Research Article

Yosef S. Razin* and Karen M. Feigh

Committing to interdependence: Implications from game theory for human–robot trust

https://doi.org/10.1515/pjbr-2021-0031
received March 31, 2021; accepted September 18, 2021

Abstract: Human–robot interaction (HRI) and game theory have developed distinct theories of trust for over three decades in relative isolation from one another. HRI has focused on the underlying dimensions, layers, correlates, and antecedents of trust models, while game theory has concentrated on the psychology and strategies behind singular trust decisions. Both fields have grappled to understand over-trust and trust calibration, as well as how to measure trust expectations, risk, and vulnerability. This article presents initial steps in closing the gap between these fields. By using insights and experimental findings from interdependence theory and social psychology, this work starts by analyzing a large game theory competition data set to demonstrate that the strongest predictors for a wide variety of human–human trust interactions are the interdependence-derived variables for commitment and trust that we have developed. It then presents a second study with human subject results for more realistic trust scenarios, involving both human–human and human–machine trust. In both the competition data and our experimental data, we demonstrate that the interdependence metrics better capture social "overtrust" than either rational or normative psychological reasoning, as proposed by game theory. This work further explores how interdependence theory – with its focus on commitment, coercion, and cooperation – addresses many of the proposed underlying constructs and antecedents within human–robot trust, shedding new light on key similarities and differences that arise when robots replace humans in trust interactions.

Keywords: human–machine trust, human–robot interaction, design and human factors, acceptability and trust, modelling and simulating humans

*** Corresponding author: Yosef S. Razin,** School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0250, United States of America, e-mail: yrazin@gatech.edu
Karen M. Feigh: School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0250, United States of America, e-mail: karen.feigh@gatech.edu

1 Introduction

Human–robot interaction (HRI) and game theory have had little interaction in the development of their respective theories of trust and collaboration. Game theory has long utilized a singular concept of trust, defined as the payoff structure of typically one-shot interactions. It thereby attempted to figure out not what trust looked like behaviorally, but what psychological motivations led to its fulfillment [1,2]. Conversely, HRI focused primarily on deconstructing the idea of trust, its underlying dimensions, antecedents, and corollaries [3,4]. This attempt to understand trust more holistically, as a system of attitudes, expectations, decisions, and behaviors, led to many insights at the cost of construct proliferation. Beyond this conceptual rift between the disciplines, HRI often viewed the eponymous "games" of game theory as reductive toy problems that did not translate well into the field; this was despite HRI's own trust research being often limited to simulations or 2D interfaces with accompanying posttask questionnaires.[1] These divisions and approaches can be traced back to the origins of these parallel paths of exploring trust.

Trust in HRI has been strongly influenced by social psychology, human factors, and teamwork, whereas trust in game theory has been more strongly influenced by philosophy, economics, and political science. While both fields have drawn liberally from others and have independently developed their own unique insights, they have yet to cross-germinate fruitfully. This article will begin to bridge that gap, starting a new conversation on what HRI (and human–machine interaction more generally) can learn from the primarily human–human interactions studied in game theory and looking at how the underlying constructs of trust from HRI relate to game-theoretic trust.

In this article, we will first present an overview of trust as it is been approached by HRI and game theory, as well as the current tenuous connections between the two fields. We will then give a brief introduction to interdependence

1 For more on this divide and its general implications, see ref. [5].

theory, focusing on how it contributes to our understanding of trust by deriving a testable definition of trust games and developing equations for commitment and a new trust index (TI). We then present two experiments: the first testing interdependence theory-based algorithms on a game theory competition data set followed by our own human subject testing, showing the power of interdependence theory over previously proposed approaches to trust prediction. Finally, we will discuss the power as well as the limits of our approach, especially with regard to human–human vs human–machine trust.

1.1 Trust and control in HRI

Early work focusing on trust in automation mainly grew out of social psychology [6–11]. It was also influenced by sociology, primarily Luhmann's *Trust and Power* [12] and Barber's *The Logic and Limits of Trust* [13] proved to be hugely influential, firmly establishing trust as multidimensional, and explicating its relation with complexity and communication. Luhmann's influence can still be identified in two major disputes within HRI trust, as far as the roles of norms and control [14]. Briefly, does the modern world and its complex technologies, such as robotics, with their inherent uncertainties and risks, preclude familiarity and norm-based trust? Furthermore, are systems of control replacements for trust in such a world instead of an integral part of trust itself? While Luhmann answered both of these in the affirmative, these questions are currently coming to the fore of debates in HRI trust. How we answer these questions will have profound implications, especially for how HRI trust is conceived in contrast to human–human trust.

While generic trust had been historically captured by a single item on survey instruments, once trust was understood as multidimensional and distinct from confidence and familiarity, early researchers of trust in automation started trying to capture these new dimensions [15,16]. Eventually, some of these axes converged around a slightly shortened form of Mayer's seminal definition of trust, as follows:

> The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor.

This is rooted in the constructs of "Ability," "Integrity," and "Benevolence" [17]. Later works created more fully fleshed out trust models, antecedents, and co-factors, incorporating concepts such as interface design, understandability, transparency, ease of use, effectiveness, accessibility, and familiarity [4,18–22]. Further developments included expanding Mayer's "Benevolence" into more general expectations concerning affective trust (e.g., cooperation vs competitiveness) [3,21,23] and re-casting Mayer's "Integrity" into structural assurance [18,20]. Finally, in order to discern whether expectations of trustworthiness truly transformed into trust, considerations of intended and actual use were considered (based on ref. [18]).

Mayer's original definition had included the final clause, "irrespective of the ability to monitor or control that party," in which Luhmann's dichotomy of external control in opposition to trust can be discerned. This clause has often been dropped in later HRI trust definitions (e.g., ref. [24]); however, more recent works have expanded and explicated what such control means. Castelfranchi and Falcone [23] have argued that while narrow, "strict" trust is antagonistic to control, a broader notion of trust that includes confidence in social systems and norms, such as laws, contracts, and ethics, actually completes and compliments trust, increasing it above what strict trust alone would suggest. Similarly, Law and Scheutz [25] understand trust as two distinct categories: performance based and relation based. Performance-based trust is relying on competence *sans* monitoring ("strict"), whereas relation-based trust expands beyond the specific situation. This latter category hews closely to Luhmann's confidence in social systems of trust as well as Castelfranchi and Falcone's concept of "broader" trust. A similar treatment of this "new," abstracted, social/normative "category" of trust is termed "structural trust" and found to be a well-defined, independent, and internally consistent dimension of HMI trust by McKnight et al. [20], Gefen et al. [18], and Malle and Ullman [26].

Thus, while the topology of trust is still contended, a consensus has emerged regarding the role of broader control via trust in social structures such as ethics and laws in HRI. This may be somewhat surprising, applying expectations of norms to robots. However, norms are a crucial part of familiarity and expectation building [12], even for non-human agents. One clear example is autonomous vehicles. For instance, Razin and Feigh [27] demonstrated that while drivers rated human and autonomous vehicles differently based on perceived performance-based trust, they had the same expectations and perceptions of both agents when it came to social expectations around driving. In other words, they believed that self-driving cars would follow the same laws and norms as humans on the road. Similar results on the importance of norms to trust around household interactions have also been identified [28,29]. The human trustor may also abstractly place structural

trust in a company that sells them robotic products, the engineers that design those products, and the laws that regulate the products and businesses involved [18,20]. Structural trust devolves upon an entire social network of actors of which the actual robot is only one node, albeit the fact at which the direct trust interaction occurs.

