the Adaptive Control of Thought—Rational (ACT-R[1]) cognitive architecture [5]. Although the field is already composed of a large volume of empirical findings, micro-theories, meta-analyses, and integrative accounts (e.g., References [6–11]), it could benefit from the kind of integration that is afforded within a cognitive architecture.

Studying trust from a cognitive architecture perspective allows not only integration of various empirical findings from the trust literature but also understanding how trust relates to other cognitive mechanisms and phenomena such as motivation, learning, and strategy choice. Using computational cognitive models as theory-building tools affords modalities of testing the validity of a theory that other ways of theorizing cannot utilize. For example, we routinely employ ACT-R models to generate *a priori* predictions. These are predictions generated by a computational model *before* a human study is conducted (they can also be referred to as *ex-ante* predictions). Typically, a model is developed based on theory, literature, or prior studies and used to generate predictions for new tasks, new experimental conditions, or new manipulations. The study design and setup are identical for model simulations and human data collection, and the simulation data are as rich and fine-grained as the human data. Then, based on the results of the human study, the model is revised, new predictions are generated, and the theory development cycle is repeated. Furthermore, having the theory expressed in computational terms could facilitate its translation into practical applications. For example, the work presented here has the potential to contribute to further developments of ACT-R in the area of robotic autonomy [12].

The starting point for the effort reported here[2] is a published model of learned trust [13], referred to as "the one-counterpart-linear model" or "the old model" in the remainder of the article for reasons that will become clear in the next section. In the next section, we review the key features of the one-counterpart-linear model and discuss its main strengths and limitations. Then, we devote another section to a revised model (also referred to as "the multiple-counterparts-non-linear model" or "the new model" in the remainder of the article) that is intended to overcome the limitations of the one-counterpart-linear model and expand its scope of applicability. Subsequently, we present our model validation efforts and also show that the multiple-counterparts-non-linear model can account for a number of results from the literature on both interpersonal and human-machine trust.

[1]See http://act-r.psy.cmu.edu for software and documentation.

[2]Parts of this paper (some of the sections referring to trust in interpersonal interaction) were presented at the International Conference on Cognitive Modeling (Juvina et al. 2016). The current paper includes an expanded theoretical background, additional model simulation results, a more extensive discussion of the results, and two new sections on individual differences and human-machine interaction, respectively.

In the last section, we discuss possible ways to further improve the multiple-counterparts-nonlinear model and suggest that it has the potential to integrate a wide range of empirical findings, and thus it can inform the development of a unified theory of learned trust.

Before introducing the two models, we specify the terminology used to describe these models (see Table T2 in Appendix A for names, acronyms, and descriptions of the model parameters). *Trait trust* is the term we use for trust propensity (also called dispositional trust in the literature). *State trust* is the trust that develops during a particular interaction in a particular situation and, thus, is a function of the *perceived evidence of trustworthiness* and *perceived evidence of trust necessity*. State trust is characterized by the values of two accumulators, trust and trust-invest, described in a previous paper [13] and in Section 2. In our view, both trait and state trust are learned; trait trust (trust propensity) is learned from the ensemble of past interactions, and state trust is learned from the current interaction. The starting value of state trust at the beginning of the current interaction is the trustor's trait trust. This reflects the finding that humans place a certain amount of trust in strangers that they know nothing about [14]. State trust is updated during an interaction depending on perceived evidence of trustworthiness and perceived evidence of trust necessity. At the end of the current (repeated) interaction, trait trust is updated with an increment that is a function of the state trust developed in the current (just ended) interaction. This reflects the finding that trait trust changes as a function of experience [15]. *Trait trust deviation* is the difference between the trait trust value at the end of the current interaction and the trait trust value at the beginning of this interaction. The trustor's *cognitive ability* is indicated by the accuracy of the trustor's judgments of trustworthiness and trust necessity. An error term is added to model individual differences in the cognitive ability to detect the actual evidence of trustworthiness and trust necessity. The value of the error is sampled from a normal distribution with a mean of 0 and a standard deviation that is a function of the assumed cognitive ability, larger standard deviations corresponding to lower ability.

## 2 DESCRIPTION AND CRITIQUE OF THE ONE-COUNTERPART-LINEAR MODEL

The one-counterpart-linear model[3] [13] was built in the ACT-R architecture and was intended to account for learning within and between two games of strategic interaction—Prisoner's Dilemma (PD) and Chicken Game (CG). These are mixed-motive non-zero-sum games that are played repeatedly between two players.[4] The individually optimal and the collectively optimal solutions may be different. Players can choose to maximize short- or long-term payoffs by engaging in defection or cooperation and coordinating their choices with each other. These features give these games the strategic dimension that makes them so relevant to real-world situations [16]. Table T1 in Appendix A presents the payoff matrices of PD and CG that were used in data collection and modeling [13]. Both PD and CG have two symmetric (win-win and lose-lose) and two asymmetric (win-lose and lose-win) outcomes. Besides these similarities, there are significant differences between the two games. In CG, either of the asymmetric outcomes is more lucrative in terms of joint payoffs than the [1, 1] outcome. This is not the case in PD, where an asymmetric outcome [10,−10] is inferior in terms of joint payoffs to the [1, 1] outcome. Long-run mutual cooperation in CG can be reached by an optimal anti-coordination strategy (i.e., alternation of [−1, 10] and [10,−1]) or a sub-optimal cooperation strategy [1, 1].[5] The optimal strategy in PD corresponds to the sub-optimal strategy in CG numerically, while the optimal strategy of alternation in CG shares

[3]Model code available at: https://science-math.wright.edu/lab/astecca-laboratory/software.
[4]Each player has only one counterpart, hence the name one-counterpart-linear model.
[5]When Chicken is played iteratively, it becomes an anti-coordination game. The long-run alternation outcome (i.e., alternation of [−1,10] and [10,−1]) Pareto-dominates the one-shot cooperation equilibrium [1,1] and is *de facto* mutual cooperation.

no surface-level similarities with the optimal strategy in PD.[6] Thus, although mutual cooperation corresponds to different choices in the two games (i.e., surface-level dissimilarity), they share a deep similarity in the sense that mutual cooperation is, in the long run, superior to competition in both games. In a previous paper [13], we present a model that explains how players learn about each other in a first game and transfer that learning to a second game, regardless of surface dissimilarities between games or the order in which games are played.

The model is not hardwired to play a particular game; it learns to play any $2 \times 2$ game [17] based on the payoff matrix that it experiences as it plays. Initial attempts to account for the transfer-of-learning effects between the two games in both directions (PD-CG and CG-PD) observed in the human data [18] based solely on the existing learning mechanisms of the ACT-R architecture were unsuccessful. A novel *trust learning mechanism* had to be added to the model to account for all the learning and transfer-of-learning effects in the data (see Reference [13] for a detailed justification of adding a trust learning mechanism to the model, including comparisons with alternative models). Essentially, this trust mechanism allows models to learn not only about the task at hand but also about other models with which they interact.

ACT-R [5] is a theory of human cognition and a cognitive architecture that is used to develop computational models of various cognitive tasks. ACT-R is composed of various modules. There are two memory modules that are of interest here: declarative memory and procedural memory. Declarative memory stores facts (know-what), and procedural memory stores rules about how to do things (know-how). The rules from procedural memory serve the purpose of coordinating the operations of the asynchronous modules. ACT-R is a hybrid cognitive architecture including both symbolic and sub-symbolic components. The symbolic structures are memory elements (chunks) and procedural rules. A set of sub-symbolic equations controls the operation of the symbolic structures. For instance, if several rules are applicable to a situation, a sub-symbolic utility equation estimates the relative cost and benefit associated with each rule and selects for execution the rule with the highest utility. Similarly, whether (or how fast) a fact can be retrieved from declarative memory depends upon sub-symbolic retrieval equations, which take into account the context and the history of usage of that fact. The learning processes in ACT-R control both the acquisition of symbolic structures (i.e., learning of new declarative memories and new procedural rules) and the adaptation of their sub-symbolic quantities to the statistics of the environment (i.e., activations of memories and utilities of rules).

Although learning in individual settings has been extensively studied, learning about others has not received much attention in the cognitive modeling field. It is not clear whether learning about other agents uses the same cognitive mechanisms as learning about inanimate entities. Yet, empirical evidence suggests that learning from others and learning about others can influence task-specific learning [19–21]. The one-counterpart-linear model learns about both the game at hand (i.e., PD or CG) and the counterpart (i.e., another model). It uses instance-based learning [22] for counterpart modeling (also referred to as opponent modeling in the literature) [23] and utility learning (a form of reinforcement learning) for action selection. Two ACT-R models run simultaneously and interact with each other (see Figure F1 in Appendix A for a flow diagram of the model).

At each round, each model gets as input the game matrix and the opponent model's previous move (as in the human study). After the two models make their moves, payoff is assigned based

[6]Surface-level similarities are based on descriptive or perceptual features. In contrast, deep similarities are based on structural or functional features. For example, the mutual cooperation outcome [1,1] is identical in PD and CG. Thus, PD and CG share this surface similarity. However, [1,1] is the most efficient outcome in the long run in PD but not in CG. In CG, the most efficient outcome in the long run is an alternation of [−1,10] and [10,−1]. Thus, on a deeper level, [1,1] in PD is similar to alternation in CG. The surface vs. depth distinction is discussed at large in the literature on learning, transfer of learning, and analogical reasoning (Holyoak & Koh, 1987; Gentner & Medina, 1998; Knez & Camerer, 2000).

on the payoff matrix. At each round, the model tries to anticipate the opponent's move based on the opponent's history of moves in similar contexts. To learn what move the opponent is likely to make at each round, the model saves instances (snapshots) of prior contexts and the corresponding moves made by the opponent in those contexts. Depending on the opponent's playing history, one of the alternative instances will be more active and more likely to be retrieved from memory. Thus, the two models try to anticipate each other's current move based on their respective histories of moves. These anticipations occur in conditions of high uncertainty due to variability of individual model behavior and the context of interdependence.

After anticipating the opponent's move, the model must decide on its own move. For this decision, the model leverages the principles of the ACT-R's procedural memory, which is composed of if-then rules. For each possible context (recent moves) and for each possible opponent's move, the model has two decision rules, one for each alternative move that the model can make. Each of these rules can fire whenever the context is instantiated and the opponent is expected to make the corresponding move. Only one rule can fire at a given time—that is, the rule with the higher utility. The utilities of production rules are updated according to the ACT-R utility learning mechanism (a reinforcement learning algorithm). After a number of rounds, one of the two rules corresponding to a context and an expectation will accrue more utility, because it maximizes the reward received by the model. A key question for this model is what the reward is. If the reward is set to the payoff received from the game matrix, the model cannot account for the deep transfer across games found in the human data.

The rewards are determined by the values of two accumulators: trust and trust-invest (see Reference [13] for more details). They change as the game unfolds, depending on the dynamics of the interaction between the two models. The players learn to trust each other, and this affects their reward structure and subsequently their strategies. The trust learning mechanism consists of a "trust accumulator," which represents the perceived trustworthiness of the other model, and a "trust-invest accumulator," which represents the perceived necessity to develop or repair trust. For example, when the two models find themselves in a self-reinforcing cycle of mutual defection, the perceived necessity to develop trust increases. This was a necessary addition to the model to overcome situations in which both players strongly distrust each other and persist in choosing a mutually destructive outcome. We have observed in our studies that humans are able to identify and (sometimes) overcome those situations.

Each accumulator starts at zero. When they both are less than or equal to zero, the model will act selfishly by trying to maximize the difference between their own payoff and the opponent's payoff. This quickly leads to the mutually destructive outcome, which decreases trust in the counterpart but increases the model's perception of trust necessity. Once the latter is positive, a model acts selflessly, trying to maximize the opponent's payoff, in an attempt to signal its willingness to develop trust. Depending on whether the counterpart reciprocates or not, this strategy can lead to mutual cooperation and development of trust, or the two models may relapse into the mutually destructive outcome. When the trust accumulator is positive, the model tries to maximize joint payoff and avoid exploitation. Thus, the model switches between three strategies, depending on its learned trust and trust necessity. Two sets of parameters were fitted to model the human data: standard ACT-R parameters and parameters that were introduced as part of the trust mechanism. The latter are shown in Table 1. The former were: activations noise (i.e., variability in activation of declarative knowledge), retrieval threshold (i.e., minimum activation of a retrievable memory), latency factor (i.e., a parameter that determines the duration of memory retrievals), utility noise (i.e., variability in utility of procedural knowledge), and learning rate (i.e., the rate of learning for procedural knowledge).

The trust learning mechanism was critical to account for this particular dataset (see Reference [13] for a thorough justification of this claim). However, a more general architectural mechanism would be necessary to handle not only trust learning but a more general kind of learning that allows cognitive architectures to perform value-based decision-making. For example, the decision to (not) trust depends on evidence of (un) trustworthiness that is valenced (i.e., positive or negative). This characteristic must be adequately reflected in mental representations that support trust decisions. Currently, ACT-R does not have a general learning mechanism for valenced values for declarative knowledge. When such values are needed (e.g., in instance-based-learning models) they are hand coded.[7] The standard sub-symbolic quantity for declarative knowledge in ACT-R (i.e., activation) cannot be used as a trust accumulator, because it cannot decrease as a function of negative evidence; activation only decreases (i.e., decays) as a function of time. A proposal for such a general architectural mechanism has been presented elsewhere [24].

### 2.1 Strengths of the One-counterpart-linear Model

The one-counterpart-linear model showed how trust learning interacts with task-specific learning to account for a range of learning effects in the human data. The same model was able to account for human data in both games including learning within each game and transfer of learning between games in both directions. This was possible because all the game-specific information was not included in the model but learned from the interaction between players during the games.

