In D. Reitter & F. E. Ritter (Eds.), Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016). University Park, PA: Penn State.

Toward a Unified Theory of Learned Trust

Ion Juvina[1] (ion.juvina@wright.edu), Michael Collins[1,2] (collins.283@wright.edu), Othalia Larue[1] (othalia.larue@wright.edu), & Celso de Melo[3] (demelo@usc.edu)
[1]Department of Psychology, Wright State University, 3640 Colonel Glenn Hwy., Dayton, OH 45435 USA
[2]Air Force Research Laboratory, Dayton, OH 45435 USA
[3]Institute for Creating Technologies, University of Southern California, 12015 Waterfront Dr., Playa Vista, CA 90094 USA

Abstract

A proposal for a unified theory of learned trust is presented. A number of limitations of a published computational cognitive model of learned trust are discussed. A solution is proposed to overcome these limitations and expand the model's scope of applicability. The revised model integrates several seemingly unrelated categories of findings from the literature and makes unintuitive predictions for future studies. The implications of the model for the advancement of the theory on trust are discussed.

Keywords: trust; trustworthiness; trust propensity; learned trust; computational cognitive model; unified theories

Introduction and Background

Newell (1990) called for unified theories of cognition specified computationally as cognitive architectures. 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 trust research, particularly on what has been referred to as learned trust (Hoff & Bashir, 2015), and attempt to integrate it in the ACT-R cognitive architecture (Anderson, 2007). Although the field of trust already comprises an impressive volume of empirical findings, micro-theories, meta-analyses, and integrative accounts (e.g., Rousseau, Sitkin, Burt, & Camerer, 1998; Mayer, Davis, & Schoorman, 1995; Schoorman, Mayer, & Davis, 2007; Lee & See, 2004; Hoff & Bashir, 2015; Schaefer, Chen, Szalma, & Hancock, 2016), 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.

The starting point for the effort reported here is a published model of learned trust (Juvina, Lebiere, & Gonzalez, 2015; referred to as "the published model" in the remainder of the paper). In the next section we briefly review the key features of the published model and discuss its main strengths and limitations. Then, we devote another section to a revised model (referred to as "the revised model" in the remainder of the paper) that is intended to overcome the limitations of the published model. Subsequently, we show that the revised model can account for a number of results from the trust literature. In the last section, we discuss possible ways to further improve the revised 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.

Critique of the Published Model

The published model (Juvina et al., 2015[1]) 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). The model is not hardwired to play a particular game; it learns to play any 2X2 game (Rapoport, Guyer, & Gordon, 1976) 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 (Juvina, Saleem, Martin, Gonzalez, & Lebiere, 2013) 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. Essentially, this trust mechanism allows models to learn not only about the task at hand but also about other models with which they interact. 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 (Biele, Rieskamp, & Gonzalez, 2009; Yaniv & Kleinberger, 2000; Harris & Corriveau, 2011). The published model uses instance-based learning (Gonzalez, Lerch, & Lebiere, 2003) for opponent modeling and reinforcement learning for action selection. In addition, the reward changes 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 mechanism consists of a "trust accumulator" that represents the perceived trustworthiness of the other model and a "trust-invest accumulator" that represents the perceived necessity to develop trust – a characteristic of the situation. For example, when the two models find themselves in a

[1] Model code available at: http://psych-scholar.wright.edu/astecca/software

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. Humans are able to identify and (sometimes) overcome those situations.

The two accumulators (trust and trust-invest) are used to determine the dynamics of the reward structure. 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 continually occurring during the game, which decreases trust in the other model 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. This can lead to mutual cooperation and development of trust or models may relapse into a mutual destructive choice. When the trust accumulator is positive, a player tries to maximize joint payoff and avoid exploitation. Thus, the model switches between three reward functions depending on the dynamics of trust between the two players.

Strengths of the published model

The main contribution of the published model was to show that trust learning interacts with task specific learning to account for a range of learning effects in the human data. 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 agreement with the literature on trust, the published model's trust is learned as a function of perceived trustworthiness (Mayer et al., 1995; Hoff & Bashir, 2015). In addition, the published model suggests 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 published model showed that both perceived trustworthiness and perceived trust necessity are important antecedents of trust formation (Collins, Juvina, & Gluck, 2016).