Beyond control, cooperation in the form of teamwork is receiving increased attention in HRI [3,30] as is coercion (both through incentives and sanctions), especially in the form of inappropriate compliance and reliance [31,32]. While this work focuses on these larger questions of control, cooperation, coercion, and commitment, we will return to discussing how performance-based and affective-based trust fit into the interdependence model (Section 5.2). In the following, we will also differentiate "rational" trust from performance-based "strict" trust given the specific meaning of rationality in game theory; indeed, the bulk of this work is aimed at explaining that both performance-based and affective-based trust are based on rational beliefs.

1.2 Interdependence theory: deconstructing control

In order to further explore the relationship of trust with control, cooperation, and coercion, we propose reviving an off-shoot of game theory, proposed by Thibaut and Kelley over a half-century ago [33]. Their interdependence theory was reintroduced into HRI trust by Wagner [34] and Robinette [35] and re-framed classical games by breaking down the relative levels of control afforded to each agent. The theory of interdependence is also broader than classical game theory as it does not assume rationality or even the attempt to maximize monetary or even concrete outcomes. Thus, it considers symbolic outcomes, such as the reputational payoff of following social norms or the pleasure of fulfilling another's needs [36]. Thibaut and Kelley also recognized that even within a single interaction, the "game" is not limited to simply the structure of the outcomes prescribed by the situation, but that actors may further process and mentally transform such situations by framing them in various temporal or social ways. These include attempting to maximize the joint outcomes of all actors or minimizing the difference between some outcomes to ensure equity. They also explored transformations instantiated by making certain externally motivated behavioral commitments (playing by the rules, turn-taking), preempting partner's choices, and accounting for future interactions [33]. These transformations act in well-characterized and prescribed ways upon outcome matrices similar to those used in game theory, such as the 2 × 2 matrix in Figure 1. Many of these steps to expand game theory would be retread starting in the late 1980s within mainstream game theory research by Geanakoplos et al. [37], when they founded psychological game theory. However, the focus on the decomposition of games by control "modes" remains a unique and crucial contribution of interdependence theory alone.

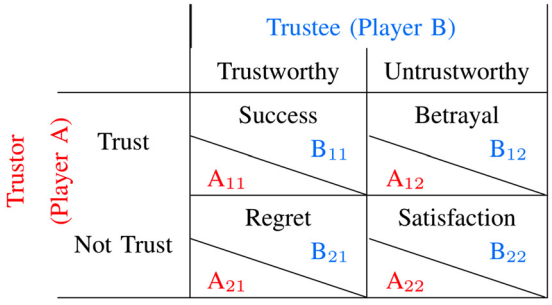

Figure 1: Payoff matrix for the trustor (red) and trustee (blue) in a trust–trustworthiness interaction. Regret here is specific to not trusting/being trusted when trust would have been fulfilled and is distinct from any emotion linked to betrayal.

1.3 Trust in game theory

Before turning to the method of deconstructing control within a game or interaction, it is worthwhile understanding how trust is even framed in game theory, which is so often focused on competitive scenarios. Unlike trust in HRI, in game theory, trust is not seen as multidimensional and there is little debate over its definition. What defines a trust game in game theory, first and foremost, is the payoff structure (Figure 1). The oft-cited requirements (e.g., refs [1,38,39]) for a trust game according to game theory are as follows:

(1) Exposure: The trustor is risking more by betrayal than if they do not trust ($A_{12} < \{A_{21}, A_{22}\}$).
(2) Improvement: The trustor stands to gain more by fulfilled trust than by not trusting ($A_{11} > \{A_{21}, A_{22}\}$).
(3) Temptation: The trustee at least is tempted to betray trust when proffered ($B_{12} > B_{11}$).
(4) Mutual Gain:[2] That the payoff for being trustworthy when trusted is higher than not being trustworthy at all ($B_{11} > \{B_{21}, B_{22}\}$).

all while assuming that $A_{21} = A_{22}$ and $B_{21} = B_{22}$.

2 Not universally accepted.

A very similar, though expanded, set of trust conditions for the *trustor* alone was independently derived by Wagner [34] as follows:

(1) The act of trust must occur in the face of uncertainty; the trustee cannot act before the trustor.
(2) Only if the trustor chooses to trust does the trustee's action matter, such that the payoff for successful trust is higher than the potential loss if the trustee is untrustworthy. Quantitatively, this means the difference between the payoffs for successful vs unsuccessful trust must be at some minimum (ε_1) dependent, reflecting some risk ($A_{11} - A_{12} > \varepsilon_1$) (Exposure)
(3) The trustor's payoffs for not trusting are independent of the trustee, such that the amount unrisked by not trusting is bounded by ε_2, such that $|A_{21} - A_{22}| < \varepsilon_2$.
(4) Successful trust is the highest outcome and betrayal is the lowest, with the non-trusting options bound by these two levels, such that $A_{11} > \{A_{21}, A_{22}\} > A_{12}$ (Improvement).
(5) The trustor must believe that the probability of the trustee acting trustworthy is greater than some trust threshold ($p^A(\mathrm{TW}) > C$).

Note that the inequalities presented by game theory only define a trust game and not how the binary decision to trust or act trustworthy is made. However, Wagner attempts to provide, at least abstractly, such a criterion in (5). One such solution for calculating that decision threshold could be the game's mixed Nash equilibrium. However, game theory generally has suggested that the "rational" solution here is the subgame perfect equilibrium (SPE), which unfortunately and unrealistically predicts that trust should rarely occur, as being untrustworthy is the trustee's weakly dominant *rational* strategy. This is clearly not how trust plays out in the real world, where trust is frequently given and fulfilled. Thus, behavioral and psychological insights are sought to fill this gap.

It is a well-known phenomenon that, even in the "toy problems" presented in game theory experiments, people choose to trust and be trustworthy more than they seemingly "should" based on payoffs and risk aversion alone [2]. Theories as to why people "over-trust" range the gamut from long-term reputation keeping, conformity to moral norms, expecting and reciprocating kindness, guilt, and inequality aversion to name but a few, with varying supporting findings in the game-theoretic trust literature [2,38–41]. Note that these theories *all* fall under the "broader" notions of structural or relation-based trust, as discussed in Section 1.1.

The gap between game theory's and HRI's approaches to trust and how they are articulated, framed, motivated, and modeled is wide indeed. In fact, the only common foundation to both approaches is that trust occurs when one is made vulnerable by exposure to risk and that it is premised on a "particular action of importance to the trustor" [17]. Game theory focuses on the binary trust decision, and, more often than not, HRI focuses on the continuously valued belief in or expectation of trust and trustworthiness. Furthermore, the very design of the game-theoretic implementation removes questions of capability, much less understandability and familiarity, and completely violates the evaluation of trust under situational normality, which are all stressed in HRI. On the other hand, one could argue that by removing these correlates, game theory examines a "purer" form of trust that goes beyond instrumentality [2]. This "pure" trust though is also strongly understood to be rooted in exclusive elements of human–human interaction, focusing on equity, kindness, and moral normativity. This approach completely disregards how trust operates when a non-human is involved (beyond anthropomorphism) or even when trust is strongly premised on performance as opposed to relational concerns. Furthermore, HRI trust rarely concerns itself with the trustee's alternative payoffs, with the exception of ref. [34], directly challenging the temptation criteria of trust games as formally defined by game theory.