The observed transfer of learning was explained based on surface and deep similarities between the two games and the players' ability to think strategically; that is, be aware of their interdependence and choose strategies that balance individual and social motives as well as short- and long-term interests. This model has the potential to inform a unified theory of learned trust, because it is implemented in a cognitive architecture and it specifies how various learning mechanisms interact with, and constrain, each other.

In addition, this model emphasizes the importance of the strategic dimension of trust development, an aspect that is often overlooked in the trust theory. For example, the broad review of the literature by Castelfranchi and Falcone [25], while providing fair criticism to classical game-theoretic approaches, does not do justice to the recent wealth of theorizing and empirical data from behavioral game theory [16]. Furthermore, in some domains, the strategic interaction aspects of trust are not immediately noticeable. Strategic interactions can be characterized as lasting, repeated, cooperative or adversarial, and involving interdependent rational agents balancing multiple motives, constraints, and so on. For example, the interaction between humans and machines is not typically seen as a strategic interaction (i.e., the interaction is usually limited in time and scope, the human is the trustor and the machine is the trustee, and the machine is just a tool—it does not typically assess the trustworthiness of the human or engage in cooperation or competition with the human user) (e.g., Reference [26]). However, when one considers the contexts in which humans and machines interact and the dynamics of this interaction, its strategic nature becomes easier to notice. Even when machines do not have goals and motives, the users and the designers of machines can be seen as strategic players: Users only spend time and effort to use a system if they feel that the designer of the system has their interests at heart [27], and successful systems are designed

[7] In instance-based learning [22], learning consists of accumulation of instances and retrieval of an instance relevant to the current decision situation. An instance contains a decision situation, the action that was taken in that situation, and the value (or utility) of that action. Typically, the instances and their values are hand coded. That is, the modeler initializes the model with a set of instances and specifies the value (for example, correct or incorrect) of the action in each instance. Learning occurs by selective activation of some of the instances. The model makes a decision by selectively retrieving an instance that matches the current situation, contains an action that has a certain value (for example, correct), and has the highest activation. Activation of an instance increases with frequency and recency of occurrence and decreases with time.

by engineers who understand the social and organizational aspects of the system's use [28]. In the near future, one could imagine machines that would be able to engage and develop strategic relationships with humans and other strategic agents. An imaginary example of what a strategic agent might look like is a companion robot that could form a relationship with a human child and maintain that relationship for a lifetime while learning to adapt to changes in the counterpart's knowledge, skills, preferences, priorities, and so on. The one-counterpart-linear model specifies how trust development is influenced by the strategic relationship between the two counterparts.

In agreement with the literature on trust, the one-counterpart-linear model's trust is learned as a function of perceived trustworthiness [7, 10]. Most of the existing computational models of trust include some form of (weighted) aggregation of past evidence of trustworthiness (e.g., References [29, 30]); our one-counterpart-linear model takes advantage of the cognitively plausible memory mechanisms of ACT-R to achieve this aggregation.

In addition, the one-counterpart-linear model asserts that a player's learned trust also depends on perceived trust necessity, which is in and of itself an important contribution to the literature. A validation study based on predictions of the one-counterpart-linear model showed that both perceived trustworthiness and perceived trust necessity are important antecedents of trust formation [15]. When trust is low and trust necessity is high, the model switches to a strategy of maximizing the counterpart's payoff in an attempt to signal its willingness to develop trust. This seemingly altruistic strategy could not be justified based on perceived trustworthiness, but it makes perfect sense in the context of perceived trust necessity. It is also consistent with Mayer et al.'s [7] definition of trust, in that the trustor acts upon their willingness to be vulnerable to the counterpart's actions.

There are many other valuable features of the one-counterpart-linear model that are discussed at length in our previous publications [13, 15]. They support our general conclusion that most of the core assumptions of the one-counterpart-linear model are valid. Next, we focus on the one-counterpart-linear model's limitations (Section 2.2) and improvements (Section 3).

### 2.2 Limitations of the One-counterpart-linear Model

The one-counterpart-linear model assumes that trust starts at zero and only the trust developed during the interaction between the two players matters. However, there is overwhelming evidence that a player may trust another player even in the absence of any interaction between the two players [31], and this initial propensity to trust determines to some extent the trust that develops during the interaction [14, 32]. In addition, trust propensity may be (at least in part) the result of learning that occurred prior to the current interaction. Collins, Juvina, and Gluck [15] measured trait trust (i.e., trust propensity) before and after participants played two games of strategic interaction with a preprogrammed confederate agent and found that trait trust changed slightly but significantly over the course of an experiment that lasted about 45mins (see their Figure 3; see also Figure 1 in this article). By extrapolation, one can conclude that someone's current level of trait trust may be the result of their prior interactions. In other words, trait trust (trust propensity) can also be considered learned trust (at least partially). A comprehensive model of learned trust cannot afford to ignore prior learning, particularly because prior learning may interact with current learning. This aspect was not relevant in the one-counterpart-linear model, because the model interacted with only one other model, but it becomes very relevant in the context of learning from interacting with multiple agents in sequence and transfer of learning from one agent to another (see the black-hat/white-hat effect in the next section).

The one-counterpart-linear model's learning equation is a linear function that increases with every instance of evidence of trustworthiness and decreases with every instance of evidence of untrustworthiness (and similarly for evidence of trust necessity). The rate of accumulation is equal

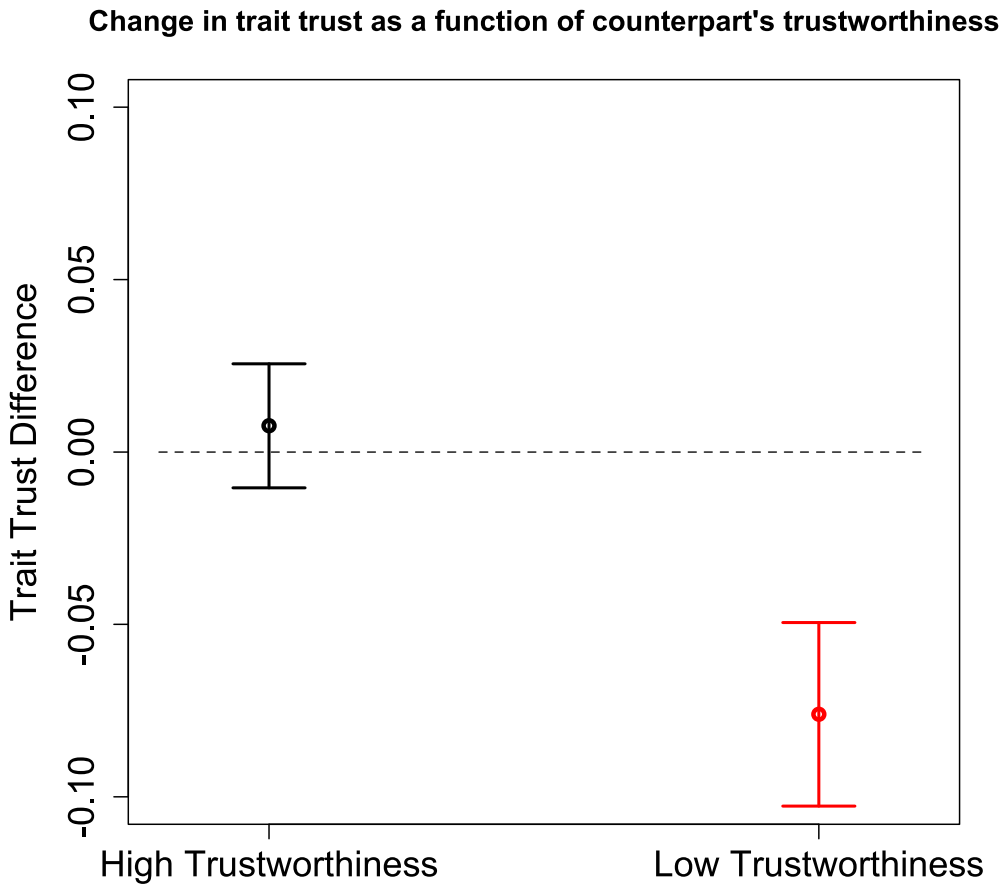


Fig. 1. Change in trait trust (trust propensity) as a function of counterpart's trustworthiness. The measurement units on the Y-axis are subdivisions of a Likert scale. For example, 0.10 is a tenth of a point (e.g., the difference between 3.1 and 3.2) on a 5-point Likert scale. Error bars represent standard errors of the mean. The difference between the two groups (high and low trustworthiness) is significant (Welch Two Sample t-test: $t(271.6) = 2.6$, $p < 0.01$).

for positive and negative evidence and is constant throughout the entire history of interaction. The following is the equation for state trust learning that was used in the one-counterpart-linear model:

$$ST_t = ST_{t-1} + PET_t, \tag{1}$$

where $ST_t$ is state trust at time $t$, $ST_{t\text{-}1}$ is state trust at time $t$-1, and $PET_t$ is perceived evidence of trustworthiness at time $t$. A similar equation was used for trust necessity. This equation worked well in the context of the one-counterpart-linear model but is problematic, because it is not in full agreement with what is known about the dynamics of trust. Trust is hard to gain and easy to lose, a characteristic that has been referred to as trust asymmetry[8] [24, 33]. Trust learners exhibit the same negativity bias that is described in the impression formation literature [35, 36]; that is, unfavorable information tends to be more influential than favorable information or experience. Evidence of

[8]Trust asymmetry refers to the different effect of trustworthiness and untrustworthiness on the dynamics of trust: Trust increases slightly following evidence of trustworthiness but decreases abruptly following evidence of untrustworthiness. It does not refer to asymmetry between trustor and trustee.

trust asymmetry was also observed in our aforementioned study [15]. Figure 1 shows changes in the trustor's trait trust (trust propensity) as a function of the trustee's trustworthiness. When the trustors interacted with high-trustworthiness confederates, their trait trust increased slightly but not significantly. When the trustors interacted with low-trustworthiness confederates, their trait trust decreased significantly. Thus, negative experiences were more influential than positive experiences. Though the difference in the trustor's trait trust made by a 45-minute interaction with a high- or low-trustworthiness counterpart is very small (i.e., less than a tenth of a point on a 5-point Likert scale), the accumulation of similar experiences over a lifetime can be substantial, explaining why people can have very high or very low trust propensity (i.e., trait trust).

Another known effect that was only partially captured by the one-counterpart-linear model was the fact that early evidence has a stronger impact on trust formation than late evidence [37]. More generally, the literature on impression formation shows that first impressions matter more than later impressions for person perception [38, 39].

Another limitation of the one-counterpart-linear model is that it assumes that all trustors are able to assess trustworthiness and trust necessity equally well. However, a trustor's cognitive ability to assess a trustee's trustworthiness has been proposed to be an important antecedent of trust [40–42]. In general, cognitive ability is an important predictor of learning, thus it is not surprising that it is also related to learned trust. Last, learning equations tend to be power functions [5, 43], and it would be surprising if trust learning were an exception.

Last, the one-counterpart-linear model was somewhat limited in scope. In this article, we begin to address the question of whether the model can be generalized to domains other than strategic interaction; for example, trust in automation or in autonomous agents.

### 2.3 Multiple-counterparts-non-linear Model

The revision[9] of the one-counterpart-linear model consists of replacing the linear function that was used to update the trustor's state trust with the following power function:

$$ST_t = ST\frac{a}{t-1} + PET_t - b * TTD, \tag{2}$$

where $ST_t$ is state trust at time *t*, $ST_{t\text{-}1}$ is state trust at time *t*-1, $a$ is a constant power exponent with a value less than 1 ($a < 1$ will be referred to as the trust decay parameter), $PET_t$ is perceived evidence of trustworthiness at time *t*, *TTD* is the trait trust deviation computed after the previous interaction with another person, and $b$ is the perception bias that scales how much $PET_t$ is adjusted as a function of the trustor's previous experience with another trustee. A similar equation was used for trust necessity.

In the multiple-counterparts-non-linear[10] model, both trait and state trust are positive or zero. A value of zero signifies the absence of trust. The evidence of trustworthiness can be positive, indicating a degree of trustworthiness, or negative, indicating a degree of untrustworthiness. The initial value of state trust is the value of trait trust that was updated after the previous interaction with another person ($ST_{t0} = TT$). In our simulations, we set the initial trait trust somewhere in the middle of the range of values that state trust can take during a repeated interaction with a specific person, depending on the range of values that the evidence of trustworthiness can take. We assume that weighting of the evidence is task-specific.

The actual evidence of trustworthiness (*AET*) may be perceived more or less accurately, resulting in perceived evidence of trustworthiness (*PET*). An error term $e$ is added to represent the trustor's

[9]Model code available at: https://science-math.wright.edu/lab/astecca-laboratory/software.

[10]The assumption here is that each player interacts with one counterpart at a time but with multiple counterparts in sequence, hence the name multiple-counterparts-non-linear model.

imperfect ability to detect and decode trustworthiness signals.

$$PET = AET + e \tag{3}$$

The continuous value of state trust is used to make categorical judgments (i.e., trust or distrust) by comparing it against the value of trait trust. If the current value of state trust is greater than the value of trait trust, then the trustor is said to trust the trustee. If the current value of state trust is less than the value of trait trust, then the trustor is said to distrust the trustee.

Trait trust (trust propensity) is updated when changing counterparts—that is, at the end of an interaction with a counterpart and before interacting with another counterpart. The update in trait trust (i.e., trait trust deviation, TTD) is a function of the state trust. At this moment, we don't have a clear idea of what this function might look like. We know from our empirical work [15] that TTD is very small (as compared to the range of state trust) and is positive after an interaction that increased state trust and negative after an interaction that decreased state trust. In other words, trait trust tends to follow the dynamics of state trust across different counterparts but on a much smaller scale. For current purposes, we use the following formula:

$$TTD = \frac{ST - TT}{k}, \tag{4}$$

where $TTD$ is trait trust deviation computed after interacting with a counterpart, $TT$ is the old value of trait trust that is being updated, $ST$ is the current value of state trust, $ST–TT$ is the value of state trust that was accrued in that interaction, and $k$ (higher or equal to 1) is a parameter that scales down the value of state trust. $TTD$ is added to the old $TT$ to compute the new $TT$.