Limitations of the published model

Most of the limitations of the published model stem from the fact the model was initially not intended to be comprehensive model of learned trust. Instead, the model had to learn trust in order to account for transfer of learning effects observed in the human data. The published 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 (McKnight, Cummings, & Chervany, 1998) and this initial propensity to trust determines to some extent the trust that develops during the interaction (Berg, Dickhaut, & McCabe, 1995; Dirks & Ferrin, 2001). In addition, trust propensity may be (at least in part) the result of learning that occurred prior to the current interaction (Collins et al., 2016) and a comprehensive model of learned trust should not ignore prior learning, particularly because prior learning may interact with current learning. This aspect was not relevant in the published 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 published 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 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 published model,

$$ST_t = ST_{t-1} + PET_t \qquad (1)$$

where ST_t is state trust at time t, ST_{t-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 published 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 (Slovic, 1993). Trust learners exhibit the same negativity bias that is described in the impression formation literature (Skowronski & Carlston, 1989; Yaniv & Kleinberger, 2000), that is, unfavorable information tends to be more influential than favorable information. In addition, early evidence has a stronger impact on trust formation than late evidence (Lount, Zhong, Sivanathan, & Murnighan, 2008). In general, learning equations tend to be power functions (Newell & Rosenbloom, 1981; Anderson, 2007) and it would be surprising if trust learning were an exception.

Another limitation of the published model is that it assumes that all trustors are able to assess equally well trustworthiness and trust necessity. However, a trustor's cognitive ability to assess a trustee's trustworthiness has been proposed to be an important antecedent of trust (Lyons, Stokes, & Schneider, 2011; Sturgis, Read, & Allum, 2010; Yamagishi, Kikuchi, & Kosugi, 1999). In general, cognitive ability is an important predictor of learning, thus it is not surprising that it is also related to learned trust.

The Revised Model

Before introducing our revisions to the published model, we specify the terminology used in this model. *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, thus, is a function of the *perceived evidence of trustworthiness*

and *perceived evidence of trust necessity*. In our view, *learned trust* includes both trait and state trust; trait trust 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 (Berg, Dickhaut, & McCabe, 1995). 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 (Collins et al., 2016). *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 the interaction. The trustor's *cognitive ability* is indicated by the accuracy of the trustor's judgments of trustworthiness and trust necessity.

The revision[2] of the published model consists of replacing the linear function that was used to update the trustor's state trust with the following power function,

$$ST_t = ST_{t-1}^{a} + PET_t - b * TTD \qquad (2)$$

where ST_t is state trust at time t, ST_{t-1} is state trust at time t-1, a is a constant power exponent with a value less than 1 (a<1), 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 revised 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 continuous value of state trust can be 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.

[2] Model code available at: http://psych-scholar.wright.edu/astecca/software

The power exponent (a) 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 1 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 (Slovic, 1993).

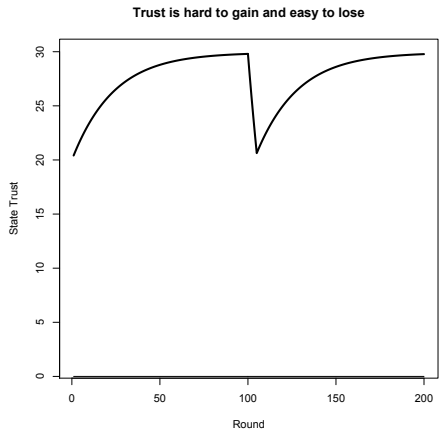

Figure 1: A hypothetical case illustrating how state trust changes over the course of 200 rounds of interaction with another player. The trustor perceives evidence of trustworthiness for the first 100 rounds, then evidence of untrustworthiness for 5 rounds, and again evidence of trustworthiness for 95 rounds.

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 (2011) 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 (Hilty & Carnevale, 1993). We implemented the explanation of the black-hat/white-hat effect that is based on the concepts of adaptation and comparison level

(Helson, 1964). 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 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 terms) of the reference point caused by the previous experience. The reference point is the trustor's trait trust. For example, if the trustor's previous experience with an untrustworthy trustee caused a large shift 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 revised model, only the trust learning mechanism has been revised, all the other mechanisms of the published model (learning to anticipate the opponent's move and to select the best move in each context, see Juvina et al., 2015) have been left unchanged.