1.4 A declaration of interdependence

Interdependence theory proposes decomposing games by determining which actor has power over which part of the total payoff structure [33]. We will illustrate each aspect of this powerful approach and its insights through the following game, as shown in Figure 2.

Imagine a human must decide whether to trust an autonomous vehicle or switch to manual mode. The payoff structure here is not only the real costs or payoffs but also incorporates emotional, reputational, and other psychological utilities. If the human does trust and the autonomous vehicle works perfectly, the human is reasonably happy especially given the cost of the car (50). If they do not trust the autonomous mode, even though they believe it generally works, and get into an accident by driving manually, they will kick themselves for not trusting and regret it (−50). However, if the human trusts the autonomous vehicle and it fails, it is catastrophic and they may never use the car again (−100). Finally, if they decide not to trust it and then hear that it actually does not work, they will feel satisfied with their justified choice (30). Note that due to psychological factors such as regret and satisfaction, typically $A_{21} \neq A_{22}$.

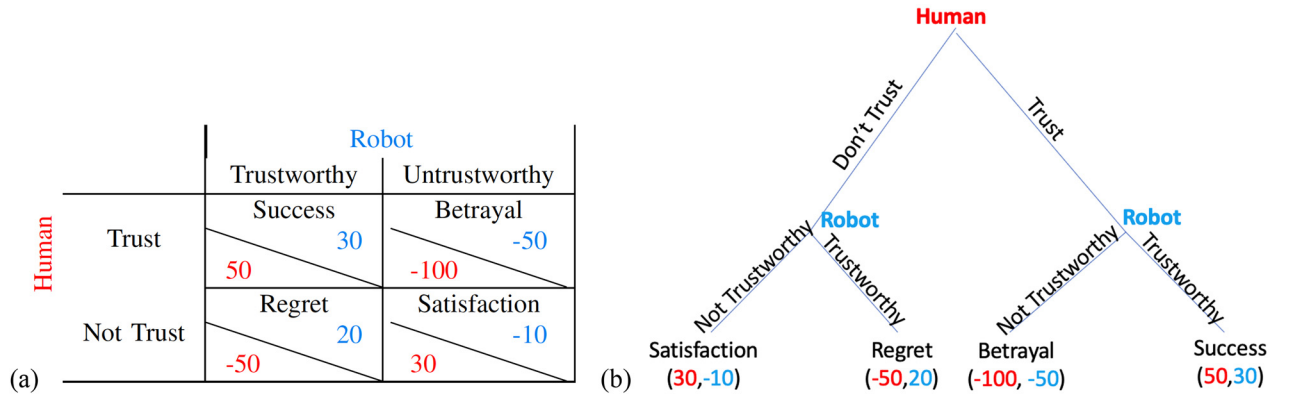

Figure 2: An illustrative example of the payoffs in a human–robot trust interaction. The human trustor's payoffs are in red and the robot's are in blue, with the game represented in two forms. (a) Example normal-form representation of the trust game, best suited for enabling calculations. (b) Extensive-form representation of the game, highlighting the sequential nature of the game, with the trustor making the first "move."

The payoffs for the robot can be seen as the utility either for it as an agent directly or for its owners, manufacturers, designers, or insurers. The robot anticipates being rewarded and used more (or perhaps its manufacturer anticipates increased share prices) for properly fulfilling trust (30) but penalized even more if it betrays the human's trust, as failure may result in discontinued use, not to speak of reputational and commercial loss (−50). While often such games assume that the trustee receives or losses nothing by being not trusted ($B_{21} = B_{22} = 0$) [2,34], that is clearly not the case, as can be illustrated. Generally, not being trusted will hurt the brand (−10) but as people get into accidents driving manually while the robot actually is demonstrably a more reliable driver than humans, then the safer autonomous vehicle will appear a better option and people will seek it out (let us say a net gain of 20).

Traditional game theory would predict that the two agents will act rationally and play the SPE. What this means is that by working backward through the example in Figure 2, the autonomous car's payoff for regret dominates satisfaction ($20 > -10$) and successfully fulfilling trust dominates betrayal ($30 > -50$). The human trustor is then left deciding between successful trust vs regret ($50 > -50$), and thus in this case, the SPE predicts that the human will indeed successfully trust. However, in cases when satisfaction dominates regret, such as when the trustee is seen as less as a tool and more as a potential teammate, the SPE indicates that one should not trust. In practice, the SPE seems to account for approximately 60–80% of trustor's decision to trust in human–human interaction [1,2,40].

Note that payoffs in game theory are known to be invariant under positive affine transformation, and thus, it does not matter if we multiply all the payoffs of the human by 1,000 or add 50 to each of those of robot. It also makes it tricky (if not impossible) to compare payoffs between the agents. However, normalizing all payoffs by each player's most extreme outcome can prove useful for understanding interdependence, as will be shown shortly.

Interdependence theory suggests that we can understand the interaction better by deconstructing the payoffs in terms of three types of control, those of each individual as well as that which arises from cooperation. Reflexive or actor control (RC) is how much unilateral power the actor has over their own outcomes, i.e. the expected difference between their choosing one action over the other. For the trustor (Player A), it is the average difference between the row-sums and for the trustee (Player B) this is transposed as the average difference between column-sums, such that

$$\begin{aligned} \mathrm{RC}_A &= 0.5((A_{11} + A_{12}) - (A_{21} + A_{22})) \\ \mathrm{RC}_B &= 0.5((B_{11} + B_{21}) - (B_{12} + B_{22})) \end{aligned} \tag{1}$$

In the human–robot game illustrated in Figure 2, the normalized reflexive control is $\mathrm{RC}_A = -0.15$ for the human trustor (Player A) and $\mathrm{RC}_B = 0.7$ for the car trustee (Player B). Thus, along this component of the payoff, the human is weakly inclined to choose not to trust and the autonomous vehicle (and its manufacturer) has a much stronger incentive to prove trustworthy.

Fate or partner control (FC) is how much unilateral power each actor has over the other's outcomes, i.e. the expected difference in one actor's outcomes when the other chooses between their actions. For Actor A, it is the average difference between their payoff's column-sums, and again this is transposed for Actor B:

Table 1: Summary of Interdependence control modes from the human-robot interaction example based on normalized payoffs

Control Mode	Human (A)	Car (B)
Reflexive Control (RC)	-.15	0.7
Fate Control (FC)	0.35	0.1
Bilateral Control (BC)	1.15	0.9

$$\begin{aligned} \mathrm{FC}_A &= 0.5((A_{11} + A_{21}) - (A_{12} + A_{22})) \\ \mathrm{FC}_B &= 0.5((B_{11} + B_{12}) - (B_{21} + B_{22})) \end{aligned} \tag{2}$$

To reiterate, FC_A is Actor A's estimation of Actor B's unilateral power over A's own outcomes. Per our example, $\mathrm{FC}_A = 0.35$ and $\mathrm{FC}_B = 0.1$, which means that the vehicle's trustworthiness has a much stronger impact on the human, than the human's choice to trust the car ($\mathrm{RC}_A = -0.15$), while the driver/consumer provides a mild positive car/manufacturer for the vehicle to be trustworthy.

Finally, bilateral or joint control (BC) is how much one actor's choice further facilitates or inhibits the other's outcomes. This set of weights is fully contingent on the partner's choice and, thus, is the result of coordination or its lack thereof. It is calculated for both partners as the average of the difference of the sums of the diagonal outcomes.

$$\mathrm{BC}_X = 0.5((X_{11} + X_{22}) - (X_{12} + X_{21})) \tag{3}$$

where X can be either A or B. Again in our interaction example, $\mathrm{BC}_A = 1.15$ and $\mathrm{BC}_B = 0.9$. Both the human and the car have a strong incentive to coordinate, fulfilling trust when trusted or not trusting the untrustworthy. Here control over payoffs via coordination is significantly stronger than any via unilateral control (Table 1).