The power exponent $a$ must be lower than 1 to make trust a leaky accumulator. It is currently set to 0.99 in all our simulations. The assumption behind this component of the equation is that the more recent values are more important than the older values of state trust. A consequence of this assumption is that trust decays in time if new evidence of trustworthiness is not perceived. Note that for $a = 1$ and $TTD = 0$, Equations (1) and (2) are identical.

Figure 2 shows a hypothetical case in which a trustor repeatedly interacts with a trustee for 200 rounds. The trustor perceives evidence of trustworthiness ($PET = 1$) for the first 100 rounds, then evidence of untrustworthiness ($PET = -1$) for 5 rounds, and then again evidence of trustworthiness ($PET = 1$) for the remaining 95 rounds. State trust accumulates rapidly in the first 50 rounds, after which it starts to approach an asymptote—that is, a state of diminishing returns for every new piece of evidence of trustworthiness. In addition, the state trust that was accumulated over 100 rounds is lost almost entirely in 5 rounds, manifesting trust asymmetry [34].

The term trait trust deviation ($TTD$ in Equation 2) becomes relevant when a trustor interacts with multiple trustees in sequence. In such cases, empirical studies suggest that the experience from a previous interaction influences how the trustor perceives the evidence of trustworthiness in the current interaction. For example, De Melo, Carnevale, and Gratch [44] review evidence and possible explanations for the black-hat/white-hat (or bad-cop/good-cop) effect from the negotiation literature: Playing a first game with an opponent with a competitive stance (black-hat) followed by a second game with an opponent with a cooperative stance (white-hat) is more effective in reducing distance to agreement than any other pairing of the black-hat and white-hat opponents [45]. We implemented the explanation of the black-hat/white-hat effect that is based on the concepts of adaptation and comparison level [46]. Theories of adaptation propose that people become accustomed to a reference point as a result of prior experience; this point then serves as a comparison for the judgment of subsequent experiences. Thus, a cooperative second bargainer should be judged as more cooperative if the first bargainer was competitive rather than cooperative. In terms of our learned trust theory, the prior experience of untrustworthiness shifted the


Fig. 2. A hypothetical case illustrating how state trust changes over the course of 200 rounds of interaction with another player. The trustor perceives evidence of trustworthiness for the first 100 rounds, then evidence of untrustworthiness for 5 rounds, and again evidence of trustworthiness for 95 rounds.

trustor's reference point toward low values of trustworthiness. In this context, evidence of trustworthiness from a new interaction is perceived as outside of the expected range, which gives it a larger subjective weight. In our model, we assume that the change in the subjective perception of the new evidence is proportional to the adjustment (i.e., adaptation in Helson's [46] terms) of the reference point caused by the previous experience. The reference point is the trustor's trait trust and the adjustment is trait trust deviation. For example, if the trustor's previous experience with an untrustworthy trustee caused a large shift (i.e., deviation) in her trait trust, the corresponding bias in her subjective perception of a new trustee will also be large (and vice versa). Thus, a trustor's previous trait trust deviation (*TTD*) determines the extent to which the perceived evidence of trustworthiness (*PET*) is adjusted.

To conclude the description of the multiple-counterparts-non-linear model, only the trust learning mechanism has been revised; all the other mechanisms of the one-counterpart-linear model (i.e., learning to anticipate the opponent's move and to select the best move in each context) [13] have been left unchanged.

Table 1. The Best Fitting Parameter Values for the Multiple-Counterparts-non-linear (MCNL) Model and the One-counterpart-linear (OCL) Model for Both Trust Accumulators (Trust and Invest) and for Each of the four Game Outcomes, Mutual Cooperation (CC), Unilateral Cooperation (CD), Unilateral Defection (DC), and Mutual Defection (DD)

| Outcome | OCL model | | MCNL model | |
|---|---|---|---|---|
| | Trust | Invest | Trust | Invest |
| CC | 3 | NA | 6 | NA |
| CD | −10 | −1 | −7 | −1 |
| DC | 10 | NA | 9 | NA |
| DD | −1 | .18 | −1* | .18* |

*Note*: An asterisk (*) indicates that a value that was held constant during the model-fitting procedure.
"Trust" Refers to the Increment (or Decrement) in the Trust Accumulator as a Result of an Instance of Perceived Evidence of Trustworthiness (or Untrustworthiness). "Invest" Refers to the Increment (or Decrement) in the Trust-invest Accumulator as a Result of an Instance of Perceived Evidence of Trust Necessity (or Lack Thereof).

## 3 MODEL VALIDATION

We expect that the multiple-counterparts-non-linear model is able to generalize to a wide range of empirical phenomena while maintaining the ability of the one-counterpart-linear model to explain the learning and transfer-of-learning effects from the original dataset.

### 3.1 Learning and Transfer of Learning in Prisoner's Dilemma and Chicken Game

Juvina et al. [13] recruited 120 participants to play Prisoner's Dilemma (PD) and Chicken Game (CG) for 200 rounds each. The participants were paired with one another and assigned to play the two games in one of two order conditions: PD-CG and CG-PD. The results revealed a wide range of within-game learning and between-game transfer-of-learning effects. The dependent variable consisted of round-by-round proportions of four outcomes: mutual cooperation, unilateral cooperation, unilateral defection, and mutual defection. The one-counterpart-linear model was fit in its entirety to this dataset by varying 11 free parameters (see Table 4 in Reference [13]). With regard to the multiple-counterparts-non-linear model, only the 6 free parameters associated with the trust mechanism were refit to the human data reported in Juvina et al. [13]. Four of the 6 parameters are associated with the "trust accumulator" that represents the perceived trustworthiness of the other player and the other 2 are associated with the "trust-invest accumulator" that represents the perceived necessity to develop trust (see Table 1). The values of these parameters specify how much perceived evidence of trustworthiness (*PET* in Equations 1 and 2) is added to (or subtracted from) the trust accumulator for each outcome of the game. Two of the 6 parameters (i.e., the parameter with the lowest absolute value for each accumulator) were kept at their values from the old model, thus allowing only 4 model parameters to freely vary to find best-fitting results. The range of parameter values for the model-fitting procedure was [−10, 10]. The fit procedure maximized the correlation ($r$) and minimized the root mean squared deviation (RMSD) between the model data and the human data.[11]

[11]High performance computing facilities at the Air Force Research Laboratory and the web service mindmodeling.org (Harris 2008) were used for the model-fitting procedure.

Table 1 shows the best-fitting parameter values for the multiple-counterparts-non-linear model and the one-counterpart-linear model. One of them did not change at all, even though it was allowed to vary freely. The other three parameters have been readjusted in the multiple-counterparts-non-linear model. These parameters were held constant for all but one of the simulations reported below. They were readjusted for Lount et al. [37] data, because a very different payoff matrix was used in that study.

The fit of the multiple-counterparts-non-linear model to the human data ($r(798) = .90$, $p < 0.01$, $RMSD = .07$) was slightly (but not significantly, $z = 1$, $p = 0.32$) better than the fit of the one-counterpart-linear model ($r(798) = .89$, $p < 0.01$, $RMSD = .09$). The multiple-counterparts-non-linear model also exhibited the same transfer-of-learning effects observed in the human data.

Collins et al. [15] conducted a follow-up study in which 320 participants recruited from the website Amazon Mechanical Turk played PD and CG for 50 rounds each in one of four possible game orders (PD-PD, PD-CG, CG-PD, or CG-CG). Participants were paired with computerized confederate agents whose behavior (i.e., strategy and trustworthiness) was manipulated to result in 16 different experimental conditions. The one-counterpart-linear model (Juvina et al. 2015) was used to generate *a priori* predictions for the Collins et al. [47] study. The predictions were published before the data were collected [47]. A majority of the model predictions across all of the 16 experimental conditions was supported and the trust mechanism was proven to be a necessary component of the one-counterpart-linear model (see Collins et al. [15] for details). Here, we test the multiple-counterparts-non-linear model against the dataset from Collins et al. [15] without any parameter optimization. The dependent variable consisted of round-by-round proportions of five outcomes: mutual cooperation, unilateral cooperation, unilateral defection, mutual defection, and alternation. The multiple-counterparts-non-linear model accounts for the human data slightly ($z = 2$, $p = 0.05$) better ($r(1598) = .68$, $p < 0.01$, $RMSD = .33$) than the one-counterpart-linear model ($r(1598) = .64$, $p < 0.01$, $RMSD = .33$).

### 3.2 Unified Account of Trust and Distrust Effects

It has been proposed that trust and distrust are different constructs [48, 49]. Here, we suggest that the different dynamics of trust and distrust can be modeled by a single equation. In Section 3, we showed how Equation (2) produces trust asymmetry (see Figure 2). A consequence of trust asymmetry is the fact that early trust breaches are more influential than late trust breaches for the overall trust that develops in a repeated interaction, which is exactly what Lount et al. [37] found. They conducted two experiments in which participants played an iterated game of Prisoner's Dilemma for 30 rounds. Participants were assigned to one of four experimental conditions (control, immediate, early, and late) and played the game with a confederate agent whom they were told was another participant. During the control condition, the confederate agent cooperated on all 30 rounds. In the other three conditions, the confederate agent cooperated on each round except for two consecutive trials on which it defected. These trust breaches could occur immediately (rounds 1 and 2), early (rounds 6 and 7), or late (rounds 11 and 12). The main finding revealed that the immediate and early breaches significantly decreased the proportion of cooperation during the last 10 rounds of the game as compared to the late breach (see Lount et al. [37] for more details).

Our multiple-counterparts-non-linear model is able to account for the basic pattern of results—that is, the different amounts of cooperation in control, immediate, early, and late conditions ($r(118) = 0.99$, $p < 0.01$, $RMSD = 0.30$). The old one-counterpart-linear model produced a poorer ($z = 5.36$, $p = 0.00$) fit ($r(118) = 0.96$, $p < 0.01$, $RMSD = 0.31$). One possible explanation for the large root mean square deviation (RMSD) is a manipulation in the experiment that was not modeled: Participants read a passage about the importance of cooperation before the start of the game. Our multiple-counterparts-non-linear model is able to explain Lount et al.'s findings based on the

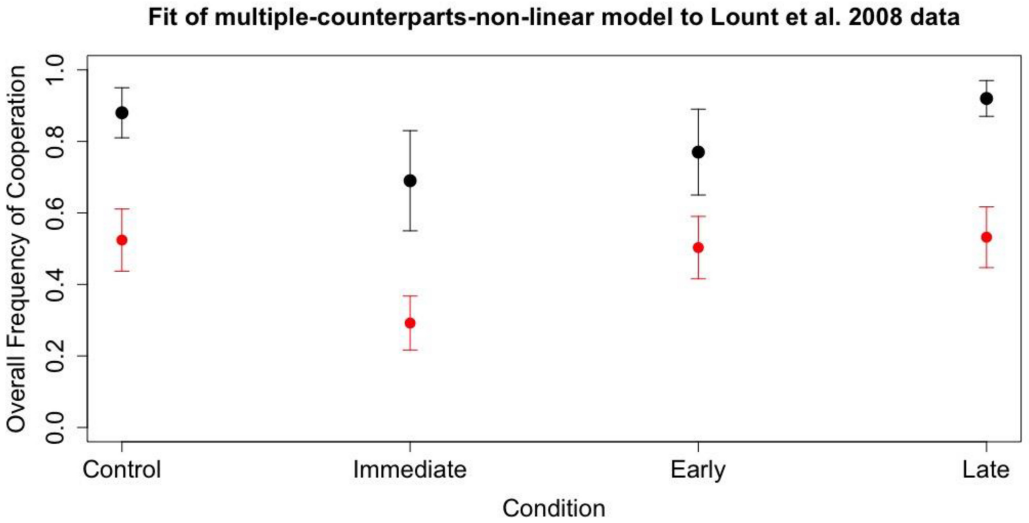


Fig. 3. Fit of multiple-counterparts-non-linear model (red dots) to Lount et al. [2008] data (black dots). Error bars represent 95% confidence intervals.

dynamics of state trust. Reestablishing trust after a breach is a long process. In the case of early breaches, most of the rounds of the interaction are used to (slowly) reestablish trust. In the case of late breaches, most of the trust accumulates before the breach, leaving a smaller number of rounds of interaction to be damaged by the breach. This is consistent with results from the impression formation literature, emphasizing the importance of making a good first impression [50].

The difference between the old one-counterpart-linear model and the new multiple-counterparts-non-linear model with regard to how they fit Lount et al.'s dataset is small, because this dataset is very simple. Of the four conditions in Lount et al.'s [37] study, only the immediate condition leads to distrust. In the other conditions (i.e., control, early, and late), when defection occurs (if at all) the accumulated trust is high enough to survive and almost never turns into distrust. The small difference is made by the fact that the new model does not start its trust accumulator at zero as the old model; instead, the initial value of trust for the new model is its trait trust (i.e., trust propensity) value. An important qualitative difference is related to how the two models account for the difference between early and late breaches observed in the human data (see Figure 3 and Figure 2F in the Appendix). Arguably, the difference between the two models would be higher in more complex datasets (see Section 4.5 for a much larger difference between the two models).