Model Validation

We expect that the revised model is able to generalize to a wide range of empirical phenomena while maintaining the ability of the published model to explain the learning and transfer of learning effects from the original dataset.

Learning and transfer of learning effects in Prisoner's Dilemma and Chicken Game

Juvina et al. (2013) recruited 120 participants to play Prisoner's Dilemma and Chicken Game 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 published model was fit in its entirety to this dataset by tweaking 11 free parameters (see Table 4 in Juvina et al., 2015). With regard to the revised model, only the six free parameters associated with the trust mechanism were refit to the human data reported in Juvina et al. (2013). Four of the six parameters are associated with the "trust accumulator" that represents the perceived trustworthiness of the other player and the other two are associated with the "trust-invest accumulator" that represents the perceived necessity to develop trust. 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 six parameters (i.e., the parameter with the lowest absolute value for each accumulator) were kept at their values from the published model, thus, allowing only four model parameters to fluctuate. The fit procedure maximized the correlation (r) and minimized the root mean squared deviation (RMSD) between the model data and the human data[3].

Table 1 shows the best fitting parameter values for the revised model and the published model. They did not change dramatically; as a matter of fact, one of them did not change at all, even though it was allowed to vary freely. Thus, only three parameters have been readjusted in the revised model. These parameters were held constant for all but one of the simulations reported below. They were readjusted for Lount et al. (2008) data because a very different payoff matrix was used in that study.

Table 1. The best fitting parameter values for the revised model and the published model for each of the four game outcomes, mutual cooperation (CC), unilateral cooperation (CD), unilateral defection (DC), and mutual defection (DD). An asterisk (*) indicates that a particular value was held constant during the model fitting procedure.

Outcome	Published model		Revised 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*

The fit of the revised model to the human data (r = .90, $RMSD$ = .07) was slightly (but not significantly) better than the fit of the published model (r = .89, $RMSD$ = .09). The revised model also exhibited the same transfer of learning effects observed in the human data.

Collins et al. (2016) 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 & trustworthiness) was manipulated to result in 16 different experimental conditions. The published model (Juvina et al., 2015) was used to generate *a priori* predictions for Collins et al. (2016) study. The predictions were published before the data were collected (Collins, Juvina, Douglas, & Gluck, 2015). A majority of the model predictions across all of the sixteen experimental conditions was confirmed and the trust mechanism was proven to be a necessary component of the published model (see Collins et al., 2016, for details). Here we test the revised model against the dataset from Collins et al. (2016) without any parameter tweaking. The data includes round-by-round proportions for five outcomes in

[3] 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.

16 conditions. The revised model accounts for the human data slightly (but not significantly) better (r = .68, *RMSD* = .33) than the published model (r = .64, *RMSD* = .33).

Unified account of trust and distrust effects

It has been proposed that trust and distrust are different constructs (Lewicki, McAllister, & Bies, 1998; Sitkin & Roth, 1993). Here we suggest that the different dynamics of trust and distrust can be modeled by a single equation. In the previous section we showed how equation 2 produces trust asymmetry (Slovic, 1993; see Figure 1). 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. (2008) found. Lount et al. (2008) conducted two experiments in which participants played an iterated game of Prisoner's Dilemma for 30 rounds. Participants were assigned to 1 of 4 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 occurred 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 frequency of cooperation during the last ten rounds of the game as compared to the late breach.

Our revised 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 = 0.99, *RMSD* = 0.33). One possible explanation for the large 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 revised model is able to explain Lount et al.'s findings based on the 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 (Ambady & Rosenthal 1993).

Black-hat/white-hat effect

De Melo, Carnevale, and Gratch (2011) had participants play Prisoner's Dilemma with two different computerized confederate agents (cooperative & 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 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 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, trustworthiness could only be 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, etc. (e.g., Strachan, Kirkham, Manssuer, & Tipper, 2016). De Melo et al. (2011) 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, as defined in the previous section. 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. 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), which allowed the model to fit the human data (r = .86, *RMSD* = .11) and reproduce the black-hat/white-hat effect.

Conclusion and Future Work

We presented 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), learning from its current interaction (state trust), and (sometimes) incidental learning from facial expressions. The model predicts that trust decays toward distrust in the absence of evidence of trustworthiness or untrustworthiness. Our future empirical work will aim to test this novel model prediction. Our future modeling work will focus on better specifying the relationship between the dynamics of trait trust in past interactions and the perception of trustworthiness in the current interaction.