If the signs of BC_A and BC_B are the same, as in the example, they are said to *correspond*, signaling that both actors share a preference for coordinated behavior. If the sign of BC matches that of RC or FC, they are said to be *concordant*, and if not, they are *discordant*. *Concordance* (*discordance*) is a measure of reinforcement (interference) between one mode of control and another. As the autonomous vehicle's interdependence weights in our example are all positive, FC_B, RC_B, and BC_B are all concordant, whereas for the human, FC_A and BC_A are concordant, but RC_A is discordant with both, signifying that the human is being coerced (in this case through incentivization). Both BC_A and BC_B are positive and thus correspond, indicating that the human and car share a preference for coordination.

In this way, interdependence theory and its associated weights can be used for a variety of analyses that illuminate many aspects of trust. This is clearly illustrated when we translate the game-theoretic trust conditions of Exposure and Improvement (or equivalently Wagner's trust conditions) into interdependence terms[3] as follows:

$$\begin{aligned} \mathrm{FC}_A > 0 \quad \mathrm{FC}_A > |\mathrm{RC}_A| \\ \mathrm{BC}_A > 0 \quad \mathrm{BC}_A > |\mathrm{RC}_A| \end{aligned} \tag{4}$$

This transformation yields the following interpretation:

> The expected additional gains of trusting that the trustee controls for the trustor, both unilaterally and through coordination, must be positive and greater in magnitude than the trustor's power over their own payoffs.

Note how this accords with our commonsense understanding of trust: if the trustor can accomplish the goal for themselves, there is no need for trust. Furthermore, the trustee exerts control over the trustor's success – and this arises from a mixture of coordination and magnanimity. This does not, however, imply altruism, as these conditions say nothing concerning the payoffs for the trustee. If we accept game theory's conditions for the trustee, Temptation can be expressed as follows:

$$\begin{aligned} \mathrm{FC}_B > \mathrm{BC}_B \\ \mathrm{FC}_B > \mathrm{RC}_B \end{aligned} \tag{5}$$

If the Mutual Gain condition is also accepted,

$$\begin{aligned} \mathrm{FC}_B > |\mathrm{BC}_B| \\ \mathrm{FC}_B > |\mathrm{RC}_B| \end{aligned} \tag{6}$$

Recall that often in game theory, trust games are designed such that $B_{21} = B_{22}$ & $A_{21} = A_{22}$, which leads to

$$\mathrm{FC}_A = \mathrm{BC}_A \tag{7}$$

$$\mathrm{FC}_B = \mathrm{BC}_B \tag{8}$$

however, neither of these two equivalence conditions are held to be actual requirements of trust games.

Note that the Temptation condition implies a somewhat cynical approach. Under it, a trust game is not simple when the trustor might consider trust as a viable strategy but when they would do so at the same time as when the trustee is tempted to betray them! From interdependence theory, we see that this essentially means that the only time they will act trustworthy in such a scenario is when the trustor's unilateral control provides an overwhelming incentive. While this may make for a "good" game, is it true that trust can only be said to occur in the presence of temptation/when it is coerced?

Leaving that question to be addressed later, observe how the interdependence analysis has allowed us to go

3 For proofs, see Appendix A.2: Theorems 1–4.

beyond the basics of the game-theoretic conditions, highlighting the cooperative aspect, as well as the power inequality between the two agents [42].

There is another gain to note from this new framing of the trust conditions. As mentioned earlier, payoffs in game theory cannot be compared between agents because they are invariant under positive affine transformations. These trust conditions contain the additional benefit of permitting normalized ranges for the interdependence weights, such that $\mathrm{RC}_A = (-1, 1)$, $\mathrm{FC}_A = (0, 1)$, $\mathrm{BC}_A = (0, 2)$, $\mathrm{RC}_B = (-2, 1)$, $\mathrm{FC}_B = (-2, 2)$, and $\mathrm{BC}_B = (-2, 1)$. Thus, various trustors and their valuations can be fruitfully compared, and likewise for trustees.

Finally, we recognize that in this section, we have introduced a number of key terms and acronyms and we will be introducing more in the following sections. Thus, for ease of reference, a glossary of terms is provided in Appendix A.1.

2 From interdependence weights to measures to trust

Armed with these re-framed theoretical constraints, it is time to forge a new path to show how trust is actually decided upon within such interactions. The initial starting points offered by game theory would be the SPE and mixed Nash equilibrium. As previously mentioned, the SPE has proven insufficient at capturing actual trust behavior for the trustor, leading to what appears to be over-trust. One solution may be the Nash equilibrium, which broadens our view from a discrete decision to the continuous domain of probabilities.

The Nash equilibrium is the point of indifference between the trustee's actions given the trustor's payoffs and *vice versa*. The probability of the trustee acting trustworthy (τ_B), which yields a Nash equilibrium if the trustee is to use a mixed strategy, is derived[4] as follows:

$$\begin{aligned} &\tau_B A_{11} + (1 - \tau_B)A_{12} = \tau_B A_{21} + (1 - \tau_B)A_{22} \\ &\Rightarrow \quad \tau_B = \frac{A_{22} - A_{12}}{A_{11} + A_{22} - A_{12} - A_{21}} = \frac{1}{2} - \frac{\mathrm{RC}_A}{2\mathrm{BC}_A} \end{aligned} \tag{9}$$

If the trustor is willing to assume that the trustee is rational, they can use τ_B as a best-response threshold to make decisions based on the trustee's trustworthiness ($p^A(TW) > \tau_B$, where $\tau_B = C$ in Wagner's fifth condition). While the basic result is well known [43], we can still glean a few key insights. First, the Nash equilibrium only holds if the trustee is assumed to be "rational." Of course, if one suspects the other as being "nasty" or a direct competitor, then the assumption of rationality does not hold [44]. Second, this approach highlights that a single act of untrustworthiness or trustworthiness may not be meaningful, but that trustworthiness is to be assessed dynamically over the relationship's span or at least from some previous expectation or likelihood, shedding light on the roles of familiarity and learning in trust. However, it does not address "thin," one-shot trust interactions, though a Bayesian *prior* over the trustor's belief may be considered as a potential alternative.