### 3.3 Black-hat/white-hat Effect

De Melo, Carnevale, and Gratch [44] had participants play Prisoner's Dilemma with two different computerized confederate agents (cooperative and individual). Each agent was represented by a different animated face. Both agents used the same strategy (tit-for-tat), but displayed different facial expressions representing different emotional reactions to particular outcomes during the game (e.g., the cooperative agent expressed joy after instances of mutual cooperation and the individual agent expressed joy after instances of unilateral defection). The authors suggested that participants used reverse appraisal to identify, from the agents' emotional displays, what the intentions of the agent were. The cooperative agent expressed emotions congruent with attempting to maximize the joint payoff of both players, whereas the individual agent expressed emotions congruent with attempting to maximize its own payoff. Participants played 25 rounds with each

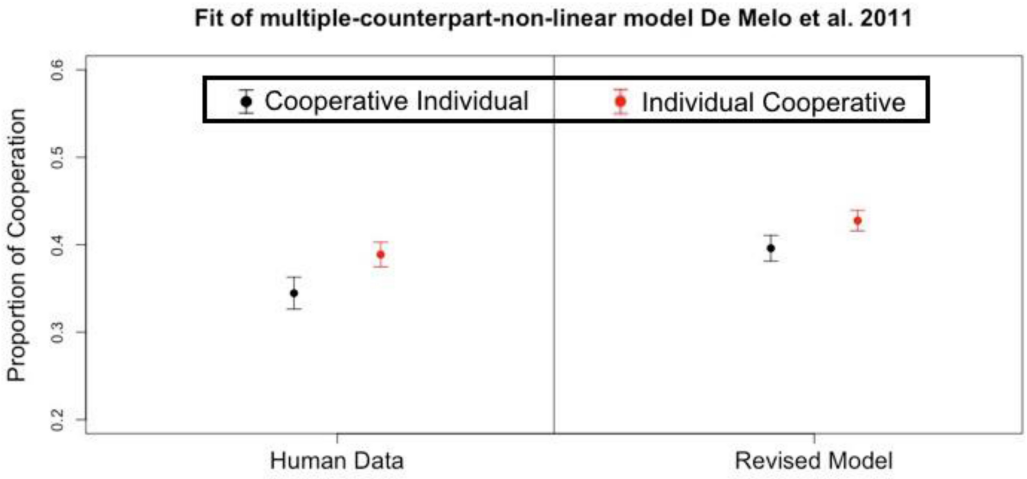


Fig. 4. Fit of multiple-counterparts-non-linear model (right panel) to De Melo et al. [37] data (left panel). Error bars represent 95% confidence intervals.

of the confederate agents in one of two orders: the cooperative agent then the individual agent (C-I), or the individual agent and then the cooperative agent (I-C). Given that the strategy of the two agents was identical, the differences in trustworthiness between the two agents could only be attributed to what was inferred from facial expressions. Other authors have also shown that the pattern of trust learning can be influenced by incidental learning from facial expression, eye gaze, and so on. (e.g., Reference [51]). De Melo et al. [44] found that participants were sensitive to the emotions displayed by the two agents: they cooperated more with the cooperative agent than with the individual one. In addition, they found evidence for the black-hat/white-hat effect; that is, cooperation was higher when playing the first game with the individual agent (black-hat) and the second game with the cooperative agent (white-hat) than vice versa.

We did not explicitly model the process of inferring trustworthiness from facial expressions. Instead, we added 12 parameters that translated particular emotions into specific amounts of evidence of trustworthiness and trust necessity (see Table T3 in Appendix A). We needed them to differentiate between the cooperative and non-cooperative agents; these agents were behaviorally identical and only differed in their emotional reactivity. However, these parameters by themselves did not make the model exhibit the black-hat/white-hat effect. The key difference was made by the trait trust deviation parameter ($TTD$ in Equation (2)) and the trust decay parameter ($a$ in Equation (2)), which allowed the model to fit the human data ($r(98) = 0.86$, $p < 0.01$, $RMSD = 0.11$) and reproduce the black-hat/white-hat effect (Figure 4). The fit of the old one-counterpart-linear model was not as good ($r(98) = 0.75$, $p < 0.01$, $RMSD = 0.14$; $z = 2.23$, $p = 0.03$). Moreover, the old model failed to produce the black-hat/white-hat effect (see Figure F3 in the Appendix). We conducted additional analyses and determined that both components of the new multiple-counterparts-non-linear model were necessary to produce the black-hat/white-hat effect: the multiple-counterparts component (parameter $TTD$ in Equations (2) and (4)) and the non-linear component (parameter $a$ in Equation (2)). Two lesioned versions of the multiple-counterparts-non-linear model were created: the no-trait-trust-deviation model ($TTD = 0$) and the no-non-linear model ($a = 1$). Both lesioned models produced significantly worse fits to the human data than the intact model ($r(98) = 0.64$, $p < 0.01$, $z = 3.73$, $p = 0.00$, $RMSD = 0.16$; $r(98) = 0.66$, $p < 0.01$, $z = 3.49$, $p = 0.00$, $RMSD = 0.16$, respectively).

### 3.4 Cognitive Ability and Trust

Up to this point, we examined the model's ability to account for effects observed in different conditions when data are averaged across individuals. Here, we consider how the model accounts for a known effect from the trust literature that focuses on differences between individuals.

Prior research found the accuracy of trustworthiness detection to be positively correlated with the participants' self- reported trait trust ($r = .48$) and sense of interdependence ($r = .55$) [42]. That is, individuals who are better or faster at detecting (un)trustworthy counterparts tend to report higher levels of trait trust and are more aware of the fact that their interests may not be independent of their counterparts' interests. The assumption is that trustworthiness is not always evident or easy to detect and trustees may have reasons to concede their actual level of trustworthiness. Trustors need to exert social intelligence [42] or strategic vigilance [52] to detect trustworthiness based on signals or cues. If we define trait trust as "default expectations of trustworthiness of others" (Yamagishi et al. [42], p. 158), the above correlation can be stated as "the better you are at detecting trustworthiness, the higher you expect it to be by default." As unintuitive as it may seem, this correlation has been shown to be robust. Sturgis, Read, and Allum [41] found that intelligence measured in childhood predicted trait trust in adulthood. Lyons et al. [2] found that cognitive ability was positively correlated with self-reported trust ($r = 0.27$) in a realistic task—a computer-based airport simulator.

This correlation becomes easier to understand if we assume that trait trust is learned from prior experiences, as we have observed empirically [15] and modeled here. Our model translates this assumption into its state trust (*ST*) and trait trust (*TT*) update equations (Equations (2) and (4), respectively). First, the model allows for the possibility that the trustee's actions are only partially observable, as in Reference [53]. The trustor needs to interpret the trustee's actions and infer trustworthiness, which requires cognitive ability. Higher levels of cognitive ability lead to more accurate estimates of trustworthiness; that is, the perceived evidence of trustworthiness (*PET* in Equations (2) and (3)) is close or equal to the actual evidence of trustworthiness (*AET* in Equation (3)), approaching perfect trust calibration [9]. Calibrated (i.e., appropriate) trust allows individuals to maximally benefit from interpersonal or human-machine interactions. Misplaced (i.e., too much or too little) trust leads to either being exploited or missing opportunities for gain. We model different levels of cognitive ability by introducing errors of different magnitudes in Equation (3). Thus, higher cognitive ability corresponds to smaller errors and vice versa.[12]

We created 51 "individual" models by varying cognitive ability (i.e., error magnitude or $e$ in Equation (3)) and allowed them to play 100 rounds of Prisoner's Dilemma with 10 different counterparts in sequence. Each counterpart was the "average" model initialized with a random level of trait trust, reflecting its unique history of playing against other counterparts. The error $e$ was fixed for a model across all its counterparts. Each individual model was initialized with a random level of trait trust at the start of the first game. Then, starting with the second game (i.e., the first counterpart change), trait trust for individual models was updated by adding trait trust deviation computed according to Equation (4). State trust was updated at each round according to Equation (2). The whole process was repeated 10 times to collect enough data. The key data were the trait trust values of the 51 models after interacting with 10 different counterparts. These values were correlated with the cognitive ability values for the 51 models. The results support the existence of a positive correlation between cognitive ability and trait trust (see Figure 5). The magnitude of this correlation is comparable to the one reported by Lyons et al. [2]. Higher correlations

[12] We did not model cognitive ability per se but the outcomes of trustworthiness assessment at different levels of cognitive ability.

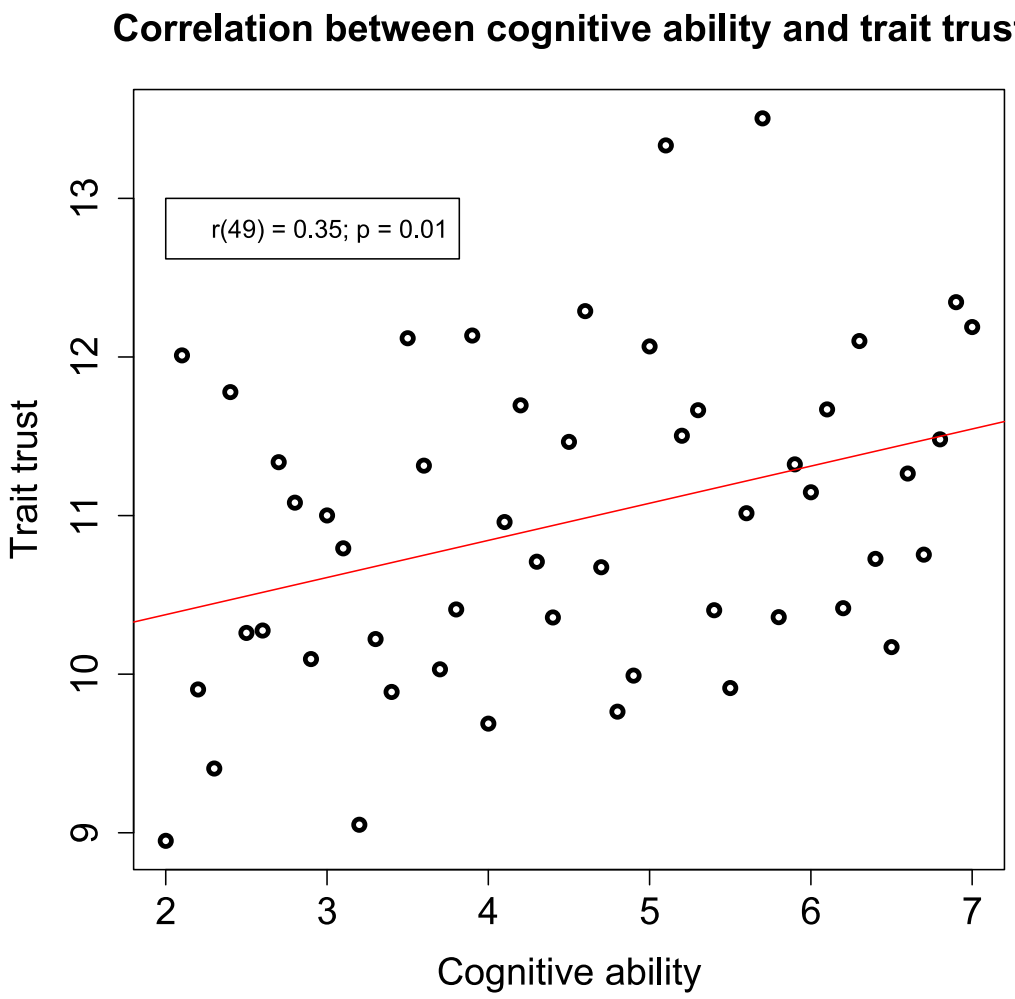


Fig. 5. Correlation between cognitive ability and trait trust.

can be obtained by decreasing the scaling parameter $k$ in Equation (4) ($k = 25$ for the data reported here) and reducing the amount of random variability in the data.

To understand the mechanism that explains this correlation, we divided the models in two groups (median split) based on their cognitive ability and plotted the average proportion of mutual cooperation for each group (Figure 6). On average, the models with higher cognitive abilities were able to engage in mutual cooperation more frequently and maintain it for longer times than models with lower cognitive abilities. Every instance of mutual cooperation provided evidence of trustworthiness that increased state trust and eventually increased trait trust as well. The low-ability model made errors in assessing trustworthiness, and these errors made it hard to maintain the mutual cooperation outcome. Thus, the model supports the explanations given by Yamagishi et al. [42] and Sturgis et al. [41] according to which socially intelligent individuals—by accurately assessing trustworthiness (or lack thereof) in their counterparts—are able to enact and maintain mutual cooperation while protecting themselves against exploitation.