Acknowledgments

This work was supported by The Air Force Office of Scientific Research grant number FA9550-14-1-0206 to IJ and Oak Ridge Institute for Science and Education to MC.

References

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of behavior and physical attractiveness. *Journal of Personality and Social Psychology*, 64, 431-441.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122-142.

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science*, *33*(2), 206–242.

Collins, M.G., Juvina, I., & Gluck, K.A. (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.

Collins, M.G., Juvina, I., Douglas, G., & Gluck, K.A. (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*. Paper presented at Behavior Representation in Modeling and Simulation (BRiMS) conference.

De Melo, C.M., Carnevale, P, & Gratch, J. (2011). The Impact of Emotion Displays in Embodied Agents on Emergence of Cooperation with People. *Presence*, *20*(5), 449–465.

Dirks, K.T., & Ferrin, D.L. (2001). The role of trust in organizational settings. *Organizational Science, 12*, 450-467.

Gonzalez, C., Lerch, F. J., & Lebiere, C. (2003). Instance-based learning in real-time dynamic decision making. *Cognitive Science* 27 (4), 591-635.

Harris, J. (2008). Maximizing the utility of MindModeling@ Home resources. Presented at Integrated Design and Process Technology Conference.

Harris, P.L., and Corriveau, K. (2011) Young children's selective trust in informants. *Phil. Trans. R. Soc. B*, 366, 1179–1187.

Helson, H. (1964). *Adaptation-level theory*. New York: Harper & Row.

Hilty, J., & Carnevale, P. (1993). Black-hat/white-hat strategy in bilateral negotiation. Organizational Behavior and Human Decision Processes, 55(3), 444–469.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors 57*(3), 407-434.

Hommel, B. & Colzato, L.S. (2015) Interpersonal trust: an event-based account. *Front. Psychol*. 6:1399.

Juvina, I., Lebiere, C., & Gonzalez, C. (2015). Modeling trust dynamics in strategic interaction. *Journal of applied research in memory and cognition. 4*(3): 197-211.

Juvina, I., Saleem, M., Martin, J. M., Gonzalez, C., and Lebiere, C. (2013). Reciprocal trust mediates deep transfer of learning between games of strategic interaction. *Organ. Behav. Hum. Decis. Process.* 120, 206–215.

Lee, J. D., See, K. A. (2004). Trust in automation: Designing for Appropriate Reliance. *Human Factors* 46(1): 50-80.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of management Review*, 23(3), 438-458.

Lount, R. B., Zhong, C. B., Sivanathan, N., & Murnighan, J. K. (2008). Getting off on the wrong foot: The timing of a breach and the restoration of trust. *Personality and Social Psychology Bulletin*, *34*(12), 1601-1612.

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

McKnight, D.H., Cummings, L.L., & Chervany, N.L. (1998). Initial trust formation in new organizational relationships. *The Academy of Management Review* 23(3), 473-490.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Rapoport, A., Guyer, M. J., & Gordon, D. G. (1976). *The 2×2 game*. Ann Arbor, MI: The University of Michigan Press.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of management review*, *23*(3), 393-404.

Schaefer, K.E., Chen, J.Y.C., Szalma, J.L., & Hancock, P.A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation. *Human Factors*.

Sitkin, S. B., & Roth, N. L. (1993). Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. *Organization science*, 4(3), 367-392.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, **105,** 131–142.

Slovic, P. (1993). Perceived risk, trust, and democracy: A systems perspective. *Risk Analysis*, **13,** 675–682.

Strachan, J., Kirkham, A., Manssuer, L., & Tipper, S. P. (2016). Incidental learning of trust: Examining the role of emotion and visuomotor fluency. *Journal of Experimental Psychology: Learning, Memory & Cognition*.

Sturgis, P., Read, S., & Allum, N. (2010). Does intelligence foster generalized trust? An empirical test using the UK birth cohort studies. *Intelligence*, 38(1), 45-54.

Yamagishi, T., Kikuchi, M., & Kosugi, M. (1999). Trust, gullibility, and social intelligence. *Asian Journal of Social Psychology*, 2(1), 145-161.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281.