2.1 In Gottman's index, trust

Unhappy with the poor accuracy of the Nash equilibrium to predict trust and its limitation to rational actors, Gottman proposed a TI based on his findings from experimental psychology [44]. This index was based on an idea from the Nash equilibrium that we want to maximize the payoffs such that no player can unilaterally choose a move that does better, but drops the rationality assumption and is neither predicated on interaction history nor predicated on the probability of trust. While Gottman's original TI was based on three potential actions per actor, we present a modified version of it here limited to the binary trust decision and translated into interdependence terms.[5] The TI is thus given by:

$$\mathrm{TI} = \frac{A_{11} - A_{22}}{A_{11} + A_{21} - A_{12} - A_{22}} = \frac{1}{2} + \frac{\mathrm{RC}_A}{2\mathrm{FC}_A} \tag{10}$$

where we recall that RC_A is the unilateral control the trustor has over their own outcomes and FC_A is the unilateral control the trustee has over the trustor's outcomes. Gottman describes this as "without regard for the trustee's gains, the trustee can be counted on to look out for the trustor's interests by changing their behavior to improve the trustor's outcomes" [44]. We can also understand this index as the equilibrium achieved between the trustor's choices given the probability, TI, that the actors will not match behaviors: trust will meet untrustworthiness and distrust with trustworthiness, as it can be derived from

$$(1 - \mathrm{TI})A_{11} + \mathrm{TI}A_{12} = (1 - \mathrm{TI})A_{22} + \mathrm{TI}A_{21} \tag{11}$$

4 For proof, see Appendix, Theorem B.1.

5 For proofs, see Appendix, Theorem A.3.

Given our derived constraint $FC_A > |RC_A|$ in equation (3), the index is no longer arbitrary, as presented by Gottman, but becomes a proper metric, such that a trust interaction can be said to not exist if TI < 0 or TI > 1. Furthermore, when 0.5 < TI < 1, trust can be said to be freely given, and when 0 < TI < 0.5, trust is forced or coerced. The latter could occur if RC_A and FC_A are *discordant* and since FC_A >0 (see equation (3)), trust must be being incentivized by the trustee against the trustor's negative inclination ($RC_A < 0$, $FC_A > 0$) [42], as in the aforementioned example with the autonomous car where TI = 0.29. One question explored further later is whether this incentivization/coercion is enough to convince the trustor to go ahead despite their misgivings. Gottman validated his TI through its positive correlation with the trustor's higher emotional attunement and lower physiological arousal and the trustee's reduced negativity and greater openness during oral relationship history interviews. Thus, he concluded, his TI does indeed reflect trust within intimate relationships [44]. As we will show, the index will also prove to be a powerful tool for predicting trust in both human–human and human–machine interactions.

2.2 Committing trust

Is the TI alone sufficient to predict trust? It seems to capture much of the interplay in the one-on-one interaction. However, it does not seem to address a central question of trust researchers from both HRI [16] and psychology [45] on how one decides to interact in the first place. Often we think of trust as a choice between doing something ourselves versus delegating to another; this has been studied in HRI classically as the self-confidence vs human–machine trust going back to Lee and Moray [16]. However, as others have pointed out [23], we often have more than one potential trustee – whether we are choosing among apps, a new car, lab partners, potential business opportunities, or people to date. How do we choose which of these avenues are worth pursuing?

In their initial work on interdependence theory, Thibaut and Kelly introduced the idea of the comparison level for the alternative, CL_{alt} [33]. This is a set point from which we compare our lowest acceptable payoffs for each interaction. The lower the CL_{alt}, the more the interaction is worth pursuing among the set of all interactions; CL_{alt} connotes the anticipated worth of the current interaction and the likelihood that it will be pursued further (i.e. the interaction's stability).

Rusbult and Buunk developed this idea further with Kelley [45], successfully validating the idea that CL_{alt} can be understood psychologically as one's commitment to the interaction. She found in her Investment Model of Commitment that as people discount or reject other CL_{alt}'s in favor of the current relationship, they become more invested in and dependent on the relationship. Likewise, increasing the worth of CL_{alt} by comparing other potential partners to one's current partner, for instance, leads to a cascade not just of distrust but ultimately betrayal. Gottman continued to build on and test this idea [44], concluding that conflict avoidance exacerbates CL_{alt}, which is reflected in further detachment. He also differentiated between commitment and trust, whereby "turning away erodes trust" but "turning away and increasing CL_{alt} erodes trust and fuels betrayal" [44].

Based on the functional requirements for CL_{alt} as described in the aforementioned works, we proposed a new transformation process [42], like those in ref. [46], to apply CL_{alt} to an interaction and understand it via interdependence, as shown in Figure 3. The CL_{alt} transformation does not affect FC_A or BC_A and thus does not directly affect the trustor's interdependence. However, increasing CL_{alt} decreases RC_A by an equivalent amount (i.e. $RC'_A = RC_A - CL_{alt}$). This transformation explains the erosion of Gottman's TI as the trustor's commitment lessens and the ratio of $RC'_A : FC'_A$ decreases.

As the trustor's commitment wanes, the percentage of the time the trustee must act trustworthy to "prove themselves" increases, as shown in equation (9). The idea of "neediness" in psychological game theory [1] is mathematically equivalent to increasing RC_A and commitment through consideration of a negative CL_{alt} but is only mentioned in passing and is less developed therein.

As explained initially by Kelley and Thibaut, the higher one actor's CL_{alt} relative to the other's, the more power they are said to have in the relationship, though this is not necessarily true for any single interaction. This is because one may choose to make themselves vulnerable (or needy as seen through the lens of game theory) in the short term, either through sacrifice or accom-

THE CL_{alt} TRANSFORMATION

O_A	TW	¬TW
T	A_{11}	A_{12}
¬T	A_{21}	A_{22}

Given Matrix (O_A)

⟶

TW	¬TW
A_{11}	A_{12}
A_{21}+CL_{alt}	A_{22}+CL_{alt}

Effective Matrix (O'_A)

$FC_A = FC'_A$ $BC_A = BC'_A$ $RC_A\text{-}CL_{alt} = RC'_A$

Figure 3: The CL_{alt} transformation. CL_{alt} only reduces RC_A and therefore commitment.

modation, in order to signal trustworthiness, without compromising their overall power. In the long term, however, doing so abdicates power and deepens one's dependence, commitment, and, indeed, "neediness." In considering alternatives, there are two further effects that we have previously derived [42], which are worth summarizing here. Given that the payoff/cost of the alternative, CL_{alt}, is inversely proportional to BC_A in the Nash equilibrium and FC_A in the TI.

(1) As the cost of alternatives grows very high ($\Rightarrow CL_{alt} < -2FC_A$), the commitment, RC'_A, increases to the point such that the TI, $\Rightarrow TI > 1$ increases above and beyond what Player B's trustworthiness indicates. This can lead to a coerced over-trust by Player A through "sunk cost" or over-commitment.
(2) Strong alternatives ($CL_{alt} \gg 0$) decrease commitment, RC'_A, to the interaction at hand, lowering the expectation of trustworthiness, τ_B, such that it may no longer meet the required threshold C to trust, where $\Delta\tau_B = \frac{CL_{alt}}{2BC_A}$.

Note that the first point provides a psychological, interdependence-based explanation of the sunk cost "fallacy." Here though it is perfectly rational and not a fallacy, *per se*. Commitment is a sunk cost as other opportunities are foregone and more personal power is ceded so that the other must be increasingly trusted.

The last point is related to a concept that Gottman entitles "turning toward/away." Recall that BC is the payoff for cooperation. It turns out that, at least theoretically, the higher the payoff for "turning toward" the other, the lower the effect of alternatives should be. In other words, "turning away" from the other decreases the robustness of relationships to alternatives, and "turning toward" the other increases its robustness, precisely the effect found in Gottman's studies [44].

3 Experiment 1: capturing human–human trust

While less fully developed, other works had previously noted the importance of coordination, commitment, and the TI in theory and experimentation, both from human–human interaction [44] as well as, to a more limited extent, from human–robot trust [34]. However, these concepts still lacked direct validation based upon quantitative data. Therefore, this is the first experimental goal of this work. The second goal is to look at the implications of our findings and indicate further directions that such game-theoretic analysis may apply to HRI and in what ways it is expected to differ from human–human trust.