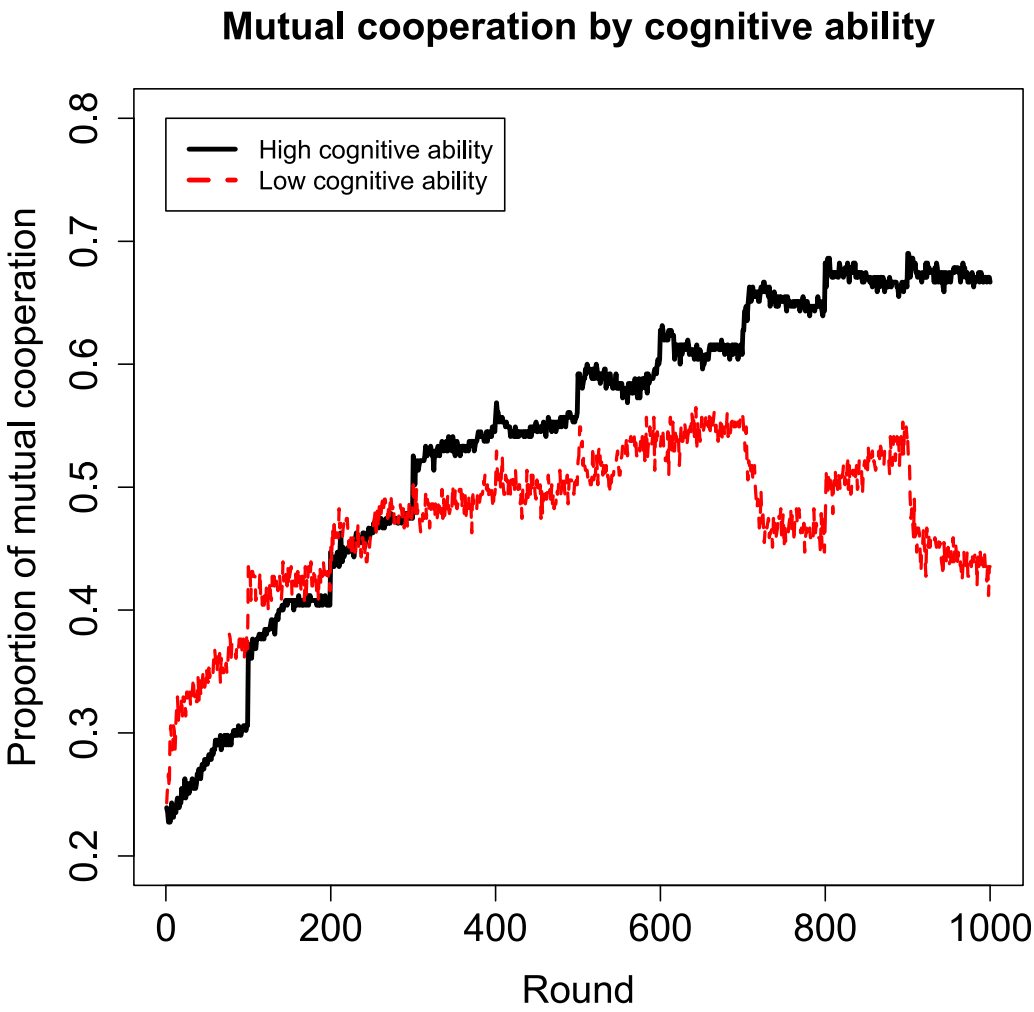


Fig. 6. Proportion of mutual cooperation by cognitive ability.

### 3.5 From Interpersonal to Human-machine Trust

Up to this point, we examined the model's ability to account for a variety of empirical effects in the domain of interpersonal trust. Even though some of the counterparts were pre-programmed computer algorithms or computational cognitive models, the assumption was that they were behaving and perceived as human-like. In this section, we begin to explore the similarities and differences between interpersonal trust and human-machine trust and test whether our model can account for some of these effects. Recently, De Visser et al. [54] studied trust dynamics (i.e., formation, violation, and repair of trust) in an experiment in which human participants received advice from a computer, an anthropomorphized computer (avatar), or another human. The task that human participants had to perform was a sequence-learning task. Specifically, they had to predict the next digit from a sequence of digits that was presented one digit at a time. At each trial, the participants made a guess for the next digit in the sequence, then received advice, had the opportunity to take the advice or keep their initial response, and received feedback with the correct response. The advice was identical in the three conditions and was gradually deteriorating in accuracy. The

results revealed an interesting difference between interpersonal and human-machine trust: Human participants manifested greater trust resilience in the human and avatar conditions than in the computer condition. Trust resilience was defined as resistance to trust violations (more details can be found in De Visser et al. [54]).

This dataset allows us to test whether our model can generalize to domains other than strategic interpersonal interaction. There are interesting similarities and differences between the task used in the De Visser et al. study (i.e., sequence learning, SL) and the game-theory-like tasks we have been using so far in developing and testing the model (e.g., Prisoner's Dilemma, PD). In both types of tasks, participants have to predict the next item in a sequence, but in PD the next item is generated by a counterpart (who is also predicting the next move of a counterpart), whereas in SL it is predetermined by the experimenter (in game theory, this case is treated as a game against nature). If we ignore the experimenter (i.e., the nature player), in both types of tasks there are two players, but the interaction between the two players is different. The interaction is symmetrical in PD and asymmetrical in SL: In PD, each player is simultaneously a trustor and a trustee, whereas in SL the advisee can only be a trustor and the advisor can only be a trustee. The critical difference seems to be about strategic interdependence: PD obviously has it and SL does not have it, because the advisor is totally independent of the advisee.

To be able to apply our model to the sequence-learning task from De Visser et al. [54], we assume that people behave as if they are in a strategic interpersonal interaction even when the strategic dimension of the interaction is not apparent. For example, (some) people may regard their computer (or car, cell phone, vacuum cleaner, etc.) as a friend (or partner) and manifest attitudes and behaviors specific to strategic interpersonal interaction when using it (e.g., be nice to a cell phone). This may seem silly, but there may be good reasons to do so; for example, it may put one in a better mood. Furthermore, we assume that the extent to which people engage in such behaviors is proportional to the degree of similarity between the machine and a generic being—that is, how human-like or animated the machine is perceived to be. Very simple features can influence perceptions of animacy; for example, task-irrelevant abstract geometric shapes are automatically perceived as intentional agents when they move in certain ways [55]. Having made these two assumptions, we can use our model to account for the difference in trust resilience between the human-human, human-avatar, and human-computer interactions. After we model the task, we will change two parameters, trait trust (i.e., trust propensity) and trust necessity, to account for this difference (see Table T4 in Appendix A for the best-fitting values of these parameters). Trait trust will be set higher for the computer and the avatar and lower for the human. This reflects the assumption that machines are designed to be trustworthy, whereas humans may show a larger range of trustworthiness values due to conflicting motives, ability to deceive, and so on. Trust necessity will be the highest for the human-human interaction, lower for human-avatar interaction, and the lowest for the human-computer interaction. Thus, the model will assume that humans are more likely to invest in trust development when they interact with another human than when they interact with a computer, even in cases where there is no apparent reason or benefit from doing so.

The sequence-learning (SL) model presented here is based on a model developed independently by one of the co-authors of this article (WK). We have only added the trust mechanism from our previous model [13, 56], including the advancements described in Section 3 (i.e., the multiple-counterparts-non-linear model) to account for the extent to which participants take the offered advice or keep their initial choice at each trial. The following is a brief description of the model (the model code is available from the authors upon request):

First, the model reads the identity of the advisor (i.e., human, avatar, or computer) from the display and encodes it in its goal. Then, it makes a first guess as to what digit in the sequence comes next. At first, this is truly a guess (i.e., random choice), then, as trials accumulate, the model learns

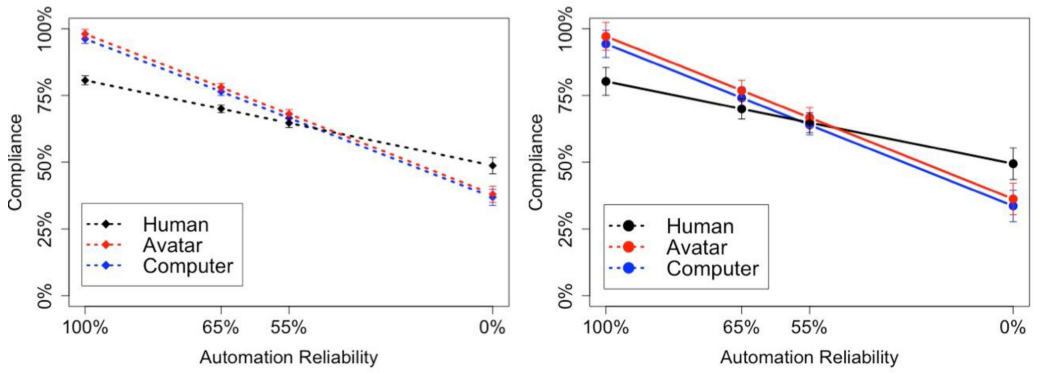


Fig. 7. The least mean square estimates (+/− standard error) for compliance with the three human (black line), avatar (red line), and computer (blue line) agents for both the model (left-side plot) and human data (right-side plot).

from the feedback given at the end of each trial, so, over time it becomes an informed guess. For this learning, the model uses the same mechanisms that it used to predict the counterpart's move in PD; that is, instance-based learning and sequence learning. Specifically, the model stores instances of context-choice pairs in its declarative memory and retrieves the most active pair to make a decision. For example, the model might have experienced five cases in which "2" was followed by "1" and three cases in which "2" was followed by "3." Based on this experience, if the correct answer in the previous trial was "2," the model is more likely to think that the correct answer in the current trial is "1." Since activations of memories in ACT-R tend to approximate the pattern of occurrences and co-occurrences in the environment [5], over time the model learns the actual sequence that was pre-specified by the experimenter. However, the pre-specified sequence includes randomness, causing the model's performance to always be (more or less) suboptimal, which justifies considering the aid's advice. After perceiving and encoding the aid's advice, the model decides whether to take the aid's advice or keep its initial choice. This decision is informed by the trust mechanism that was imported from the strategic interaction model. The SL model learns to trust by monitoring the aid's trustworthiness and trust necessity as in the strategic interaction model. In addition, the SL model learns to self-trust by monitoring its own trustworthiness and trust necessity. Trustworthiness increases with correct guesses and decreases with incorrect guesses. Trust necessity for the aid (i.e., need to trust the aid) increases when both the model and the aid are incorrect and decreases when the model is correct and the aid is incorrect. Trust necessity for the model (i.e., need to trust itself) increases when both the model and the aid are incorrect and decreases when the model is incorrect and the aid is correct. Thus, in the SL model, accuracy was used as perceived evidence of trustworthiness and trust necessity (PET in Equation (2)) and no perception bias or variability in cognitive ability was assumed. The model decision to take the aid's advice or not is based on the values of trust and self-trust: The advice is taken when trust is higher than self-trust and not taken otherwise. If both trust and self-trust are decreasing, the decision is based on trust necessity and self-trust necessity: The advice is taken when trust necessity is higher than self-trust necessity and not taken otherwise. The free parameters for the SL model were the initial trait-trust (trust propensity) values and the increments for updating trustworthiness and trust necessity.

The SL model fits the data from De Visser et al.'s [54] experiment 1 very well ($r(10) = .99$, $p < 0.01$, $RMSD = .02$). Figure 7 shows the least mean squared estimates for the model's rate of compliance with the counterpart's advice over the course of the experiment compared to the human

data from De Visser et al.'s [54] experiment 1.[13] During the initial block, when the agent's advice was 100% accurate, the differences in compliance are due to the model's initial trait trust (trust propensity). As the experiment progresses, the model learns and its assessment of the different automated aids changes. During the final block, when the advisor's accuracy is 0%, the opposite pattern of compliance is observed: The model shows a higher level of compliance with the human agent compared to the avatar or the computer agent. The difference in compliance across the three aids arises from the model's assessment of trust necessity for each of the automated aids. Trust necessity is assumed to be higher in interpersonal interactions as compared to human-machine (i.e., avatar and computer) interactions.

The one-counterpart-linear model produced a worse fit ($z = 3.34$, $p = 0.00$; see Figure F4 in the Appendix) to De Visser et al.'s [54] data ($r(10) = .79$, $p < 0.01$, $RMSD = .42$) than the multiple-counterparts-non-linear model ($r(10) = .99$, $p < 0.01$, $RMSD = .02$). The difference in the ability of the two models to account for the human data from De Visser et al. [54] stems from the trust decay parameter ($a$ in Equation (4)). The rate of compliance in the one-counterpart-linear model remains high over the course of the entire experiment. In contrast, the multiple-counterparts-non-linear model shows a steady decline in compliance as the agent's reliability decreases over the course of the experiment (Figure 7). The difference in the dynamics of compliance between the two models results from three different effects of the trust decay parameter.

First, before the model interacts with any of the decision aids, it undergoes a training period where it interacts with a neutral agent. During the initial training period the model begins to learn the number sequence of the experiment and to assess its trust in itself (i.e., self-trust). The model's self-trust is based on the accuracy of its initial answers. Due to the fact that the model is initially learning the number sequence, the model's initial performance is poor. The model's poor performance leads to a low self-trust evaluation. This low initial self-trust occurs in both the one-counterpart-linear model and the multiple-counterparts-non-linear model, but the trust decay parameter of the multiple-counterparts-non-linear model causes an important difference. In the multiple-counterparts-non-linear model, self-trust decreases according to a power law, eventually plateauing at a particular value. When the model's performance on the task increases, the model's self-trust increases rapidly. In contrast, the old one-counterpart-linear model's self-trust decreases as a linear function and does not plateau. When its performance increases, its self-trust has a very low starting point and takes a long time to reach a higher level. As a result, the old model maintains a higher rate of compliance with the aid's advice, even when the aid's accuracy decreases.

Second, due to the fact that each agent starts the experiment with 100% accuracy, the model's trust in each agent increases to a high level at the start of the experiment. However, as with the model's self-trust, the trust decay parameter makes a large difference in how quickly the trust in the agents can change. The old model's trust in each of the agents increases linearly and is not bound by an asymptote. As a result, it takes longer for the old model to be sensitive to the aid's decrease in performance and change its compliance behavior. By contrast, in the new multiple-counterparts-non-linear model, the trust accumulator plateaus, which allows the model to be more sensitive to the aid's decrease in performance.

Finally, the third reason for the high rate of compliance in the old model is the lack of decay for its trust necessity parameter. This parameter stays at relatively high values and intervenes

[13]The empirical data were reported in De Visser et al. (2016) as least mean squared estimates from a linear mixed effects model. We followed the same procedure for the model simulation data. This is one of the reasons we get such a good fit: The linear mixed effects model linearized the data for both the human and the model datasets. Most likely, the fit would not have been as good had we attempted to fit the raw data. The second reason for obtaining a good fit is the model-fitting process itself; that is, finding the best set of parameter values to fit this particular dataset. Thus, the model describes the data rather than predicts them. Potentially, this model could generate equally good fits for a range of empirical outcomes.

Table 2. The AICc Metrics for the One-counterpart-linear Model and the Multiple-counterpart-non-linear Model for Five Separate Datasets

| | AICc | |
|---|---|---|
| Dataset | One-Counterpart Linear Model | Multiple-Counterpart Non-Linear Model |
| Juvina et al. [13] | −3,828.32 | −4,228.25 |
| Collins et al. [15] | −35,453.18 | −35,451.18 |
| Lount et al. [37] | −254.16 | −259.52 |
| De Melo et al. [44] | −329.22 | −370.22 |
| De Visser et al. [54] | −65.13 | −136.74 |

Table 3. The Overall AICc and ΔAICc Metrics for the One-counterpart-linear Model and the Multiple-counterpart-non-linear Model Across Five Datasets

| Model | Overall AICc | ΔAICc |
|---|---|---|
| Single-counterpart-linear model | −39,180.63 | 470.64 |
| Multiple-counterpart-non-linear model | −39,651.27 | 0 |

to repair trust more often than in the new model. By allowing trust necessity to decay, the new model maintains its ability to detect consistent drops in the aid's trustworthiness. The combination of these three factors leads the old model to over-comply with the decision aids over the course of the experiment, fail to calibrate its trust to the aid's decreasing levels of trustworthiness, and thus fail to fit the human data. Additionally, these results reveal the importance of the new model's trust decay parameter in dynamic situations where a trustee's behavior changes rapidly. When the trust accumulator plateaus, the model can quickly recalibrate its trust to the recent evidence of trustworthiness.

### 3.6 Model Comparison

So far, we have shown that the multiple-counterpart-non-linear model can account for important trust phenomena (trust asymmetry, trust resilience, black-hat/white-hat effect, etc.) better than the one-counterpart-linear model. However, the multiple-counterpart-non-linear model has more free parameters. In this section, the two models are compared based on the Akaike Information Criterion (AIC) metric [57] that takes into account both the model fit and the number of free parameters. Given similar fit, AIC favors models that have fewer parameters. We actually used the AICc metric, which is a small sample correction of the standard AIC metric. The necessity of the parameters added to the multiple-counterparts-non-linear model can be assessed by comparing its AICc against the one-counterpart-linear model's AICc.

Two sets of AICc values were calculated for both the one-counterpart-linear model and the multiple-counterpart-non-linear model. The first set of AICc values was calculated based on a model fit to each of the five datasets. Table 2 shows that the multiple-counterpart-non-linear model had a lower AICc compared to the one-counterpart-linear model for four of the five datasets. The second set of AICc values was calculated for each model across all the datasets. This overall AICc takes into account a model fit and the number of free parameters across a range of datasets [60]. Table 3 shows an overall lower AICc value for the multiple-counterpart-non-linear model compared to the single-counterpart-linear model. Finally, a ΔAICc [58] was calculated for each of the two overall AICc metrics by taking the difference between each models' AICc and the minimum

AICc of the set. Computing the ΔAICc value aids in comparing the AICcs between models. The lowest AICc and best model will have a ΔAICc of zero. In addition, the larger the ΔAICc between the two models, the less empirical support there is for a model when compared to another. According to the guidelines of Burnahm and Anderson [58], a ΔAICc larger than 10 suggests little empirical support for a model compared to another model. Examining the ΔAICc values, we can conclude that there is little support for the single-counterpart-linear model when compared to the multiple-counterpart-non-linear model.

In summary, a comparison of the two models based on AICc reveals that, despite the additional parameters, the multiple-counterpart-non-linear model is better than the single-counterpart-linear model. The added parameters (trust discounting and trait trust deviation) are particularly useful to generalize the model beyond Prisoner's Dilemma and other $2 \times 2$ strategic interaction games, particularly to interacting with multiple trustees in sequence (De Melo dataset and cognitive ability simulation) and dealing with increasingly untrustworthy advice from humans or machines (De Visser's [54] dataset).

## 4 GENERAL DISCUSSION AND CONCLUSION

We presented an updated version of a cognitive model of learned trust that integrates several seemingly unrelated categories of findings from the literature and thus makes headway toward a unified theory of learned trust. The model cumulates learning from its history of interactions with multiple other models (trait trust or trust propensity) and learning from its current interaction (state trust). The integration between trait and state trust that we propose here has the potential to unify research directions that are currently somewhat disjointed: the psychological literature emphasizing trait trust (i.e., trust propensity or general trust) (e.g., Reference [61]) and the economics and computation literatures emphasizing state trust (e.g., Reference [62]). It also suggests computational solutions to the *cold-start problem*; that is, the inability of a model to generate predictions before having experienced several rounds of interaction with a counterpart [30].

The model assumption that trait trust is learned (at least in part) from the lifelong history of interaction with multiple counterparts was justified by our empirical finding that trait trust changes slightly over the course of a study in a specific condition in which evidence of untrustworthiness is frequently observed ([15]; see Figure 1 above). This finding may be seen as inconsistent with the literature that describes trust propensity as a relatively stable personality trait [7, 63–66]. To assess the robustness of our finding, we have replicated this finding in a large study with 627 participants. In addition, careful reading of a larger body of literature suggests that there is no inconsistency between our findings and the literature on trust propensity. Typically, the stability of trust propensity as a personality trait is assessed through test-retest reliability, which essentially is a correlation between trust propensity scores observed at different time points [66]. We also observed high test-retest reliability for trait trust in our studies.

However, when researchers measure and report changes in means and distributions of trust propensity scores over time, differences are typically observed. Players' investments in one-shot trust games, thought to represent their trait trust, have been found to increase with the age of players [67] and vary based on geographic region [68]. Twenge, Campell, and Carter [69] examined longitudinal survey research and found that trait trust has declined in the United States since the 1970s and was moderated by age, birth year, and income. Additionally, differences in trait trust have been noted between cultures, being on average higher in the West than in the East [70]. Yamagishi and Yamagishi [71] have proposed that the difference in average trait trust is moderated by particular cultural practices (e.g., reliance on established relationships) and that these practices lead to trait trust being learned at different rates across different cultures. More recently, Baer, Matta, Kim, Welsh, and Garud [72] have shown that trait trust is affected by particular social contexts.

The difference between our findings and the ones mentioned above is one of timescale: We observed a change in trait trust during a typical laboratory session (45mins), whereas the changes mentioned in the literature happened over the course of years. We conjecture that this difference can be reconciled if we consider the relative frequencies of the different types of evidence (i.e., trustworthiness and untrustworthiness). In real-world settings, the ratio between the two different types of evidence might change very slowly. For example, if the ratio between negative and positive evidence of trustworthiness is 2/3 over a time span of 10 years, the net effect of this mix of evidence may be zero (because of the asymmetry between negative and positive evidence; see Figure 2). This may give the appearance of relative stability of trust propensity over time. In other words, to the extent that the mix of evidence of trustworthiness and untrustworthiness is relatively constant, trust propensity appears as a stable trait.

In our studies, we manipulated the ratio between evidence of trustworthiness and untrustworthiness: In the high trustworthiness condition, the ratio was disproportionately in favor of evidence of trustworthiness; whereas in the low trustworthiness condition, the ratio was disproportionately in favor of evidence of untrustworthiness (see Reference [15] for details). The change in trait trust was observed only in the low trustworthiness condition (see Figure 1). Arguably, this condition occurs very rarely in real-world settings, because of cultural practices and institutions that discourage untrustworthiness. For example, in real-world settings, trustors would discontinue relationships with trustees who produce repeated evidence of untrustworthiness. Our future modeling work will focus on better specifying the relationship between the dynamics of trait trust (trust propensity) in past interactions and the perception of trustworthiness in the current interaction.

The current version of the model defines the trust learning equation as a power law (and consequently, the trust accumulator as a leaky accumulator), making it consistent with other more general learning mechanisms of the cognitive architecture [5]. This gives rise to interesting model behaviors that match empirical effects observed in human studies such as trust asymmetry, the higher impact of early trust breaches, and potentially other effects that were not explored here, such as "surprise" [73]. In addition, the model predicts that trust decays in the absence of evidence of trustworthiness or untrustworthiness. Although computational models of trust tend to include some form of trust decay, we do not know of any empirical evidence for this effect in the trust literature. Our future empirical work will aim to test this novel model prediction. If this prediction is corroborated by empirical data, it will strengthen the intuitive wisdom that trust can only be maintained if the flow of information between the two protagonists is uninterrupted. Suggestive support for this prediction comes from research on virtual, geographically distributed teams: Jarvenpaa and Leidner [74] found that trust in such conditions was "swift" but very fragile; regular and timely communication feedback was critical for building trust and commitment in distributed teams.

Trust theorists seem to agree that trust is based on "good reasons" [7, 63, 75], but all the good reasons seem to belong to the categories that were referred to here as perceived evidence of trustworthiness (PET in Equation (2)) and trait trust (TT in Equation (4), trustor's propensity in Mayer et al. [7]). Our model suggests that perceived trust necessity and the trustor's cognitive ability could also be considered among the good reasons (or antecedents) of trust. Figure 8 shows a modified and simplified version of Mayer et al.'s [7] model of trust. The factors that pertain to the trustor are presented separately from those that pertain to the trustee to show that Mayer's model emphasizes factors pertaining to the trustee; that is, components of trustworthiness. This emphasis on trustworthiness as the most important antecedent of trust seems to permeate the entire trust literature [26]. Our work suggests that factors pertaining to the trustor (i.e., perceived trust necessity and cognitive ability, colored in red in Figure 8) are also important. They may even be critical

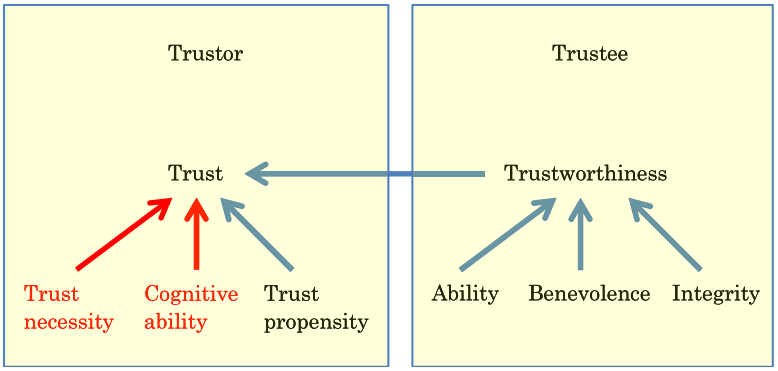


Fig. 8. Antecedents of trust.

in certain situations. For example, a correct assessment of trust necessity is critical to overcome cases of escalating mutual distrust or to prevent hasty disposal of imperfect but useful technology, and the ability of the trustor to detect the actual trustworthiness behind ambiguous or misleading cues can be essential for appropriate trust calibration.

Although there are many computational models that deal with trust and related phenomena (e.g., References [30, 76–78]), our attempt presented here is unique because it tries to integrate trust in a cognitive architecture and a unified theory of cognition. Given that trust has an important cognitive component [25], it would make sense to try to make use of the established theory on cognition and its computational instantiation. We developed a computational model of trust that used the ACT-R cognitive architecture's established cognitive mechanisms such as instance-based learning to develop a dynamic representation of the counterpart's behavior and procedural learning to develop best-response strategies.

However, using the existing ACT-R mechanisms proved to be insufficient to account for the complex dynamics of the strategic interaction between counterparts. We added a new trust learning mechanism to the model that specifies how the counterparts learn to trust each other based on observed evidence of trustworthiness. This trust learning mechanism interacts with the existing ACT-R learning mechanisms by influencing what strategies are learned and when strategy shifts occur. We added the trust learning mechanism to the model as a proof of concept, but suggested that a similar (perhaps more general) mechanism should eventually (after thorough validation) be added to the architecture. Even though our work presented here is somewhat limited in scope, it can be characterized as a step toward a general architectural mechanism that specifies how agents learn about both the task at hand and other agents with which they interact in complex cooperative or competitive environments.

We also made some headway toward integrating multiple task paradigms. The model focus was initially on $2 \times 2$ strategic games (i.e., "Prisoner's Dilemma" type of games). However, we hope that the new trust learning mechanism will apply more widely to interactive decision-making. We have shown that the trust mechanism applies to another interactive decision-making phenomenon: advice taking in a sequence-learning task [54]. The core idea that we hope will generalize to other interactive decision-making situations is that players form beliefs about their counterparts, and these beliefs guide strategy choice and strategy shifts.

A cautionary word is necessary here to avoid unrealistic expectations: We are still far from a truly unified, all-encompassing theory of learned trust. However, following Newell's [3] advice, we believe it is important to work toward developing unified theories. As the term "toward" in the title of this article suggests, we are outlining here a desideratum rather than an accomplishment.

To remain general, our model had to be left largely underspecified. The trust learning mechanism we proposed here did not deal with the knowledge level. We only specified how trust could be learned, but the actual knowledge that was learned came from perceptual input and interaction with other agents. For example, we only suggested that the perceived evidence of trustworthiness or untrustworthiness accumulated in a certain way. The evidence of trustworthiness varied in the different instantiations of our model. For example, in the PD and CG games, the payoff matrix suggests what may constitute evidence of trustworthiness (or untrustworthiness) and its magnitude. In De Melo's task, facial expressions were translated into evidence of trustworthiness. In De Visser's task, it was the quality of the advice that was considered evidence of trustworthiness. All this is information that is perceived in the environment and becomes knowledge for the model. As for the values (magnitudes) of this evidence, they were estimated in the process of model fitting; that is, they were considered to be free parameters for the model. Translating skill and knowledge phenomena into free parameters was done to keep the model simple and focus on general principles and mechanisms. We also proposed a way to "interpret" the evidence about a trustee relative to the experience the trustor had with a previous trustee. To work in a different domain, the model will need to incorporate specifications of what counts as evidence of (un)trustworthiness, how the evidence is perceived and decoded, and so on. For example, the extent to which one might decrease their trust may be different when they encounter defection in Prisoner's Dilemma as compared to when they receive bad advice in a sequence-learning task.