3.1 Experimental procedure 1

To test and validate our work, we turned to a competition data set [40] that contained 240 unique, non-trivial games generated from 10 "classical" non-trivial game types, such as "trust," "near dictator," "costly punishment," and "safe shot." Each of the 240 generated games was played between 116 students that were paired off, but blind to each other's choices, with pairings changed for each game. Students were drawn from a business school subject pool and compensated based on the payoffs and choices made in one of the played games, chosen at random. The games were divided into 120 for training the estimation algorithms and 120 to be used to validate prediction accuracy. Results from the top 15 performing algorithms in both the estimation and prediction components of the competition were publicly reported as well as baseline results and a coding template for implementation [40].

To this data set, which included over a dozen strategies of gameplay, we added our various interdependence theory-derived variables as well as the TI previously mentioned. The full list of algorithms and variables can be found in Tables 2 and 3, respectively. We normalized all payoffs by their most extreme value, as discussed earlier, to counter issues that could arise given game theory's utility invariance under strictly positive affine transformations. After validating the baseline code, all games that did not fit our minimal criteria for defining trust games (e.g., exposure and improvement, equation (3)) were removed, resulting in a reduced set of 47 estimation games and 59 prediction games. The data sets were not re-equalized by size, so that comparisons could be made against the baseline results from the competition. Games that did not fulfill the Temptation criteria were retained, in part due to previous HRI work not including that requirement in trust interactions [34,42] and furthermore because temptation is directly related to the trustee's commitment, a condition that we wished to test and not simply exclude. All parameter values in the baseline algorithms were reoptimized with the goal of minimizing the mean squared error. Our models did not make the strong presumption of ref. [40] to remove the intercept *a priori*, since there was no reason to believe that the mean of trust on the y-axis should be 0. In fact, if trust is examined independently of any antecedents such as

Table 2: Previously proposed strategies of trusting and trust fulfillment

Algorithm	Strategy
SPE	Players follow "rational" strategies
Inequality aversion [40]	Players avoid inequality but weight disadvantageous and advantegous inequality differently
Equality reciprocity (ERC) [47]	Mixing SPE, gains from co-ordination (trustor), and tit-for-tat (trustee). All material payoffs being equal, players prefer equal distribution
Charness–Rabin (CR) [40,48]	Combining SPE with the idea of fairness/kindness (tit for tat)
"Seven Strategies" [40]	Regression analysis of strategies that one or both players may employ. See Table 3 for full list.

familiarity and faith in society, then it must account somewhere for potential background bias, which is at least expected on the part of the trustor. All regression algorithms were tenfold cross-validated.

An important caveat of this data set (and in fact all game-theoretic and interdependence-based games in the literature) is that $A_{21} = A_{22}$ and $B_{21} = B_{22}$, which implies, for those still following, that $FC_A = BC_A$ and $RC_B = BC_B$. Thus, the trustee's commitment is equal to their additional incentive to cooperate, and the control the trustee has over the trustee's payoffs is an even mixture of unilateral and joint control.

3.2 Results

Due to concerns of multicollinearity among the 16 variables the variance inflation factors (VIF) for the data set were checked (Table 3). Gottman's TI and the commitment of the trustor (RC_A) showed a correlation of 97% and the trustee's SPE (b1) and the trustee's strategy of maximizing "niceness" (mn1) were heavily correlated at 93%. The trustor's SPE (ri) had medium strength correlations with both the trustor assuming a malicious trustee (maxmin) and the trustee's commitment (RC_B) (51 and 55%, respectively). Given the importance of SPE and our hypothesis, we dropped maxmin and tested both dropping RC_A and TI, settling on TI as it showed stronger results, which brought all VIF below 3 except for ri (VIF = 3.65).

After our data were checked for various statistical assumptions, all of the game play strategies of ref. [40] and several machine learning regressions led to the results shown in Table 4. As in ref. [40], playing the SPE alone still accounted for 75.3% of the variance in the trustor's trusting response and 97.6% of the variance for the trustee's fulfillment of trust in the reduced data set.

Interestingly even with reoptimizing parameters, many of the methods tested by ref. [40] did not perform as well as or only slightly better than the SPE for prediction when only

Table 3: Seven strategies and interdependence variables and their initial and final VIF, after strong multicollinear variables were dropped

Variable	Meaning	VIF init.	VIF final
	"Seven Strategy" variables [40]		
ri	SPE for trustor	**4.48**	**3.65**
lev1	Trustor maximizing self-payoffs given total uncertainty	**4.07**	2.78
mm1	Trustor maximizing payoffs of weakest player (kindness)	2.04	1.84
maxmin	Trustor maximizing payoff assuming Player 2 is malicious	**3.60**	Dropped
jm1	Maximizing joint payoffs	2.97	2.10
ia1	Minimizing payoff differences (equality)	1.97	1.54
b1	SPE for trustee	**10.58**	1.71
mn1	Trustee maximizing trustor's payoff if rational choice is indifferent	**11.07**	Dropped
mm2	Trustee maximizing payoffs of weakest player (kindness)	2.20	1.88
ia2	Minimizing payoff differences (equality)	2.02	1.57
	Interdependence variables		
RC_A	Trustor's commitment	**25.75**	Dropped
FC_A / BC_A	Trustee's unilateral and joint control over trustor	2.47	2.46
RC_B / BC_B	Trustee's commitment and joint control	1.68	1.36
FC_B	Trustor's control over trustee	2.21	1.95
TI	Gottman's TI	**27.59**	2.26

Note: A VIF over 3 is indicative of high multicollinearity. Values above this are indicated in bold.

Table 4: Mean squared error for trust and trust fulfillment: All regressions were run at least initially with all Seven Strategy and interdependence variables. The three best performers along each category are highlighted

	Estimation		Prediction	
Method	Trustor	Trustee	Trustor	Trustee
Subgame Perfect Equilibrium (SPE)	0.1288	0.0184	0.0432	0.0065
Inequality Aversion	0.0336	0.0249	0.0229	0.0071
Equality Reciprocity (ERC)	0.0378	0.0176	0.0509	**0.0057**
Charness-Rabin	0.0729	0.0036	0.0626	0.0263
Seven Strategies	0.0802	0.0035	0.0373	0.0077
Linear Reg.	0.0183	0.0098	**0.0218**	0.0069
Reg. SVM	**0.0075**	**0.0020**	0.0263	**0.0051**
Reg. Tree	0.0144	0.0035	**0.0210**	0.0124
Gaussian Proc. Reg.	**0.0065**	**0.0021**	0.0219	**0.0057**
Ensemble Reg.	**0.0052**	**0.0021**	**0.0141**	0.0077

trust games were analyzed. This was despite their sometimes significant improvement over the estimation set. In addition, when the full set of seven strategies and interdependence variables were fitted and tested in various regression schemes, the seven strategy variables were almost always discarded by the models as insignificant with regard to the trustor. All of the best performing algorithms (linear regression, support vector machine [SVM], Gaussian process regression [GPR], and ensemble regression) showed that the interdependence terms better captured the likelihood of trust, both in terms of lowest error rates and fewest terms. While SVM and GPR prevent us from examining which variables were most impactful, we can use linear regression and the tree-based methods (regression tree and the boosted ensemble) to draw some meaningful conclusions.

3.3 When to trust

Starting with the dropping of terms, as recommended by VIF, and then performing stepwise improvement, the final linear regression model (shown in Figure 4) found that there was a significant bias toward trusting (0.423) and that the most important variables were Gottman's TI and the trustee's commitment/cooperation (RC_B/BC_B). Both FC terms were of borderline significance ($p = 0.057$ and 0.050). Marginal improvement in the mean squared error occurred if FC_A/BC_A was dropped (0.0244 to 0.0218), but there were no gains if FC_B was removed. Since all variables are normalized, the regression weights can be compared against each other, indicating that while FC_B may have borderline significance, its effect is an order of magnitude weaker than the other terms.