This work is relevant to the area of human-autonomy teaming. The science and practice of human-machine interaction have departed from the traditional function allocation methods (who-does-what or men-are-better-at/machines-are-better-at; Fitts [79]) and is currently moving toward a *human-autonomy teaming* approach in which the focus is on how machines can become effective team players [80] and how humans and technology co-evolve [28]. Empirical and theoretical work on interpersonal and human-machine trust can inform design and evaluation of human-autonomy teaming. For example, De Visser, Pak, and Shaw [81] argue for developing autonomous systems that possess trust repair capabilities. The work presented here suggests that trust is more resilient (i.e., resistant to breaches) when the trustor perceives the relationship with the trustee as a strategic relationship (see Sections 2.1 and 4.6) and engages in costly and risky trust repair strategies (see Section 2 for a description of how models sometimes manage to escape the self-reinforcing cycle of mutual defection; see also Juvina et al. [13] for more details).

In conclusion, this article reports on the incremental progress we have made from a post hoc model of strategic interaction to a more general model that is able to make *a priori* predictions and account for seemingly unrelated results from the literature on interpersonal and human-machine interactions.

## APPENDIX A

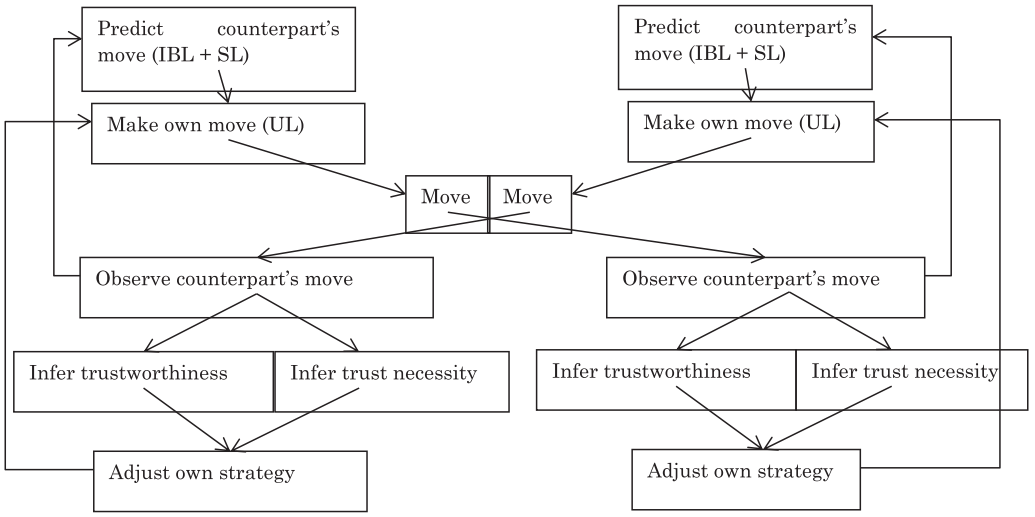


Fig. F1. Diagram of the cognitive model that explains learning in the Prisoner's Dilemma (PD) and Chicken Game (CG) and transfer of learning between the two games in both directions (PD-CG and CG-PD). Two models play simultaneously. Each model tries to predict the counterpart's move using instance-based learning (IBL) and sequence-learning (SL). Once they predict their counterpart's move, they make their own move based on what they have learned through utility learning (UL) to be the best move in a given context. Then players observe each other's moves and use this information to inform future predictions and infer the counterpart's trustworthiness and the necessity to develop (invest in) trust. Trustworthiness and trust necessity determinewhat reward function is used for learning a strategy to best respond to the counterpart's predicted move.

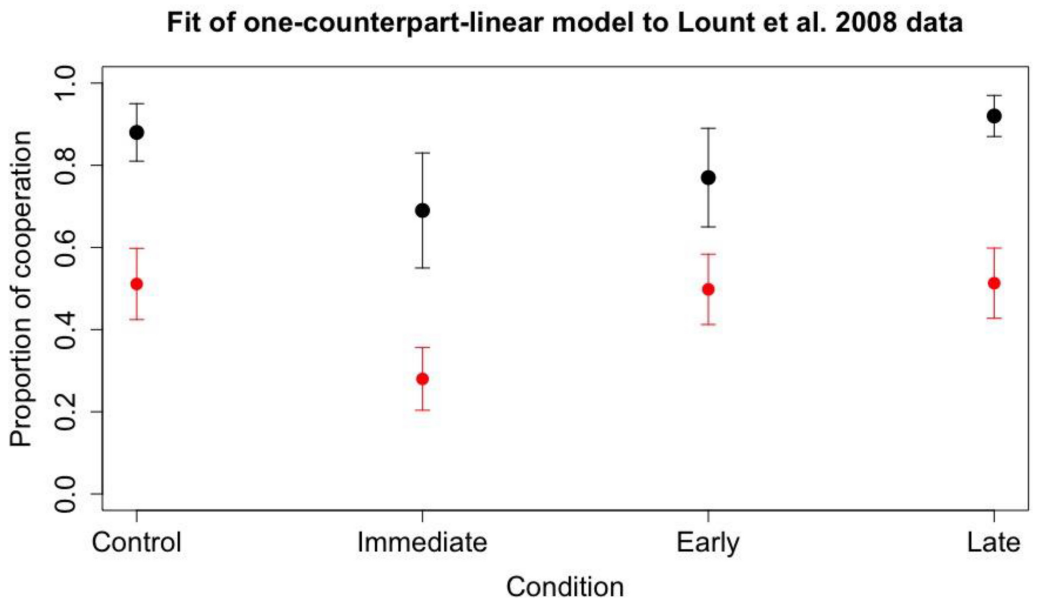


Fig. F2. Fit of the one-counterpart-linear model (red dots) to Lount et al. (2008) data (black dots). Error bars represent 95% confidence intervals.

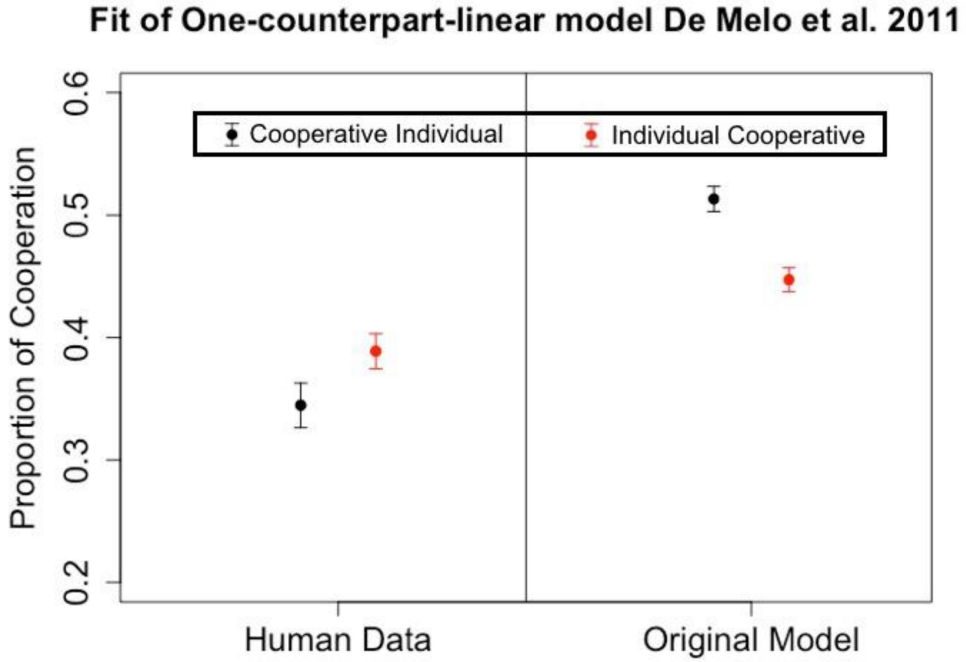


Fig. F3. Fit of the one-counterpart-linear model (right panel) to De Melo et al. (2011) data (left panel). Error bars represent 95% confidence intervals.

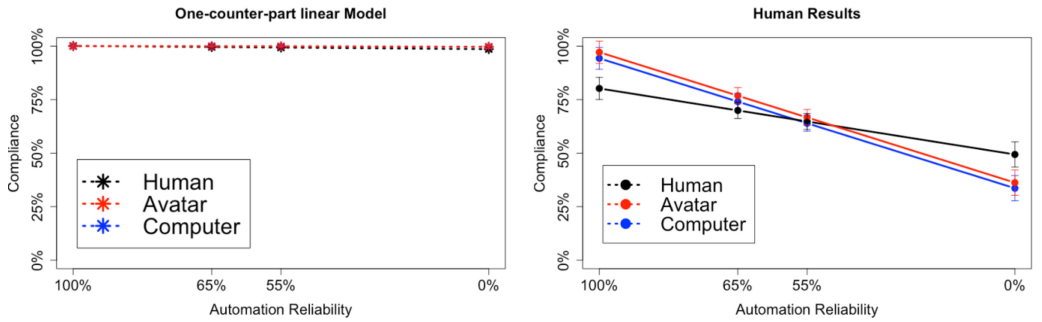


Fig. F4. The least mean square estimates (+/− standard error) for compliance with the three human (black line), avatar (red line), and computer (blue line) agents for both the one-counterpart-linear model (left-side plot) and human data (right-side plot).

Table T1. Payoff Matrices of Prisoner's Dilemma (PD) and Chicken Game (CG)

| PD | A | B | CG | A | B |
|---|---|---|---|---|---|
| A | −1,−1 | 10,−10 | A | −10,−10 | 10,−1 |
| B | −10,10 | 1,1 | B | −1,10 | 1,1 |

Table T2. Names, Acronyms, and Descriptions for the Parameters of the Multiple-counterparts-non-linear Model

| Parameter | Acronym | Description |
|---|---|---|
| Trait Trust | TT | Trust propensity or dispositional trust |
| State Trust | ST | The trust that develops during a particular interaction |
| Actual evidence of trustworthiness | AET | Actions, statements, or other indicators (e.g., facial expressions) of trustworthiness |
| Perceived evidence of trustworthiness | PET | The trustor's representation of the AET |
| Perceived evidence of trust necessity | PETN | The trustor's representation of the indicators of situations that require trust development |
| Trait trust deviation | TTD | Difference between the trait trust value at the end of the current interaction and the trait trust value at the beginning of this interaction |
| Cognitive ability | e | Accuracy (error) in assessing trustworthiness signals |
| Trust decay | a | Power exponent that determines how fast state trust decays over time ($a < 1$) |
| Perception bias | b | A parameter that scales how much PET is adjusted as a function of the trustor's previous experience with another trustee |
| State trust divider | k | A parameter that scales down the value of state trust before it accumulates into trait trust |

Table T3. The List of Parameters for the Trust and Trust-invest Accumulators Based on the Behavioral Outcomes that Occurred in De Melo et al.'s (2010) Study, that is, Whether Both Agents Cooperated (CC), Both Defected (DD), the Model Cooperated and the Agent Defected (CD), the Model Defected and the Agent Cooperated (DC), and the Emotion Displayed by the Confederate Agent

| | | Behavioral Effects on Trust | | Emotional Effects on Trust | |
|---|---|---|---|---|---|
| Outcome | Emotion | Trust Accumulator | Trust Invest Accumulator | Trust Accumulator | Trust Invest Accumulator |
| CC | Joy | 6 | NA | 1 | NA |
| CC | Neutral | 6 | NA | −3 | NA |
| CD | Shame | −7 | −1 | 1 | NA |
| CD | Joy | −7 | −1 | −6 | −2 |
| DC | Anger | 9 | NA | −6 | .5 |
| DC | Sadness | 9 | NA | −4 | NA |
| DD | Sadness | −1 | .18 | −3 | .32 |
| DD | Sadness/Neutral | −1 | .18 | −3 | .32 |

For example, if the outcome was mutual cooperation (CC) and the counterpart expressed joy, an increment of 1 will be added to the trust accumulator; if the outcome was unilateral cooperation (CD) and the counterpart expressed joy, a decrement of −6 will be added to the trust accumulator.

Table T4. The List of Initial Trait Trust (Trust Propensity) and Perceived Evidence of Trustworthiness (PET) Parameter Values for the Trust and Trust Invest Accumulators for De Visser et al.'s (2016) Study for Both the ACT-R Model (Model) and the Different Decision Agents (Human, Avatar, and Computer)

| Agent | Initial Trust | Initial Invest | Correct | Incorrect | Invest | Uninvest |
|---|---|---|---|---|---|---|
| Human | 5 | 5 | .75 | −3 | 1.5 | 0 |
| Avatar | 6 | 5 | .75 | −3 | 1.18 | 0 |
| Computer | 7.5 | 5 | .75 | −3 | 1.2 | 0 |
| Model | 0 | 5 | 1 | 0 | .18 | 0 |

## ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers of the ICCM 2016 conference for their feedback on the conference paper and the anonymous reviewers of ACM TiiS journal for feedback on two earlier drafts of this article. We also thank Kevin Gluck from AFRL and the members of the ASTECCA laboratory at WSU for the many insightful conversations we had on trust and cognitive modeling.

## REFERENCES

[1] S. L. Jarvenpaa, T. R. Shaw, and D. S. Staples. 2004. Toward contextualized theories of trust: The role of trust in global virtual teams. *Inform. Syst. Res.* 15, 3 (2004) 250–267.

[2] J. B. Lyons, C. K. Stokes, and T. R. Schneider. 2011. Predictors and outcomes of trust in teams. In *Trust in Military Teams*. N. A. Stanton, (Ed.) Ashgate Publishing Ltd.

[3] A. Newell. 1990. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.

[4] K. A. Hoff and M. Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Fact.* 57, 3 (2015), 407–434.

[5] J. R. Anderson. 2007. *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, New York.

[6] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Acad. Manag. Rev.* 23, 3 (1998), 393–404.