The best performer, the least-squares boosted ensemble, showed similar results to the linear regression analysis, with the trustee's commitment/cooperation gains (RC_B/BC_B) playing the largest role followed by the TI, as shown in Figure 5. Note that as long as the trustee seems at least indifferent to commitment/cooperation ($RC_B/BC_B > -0.12$), TI is sufficient for indicating whether trust is bestowed.

```
Linear regression model:
    pr ~ 1 + nFC_A_BC_A + TI + nRC_B_BC_B + nFC_B

Estimated Coefficients:
                    Estimate       SE       tStat      pValue
                    ________    ________    _______    _________

    (Intercept)      0.42318    0.082217     5.1471    6.5952e-06
    nFC_A_BC_A      -0.20955     0.10726    -1.9536      0.057428
    TI               0.35616    0.075656     4.7077    2.7343e-05
    nRC_B_BC_B       0.45106     0.04027     11.201    3.4146e-14
    nFC_B           0.046051    0.022839     2.0164      0.050186


Number of observations: 47, Error degrees of freedom: 42
Root Mean Squared Error: 0.144
R-squared: 0.789,  Adjusted R-Squared: 0.768
F-statistic vs. constant model: 39.2, p-value = 1.18e-13
```

Figure 4: Final tenfold cross-validated and stepwise-improved linear regression for the trustor. nFC_A_BC_A is FC_A/BC_A, nRC_B_BC_B is RC_A/BC_B, nFC_B is FC_B, and *pr* is the probability that trust was given.

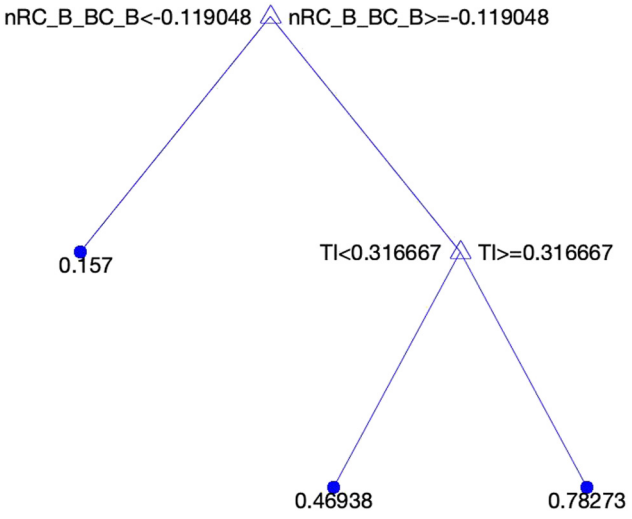

Figure 5: Boosted ensemble tree for trustor–the two retained variables are the trustee's commitment/joint control, RC_B/BC_B, and the TI.

Of further interest is that trust is bestowed even if the TI is below 0.5. In that regime, RC_A and FC_A have opposite signs, indicating that trust is being forced; in the case of this data set, FC_A is always positive, which means that when TI < 0.5, RC_A must be negative and trust is being incentivized. Since neither term can be greater than one, we also see that the trustor's negative commitment is no more than 0.37 ($RC_A > -0.37$), so the trustor's lack of commitment in such cases may be understood as bordering on indifference. To summarize, the trustee's control over the trustor's outcome greatly overrides lack of commitment as long as the incentive to trust/cooperate is about 2.7 times greater. Furthermore, the trustor is likely to strongly trust the trustee when their own commitment aligns with that of the trustee's incentivization/cooperation. Thus, for the trustee, not only does incorporating the interdependence results better predict the probability of trusting but also it appears that just a small subset of the interdependence variables alone gives a more accurate, simpler, and common-sense model of trust, than the "seven strategies" or models of trust based on (in)equality or fairness.

3.4 On being trusted

The results for the trustee display a rather different pattern. The various regression methods, including the interdependence variables, generally performed much better than the baselines [40], especially on the estimation set, but only carefully optimized equality-reciprocity (ERC),

```
Linear regression model:
    pb ~ 1 + b1 + mm2

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)    0.048775    0.026568    1.8359     0.073137
    b1              0.84148    0.034138    24.649     2.193e-27
    mm2             0.14264     0.03348    4.2604    0.00010583


Number of observations: 47, Error degrees of freedom: 44
Root Mean Squared Error: 0.102
R-squared: 0.946,  Adjusted R-Squared: 0.943
F-statistic vs. constant model: 382, p-value = 1.58e-28
```

Figure 6: Final tenfold cross-validated and stepwise-improved linear regression for the trustee. b1 is the trustee's SPE and mm2 is the trustee's strategy to help the weakest player.

SVM, and GPR algorithms could outperform the subgame equilibrium during prediction. All methods are dominated by the SPE but the linear regression analysis (Figure 6) and tree-based methods (Figure 7) revealed that the next most important variable is mm2, when the trustee maximizes the payoffs of the weakest player. The tree-based methods showed that this is especially important when they are already inclined to fulfill trust (b1 > 0.75). This may also explain why ERC or Rabin's kindness algorithms prove so strong on the estimation set- equality/fairness is incorporated into the trustworthiness decision but only after narrow self-interest is considered. While the interdependence parameters are not invoked in these models, the regression tree (not shown) does suggest that the trustor's control over the trustee (FC_B) may play some small role in encouraging cooperation when the SPE tends toward defection.

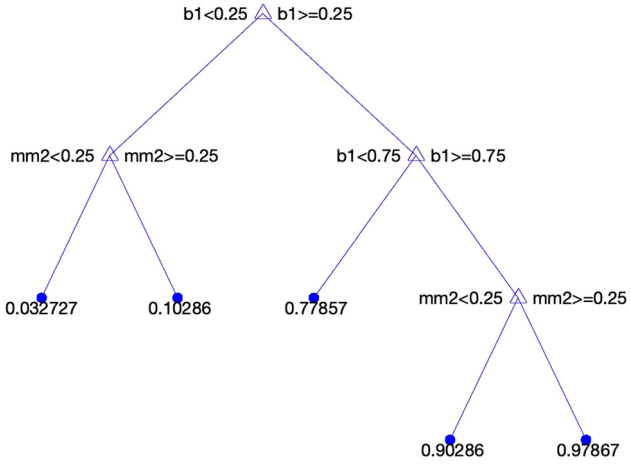

Figure 7: Regression tree for the trustee on deciding when to fulfill trust. b1 is the trustee's SPE and mm2 is the trustee's choosing to help the weakest player (kindness).

3.5 Discussion

Based on these results, it appears that the interdependence-based models best capture the response of the trustor compared to all other strategies and methods from a purely modeling perspective. This result further validates the theoretical development of the TI in field experiments [44]. It also lends credence to our commitment model [42]; initially derived from field experiments [44,45], the commitment model is replicating well in the lab. We also see some strong support for common-sense theories that often get less play in the game theory or HRI trust literature. This is especially true of the unilateral control each agent has in incentivizing or penalizing the other (FC), not only as a strategy but also as a second-order consideration for both players [41].