[7] R. C. Mayer, J. H. Davis, and F. D. Schoorman. 1995. An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 3 (1995), 709–734.
[8] F. D. Schoorman, R. C. Mayer, and J. H. Davis. 2007. An integrative model of organizational trust: Past, present, and future. *Acad. Manag. Rev.* 32, 2 (2007), 344–354.
[9] J. D. Lee and K. A. See. 2004. Trust in automation: Designing for appropriate reliance. *Hum. Fact.* 46, 1 (2004), 50–80.
[10] K. A. Hoff and M. Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Fact.* 57, 3 (2015), 407–434.
[11] K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Hum. Fact.* 58, 3 (2016), 377–400.
[12] K. Gluck. 2016. *Pace, Persistence, and Scale.* Talk at the twenty-third ACT-R postgraduate summer school. Lancaster, PA. Retrieved on Aug 2017 from http://act-r.psy.cmu.edu/?post_type=workshops&p=20531.
[13] I. Juvina, C. Lebiere, and C. Gonzalez. 2015. Modeling trust dynamics in strategic interaction. *J. Appl. Res. Mem. Cog.* 4, 3 (2015), 197–211.
[14] J. Berg, J. Dickhaut, and K. McCabe. 1995. Trust, reciprocity, and social history. *Games Econ Behav.* 10, 1 (1995), 122–142.
[15] M. G. Collins, I. Juvina, and K. A. Gluck. 2016. Cognitive model of trust dynamics predicts human behavior within and between two games of strategic interaction with computerized confederate agents. *Front. Psychol.* 7:49.
[16] C. F. Camerer. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*; Princeton University Press: Princeton, NJ.
[17] A. Rapoport, M. J. Guyer, and D. G. Gordon. 1976. *The 2×2 Game.* The University of Michigan Press, Ann Arbor, MI.
[18] I. Juvina, M. Saleem, J. M. Martin, C. Gonzalez, and C. Lebiere. 2013. Reciprocal trust mediates deep transfer of learning between games of strategic interaction. *Organ. Behav. Hum. Decis. Process.* 120, 206–215.
[19] G. Biele, J. Rieskamp, and R. Gonzalez. 2009. Computational models for the combination of advice and individual learning. *Cog. Sci.* 33, 2 (2009), 206–242.
[20] I. Yaniv and E. Kleinberger. 2000. Advice taking in decision making: Egocentric discounting and reputation formation. *Organiz. Behav. Hum. Dec. Proc.* 83, 2 (2000), 260–281.
[21] P. L. Harris and K. Corriveau. 2011. Young children's selective trust in informants. *Phil. Trans. R. Soc. B*, 366, 1179–1187.
[22] C. Gonzalez, F. J. Lerch, and C. Lebiere. 2003. Instance-based learning in real-time dynamic decision making. *Cog. Sci.* 27, 4 (2003), 591–635.
[23] H. He, J. Boyd-Graber, K. Kwok, and H. Daume 'III. 2016. Opponent modeling in deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning.*
[24] I. Juvina, O. Larue, and A. Hough. 2018. Modeling valuation and core affect in a cognitive architecture: The impact of arousal and valence on memory and decision-making. *Cog. Syst. Res.* 48: 4–24.
[25] C. Castelfranchi and R. Falcone. 2010. Trust theory: A socio-cognitive and computational model. *Wiley Series in Agent Technology.* John Wiley & Sons Ltd., Chichester, UK.
[26] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. de Visser, and R. Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Fact.* 53, 5 (2011), 517–527.
[27] J. Grudin. 1988. Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. In *Proceedings of the Association for Computing Machinery in Conference on Computer-supported Cooperative Work.* 85–93.
[28] M. S. Ackerman. 2000. The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Hum.-Comput. Interact.* 15 (2000), 179–203.
[29] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. 2006. An integrated trust and reputation model for open multi-agent systems. *Autonom. Agents Multi-Agent Syst.* 13 (2006), 119–154.
[30] Q. V. Dang and C. L. Ignat. 2016. Computational trust model for repeated trust games. In *Proceedings of the International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom'16).*
[31] D. H. McKnight, L. L. Cummings, and N. L. Chervany. 1998. Initial trust formation in new organizational relationships. *Acad. Manag. Rev.* 23, 3 (1998), 473–490.
[32] K. T. Dirks and D. L. Ferrin. 2001. The role of trust in organizational settings. *Organiz. Sci.* 12 (2001), 450–467.
[33] R. M. Kramer. 1999. Trust and distrust in organizations: Emerging perspectives, enduring questions. *Ann. Rev. Psych.* 50 (1999), 569–598.
[34] P. Slovic. 1993. Perceived risk, trust, and democracy: A systems perspective. *Risk Anal.* 13 (1993) 675–682.
[35] J. J. Skowronski and D. E. Carlston. 1989. Negativity and extremity biases in impression formation: A review of explanations. *Psych. Bull.* 105 (1989), 131–142.
[36] I. Yaniv and E. Kleinberger. 2000. Advice taking in decision making: Egocentric discounting and reputation formation. *Organiz. Behav. Hum. Decis. Proc.* 83, 2 (2000), 260–281.

[37] R. B. Lount, C. B. Zhong, N. Sivanathan, and J. K. Murnighan. 2008. Getting off on the wrong foot: The timing of a breach and the restoration of trust. *Person. Soc. Psych. Bull.* 34, 12 (2008), 1601–1612.
[38] N. H. Anderson and A. A. Barrios. 1961. Primacy effects in personality impression formation. *J. Abnorm. Soc. Psych.* 63, 2 (1961), 346–350.
[39] J. DeCoster and H. M. Claypool. 2004. A meta-analysis of priming effects on impression formation supporting a general model of informational biases. *Person. Soc. Psych. Rev.* 8, 1 (2004), 2–27.
[40] J. B. Lyons, C. K. Stokes, and T. R. Schneider. 2011. Predictors and outcomes of trust in teams. N.A. Stanton (Ed.), *Trust in Military Teams*. Ashgate Publishing Ltd.
[41] P. Sturgis, S. Read, and N. Allum. 2010. Does intelligence foster generalized trust? An empirical test using the UK birth cohort studies. *Intelligence* 38, 1 (2010), 45–54.
[42] T. Yamagishi, M. Kikuchi, and M. Kosugi. 1999. Trust, gullibility, and social intelligence. *Asian J. Soc. Psych.* 2, 1 (1999), 145–161.
[43] A. Newell and P. S. Rosenbloom. 1981. Mechanisms of skill acquisition and the law of practice. In *Cognitive Skills and Their Acquisition*, J. R. Anderson (Ed.), (1–55). Erlbaum, Hillsdale, NJ.
[44] C. M. De Melo, P. Carnevale, and J. Gratch. 2011. The impact of emotion displays in embodied agents on emergence of cooperation with people. *Presence* 20, 5 (2011), 449–465.
[45] J. Hilty and P. Carnevale. 1993. Black-hat/white-hat strategy in bilateral negotiation. *Organiz. Behav. Hum. Decis. Proc.* 55, 3 (1993), 444–469.
[46] H. Helson. 1964. *Adaptation-level Theory*. Harper & Row, New York.
[47] M. G. Collins, I. Juvina, G. Douglas, and K. A. Gluck. 2015. Predicting trust dynamics and transfer of learning in games of strategic interaction as a function of a player's strategy and level of trustworthiness. In *Proceedings of the International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRiMS'15)*.
[48] R. J. Lewicki, D. J. McAllister, and R. J. Bies. 1998. Trust and distrust: New relationships and realities. *Acad. Manag. Rev.* 23, 3 (1998), 438–458.
[49] S. B. Sitkin and N. L. Roth. 1993. Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. *Organiz. Sci.* 4, 3 (1993), 367–392.
[50] N. Ambady and R. Rosenthal. 1993. Half a minute: Predicting teacher evaluations from thin slices of behavior and physical attractiveness. *J. Person. Soc. Psych.* 64 (1993), 431–441.
[51] J. W. Strachan, A. J. Kirkham, L. R. Manssuer, and S. P. Tipper. 2016. Incidental learning of trust: Examining the role of emotion and visuomotor fluency. *J. Exper. Psych.: Learn., Mem., Cog.* 42, 11 (2016), 1759.
[52] C. Heintz, M. Karabegovi, and A. Molnar. 2016. The co-evolution of honesty and strategic vigilance. *Front. Psych.* 7:1503.
[53] W. Thompson and S. Kaufmann. 2010. Signaling games with partially observable actions as a model of conversational grounding. In *Proceedings of the 3rd AAAI Conference on Interactive Decision Theory and Game Theory*. AAAI Press.
[54] E. J. de Visser, S. S. Monfort, R. McKendrick, M. A. B. Smith, P. E. McKnight, F. Krueger, and R. Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *J. Exper. Psych.: Appl.* 22, 3 (2016), 331–349.
[55] B. Van Buren, S. Uddenberg, and B. J. Scholl. 2016. The automaticity of perceiving animacy: Goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychon. Bull. Rev.* 23, 3 (2016), 797–802.
[56] I. Juvina, M. G. Collins, O. Larue, and C. de Melo. 2016. Toward a unified theory of learned trust. In *Proceedings of the 14th International Conference on Cognitive Modeling*. D. Reitter and F. E. Ritter (Eds.). (188–193).
[57] H. Akaike. 1973. Information theory as an extension of the maximum likelihood principle. In *Proceedings of the International Symposium on Information Theory*, B. N. Petrov and F. Csaki (Eds.). 267–281.
[58] K. P. Burnham and D. R. Anderson. 2002. Model selection and multimodel inference: A practical information-theoretic approach. Springer-Verlag, New York.
[59] M. A. Navakatikyan. 2007. A model for residence time in concurrent variable interval performance. *J. Exper. Anal. Behav.* 87, 1 (2007), 121–141.
[60] M. A. Navakatikyan. 2007. A model for residence time in concurrent variable interval performance. *J. Exper. Anal. Behav.* 87, 1 (2007), 121–141.
[61] T. Yamagishi, S. Akutsu, K. Cho, Y. Inoue, Y. Li, and Y. Matsumoto. 2015. Two-component model of general trust: Predicting behavioral trust from attitudinal trust. *Soc. Cog.* 33, 5 (2015), 436.
[62] A. Das and M. M. Islam. 2012. SecuredTrust: A dynamic trust computation model for secured communication in multiagent systems. *IEEE Trans. Depend. Sec. Comput.* 9, 2 (2012), 261–274.
[63] J. A. Colquitt, B. A. Scott, and J. A. LePine. 2007. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *J. Appl. Psych.* 92, 4 (2007), 909–927.
[64] H. Gill, K. Boies, J. E. Finegan, and J. McNally. 2005. Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust. *J. Bus. Psych.* 19, 3 (2005), 287–302.

[65] M. K. Lee and E. Turban. 2001. A trust model for consumer internet shopping. *Int. J. Electron. Comm.* 6, 1 (2001), 75–91.
[66] J. B. Rotter. 1971. Generalized expectancies for interpersonal trust. *Amer. Psych.* 26, 5 (1971), 443.
[67] M. Sutter and M. G. Kocher. 2007. Trust and trustworthiness across different age groups. *Games Econ. Behav.* 59, 2 (2007), 364–382.
[68] N. D. Johnson and A. A. Mislin. 2011. Trust games: A meta-analysis. *J. Econ. Psych.* 32, 5 (2011), 865–889.
[69] J. M. Twenge, W. K. Campbell, and N. T. Carter. 2014. Declines in trust in others and confidence in institutions among American adults and late adolescents 1972–2012. *Psych. Sci.* 25, 10 (2014), 1914–1923.
[70] E. M. Krockow, M. Takezawa, B. D. Pulford, A. M. Colman, S. Smithers, T. Kita, and Y. Nakawake. 2018. Commitment-enhancing tools in centipede games: Evidencing European-Japanese differences in trust and cooperation. *Judg. Decis. Making* 13, 1 (2018), 61.
[71] T. Yamagishi and M. Yamagishi. 1994. Trust and commitment in the United States and Japan. *Motivat. Emot.* 18, 2 (1994), 129–166.
[72] M. D. Baer, F. K. Matta, J. K. Kim, D. T. Welsh, and N. Garud. 2018. It's not you, it's them: Social influences on trust propensity and trust dynamics. *Personn. Psych.* 71, 3 (2018), 423–455.
[73] F. Paglieri, C. Castelfranchi, C. da Costa Pereira, R. Falcone, A. Tettamanzi, and S. Villata. 2014. Trusting the messenger because of the message: Feedback dynamics from information quality to source evaluation. *Comput. Math. Organiz. Theory* 20: 176–194.
[74] S. L. Jarvenpaa and D. E. Leidner. 1999. Communication and trust in global virtual teams. *Organiz. Sci.* 10, 6 (1999), 791–815.
[75] J. D. Lewis and A. Weigert. 1985. Trust as a social reality. *Soc. Forces* 63 (1985), 967–985.
[76] J. Sabater and C. Sierra. 2005. Review on computational trust and reputation models. *Artific. Intell. Rev.* 24, 1 (2005), 33–60.
[77] J. Gao and J. D. Lee. 2006. Extending the decision field theory to model operators' reliance on automation in supervisory control situations. *IEEE Trans. Syst., Man, Cyber.—Part A: Syst. Hum.* 36, 5 (2006), 943–959.
[78] S. Farrell and S. Lewandowsky. 2000. A connectionist model of complacency and adaptive recovery under automation. *J. Exper. Psych.: Learn., Mem. Cog.* 26, 2 (2000), 395.
[79] P. M. Fitts. (Ed.) (1951). Human engineering for an effective air navigation and traffic control system. National Research Council, Washington, DC.
[80] S. W. A. Dekker and D. D. Woods. 2002. MABA-MABA or abracadabra? Progress on human–automation coordination. *Cog., Technol. Work* 4:240–244.
[81] E. J. de Visser, R. Pak, and T. H. Shaw. 2018. From "automation" to "autonomy": The importance of trust repair in human–machine interaction. *Ergonomics* 61, 10 (2018), 1409–1427.

Received November 2016; revised November 2018; accepted April 2019