These second-order beliefs, that is the beliefs the trustor has of what the trustee believes of them, and *vice versa*, have been considered by game theory for many years, but understandably have garnered less attention from HRI. Yet it is precisely these "auxiliary" beliefs that trustors must consider. The trustee reveals them in their extraneous use of fairness, helping the weaker player/reciprocating equality when it is already in their best interest to fulfill trust. Furthermore, the trustor seems to place the commitment of the trustee before any other considerations (RC_B being the root of their regression tree and the highest weighted term in the linear regression). This is only furthered by the trustee's beliefs concerning the power of the trustee in sanctioning or incentivizing the trustor.

If games had been excluded based on the Temptation condition, then the role of the TI and the trustee's kindness would have been completely obscured, as then $RC_B < 0$ and b1 = 0. This scenario would likely be extremely rare in human–machine trust but its further implications are left to future work.

One weakness of this data set, and game-theoretic trust in general, as well as much HRI work on trust, is the assumption that the trustor's and trustee's payoffs are equal if the trustor chooses not to trust in the first place ($A_{21} = A_{22}$ and $B_{21} = B_{22}$). From a technical perspective, this prevents us from determining whether it is RC_B or BC_B that matter for the trustor and confounds whether BC_A does have a significant effect due to its somewhat synthetic "perfect correlation" with FC_A, and thus with TI. As discussed in ref. [42], in psychological game theory and interdependence theory, where psychological costs such as regret and satisfaction are included, these values are rarely equal. New potential relationships generally mean regret is more costly (lowering A_{21}), whereas one-off interactions with strangers and team projects generally value satisfaction with coordination more highly (raising A_{22}). Both of these situations lead to $BC_A > FC_A$ [42]. In comparison, one's overall optimism or brand trust can increase A_{21}, and conversely, pessimism can decrease A_{22}, such that $A_{21} > A_{22}$. This conclusion is reasonable, as commitment is relative to the specific relationship and would thus be moderated by one's overall sense of other potential trust relationships [42].

An illustrative example further validating these points can be found in Dunning et al. [2]. That series of experiments looked at trustor's risk tolerance, whether they wanted to trust, felt like they should trust, and the guilt and agitation they anticipated feeling at not trusting (when the other may be trustworthy). As in many other studies, they found that people "over-trusted" based on rationality (SPE) and risk tolerance alone. In general, the choice to trust was closer to what people felt like they should do vs what they wanted to do. This choice was therefore understood to be partially motivated by anticipated agitation at not trusting, as well as perceived approval of normative behavior from authority. From our results, we can understand both of these unilaterally as increasing the trustor's commitment. Familiarity increased repayment expectations, seemingly through the improved calibration of the threshold for FC_A. Furthermore, they found that when the trustee is seen as making a thoughtful decision to trust instead of just choosing at random, they are more likely to be trusted, illustrating the importance of second-order considerations. However, participants also preferred to give others the opportunity to be trustworthy, which they perceived as a sign of respect for autonomy. In our experiment, this may point to the small effect of equality/fairness in amplifying the trustee's SPE. Once trustworthiness is called for, it pays to more strongly signal one's commitment to fairness/equality as the trustee. Further, evidence from Dunning et al.'s trials pointed to trusting above rationality to be predicated on self-perceived moral norms of fulfilling one's social duty and to avoid casting aspersions on another's character. However, taken together these last points posit an alternative account that would suggest a key testable difference in modeling human–human vs human–robot trust interactions as norm fulfillment.

4 Experiment 2: breaking down trust

This experiment looked to rectify the shortcomings of the previously tested data set by (a) considering trust

between humans as well as between humans and various technologies, (b) employing more realistic scenarios, (c) taking into account various types and quantities of risk, and (d) breaking the $FC_A = BC_A$ and $RC_B = BC_B$ assumption of previous game theory and HRI trust research.

4.1 Experimental procedure 2

In this experiment, 34 different scenarios were composed across 9 different types of risks: physical, psychological, social, time loss, performance, financial, ethical, privacy, and security, based on ref. [49]. We assigned each participant 8 scenarios, drawn from 2 of the 9 risk types, with an equal balance of human–machine and human–human scenarios, leading to a $2 \times 2 \times 2$ within- and between-subjects design. Examples of human–human trust included taking a friend's suggested route to avoid traffic, having a stranger watch luggage briefly, dividing up work with classmates, and participating in pharmaceutical trials. Examples of human–machine trust included following GPS guidance, using a dating app, driving an autonomous vehicle, taking emergency guidance from a robot during a fire, and trusting enemy classification from a military drone.

Each scenario was composed of two elements: a payoff table and a scenario written out in prose. Payoffs were created randomly but some scenarios dictated certain constraints, beyond those of equation (3), that we coded for. The general nature of these constraints will be discussed later. Scenarios were also designed to reflect a wide range of scales (10^0–10^7). To maintain consistency, participants only acted as trustors. Given the high level of convergence in human trustee behavior in Experiment 1 to the SPE, trustee behavior for both human and machine scenarios was algorithmically determined with some noise injected.

Sixty participants took part in this experiment (55% male, 43% female, 2% non-identified), ranging from 18–50+ (85% between 18 and 39), and 78% having at least some post-secondary education. Before the experiment, participants underwent training including a practice round to become familiarized with the layout, expectations, and most importantly, how to read and understand the payoff table. Their understanding of gameplay was assessed both after training and at the end of the experiment. After the experiment, general feedback was solicited and a number of insights into carrying out such experiments in the future were collected. Given our desire that participants understand each task, there was no time limit, and most spent 2–4 min per scenario.

The experiment was carried out using the Gorilla Experiment Builder (www.gorilla.sc) to create and host our experiment [50]. All research performed with human participants was done in compliance with all relevant national regulations, institutional policies, and in accordance with the tenets of the Helsinki Declaration, and was approved by the Georgia Institute of Technology's IRB. Participants were recruited through Prolific, and the data were collected between March 22 and March 23, 2021.

All the same algorithms deployed in Experiment 1 were tested again here, with the exception of GPR which was replaced with binomial regression. All algorithms were modified to accommodate $A_{21} \neq A_{22}$ and $B_{21} \neq B_{22}$. Furthermore, although earlier we had a regression problem to solve, now with every participant having different payoffs, we approached the experiment as a classification problem. While this will affect the meaning of the error rate, the overall patterns of performance and variable importance should remain clear.

4.2 Results

In this experiment, we only modeled the trustor and not the trustee. Thus, we did not have to consider the final three of the "Seven Strategies" from Table 3. VIF was once again performed, resulting in maxmin, lev1, and RC_A being dropped, keeping the remaining VIFs <3.5). Inequality Aversion, ERC, CR, and the SPE were all strongly correlated with each other (corr $= 0.65 - 0.83$) and exhibit multicollinearity, so only the SPE was retained.

The total variance that could be explained by the SPE alone was 55.6%. Like in Experiment 1, the strategies from game theory only performed slightly better than the SPE, at best. The linear and binomial regressions worked somewhat better than these strategies, reaching 75% accuracy. However, the machine learning classifiers performed estimation significantly better, all achieving over 85%. These classifiers were tested on the Interdependence terms and indices as well as a set combining the Seven Strategies with the Interdependence terms. Once again, the classifiers all performed as well or better with the Interdependence terms alone, with the SVM reaching a maximum of 92% accuracy and the ensemble KNN 100% for estimating the decision to trust based on the Interdependence terms alone over the whole data set. The estimations for human–human vs human–machine were extremely similar across the board, with the more traditional game-theoretic strategies and regressions per-

Table 5: Mean squared error for trust and trust fulfillment: All classifications were run at least initially with all Seven Strategies and interdependence variables. H–H are human–human trust scenarios and H–M are human–machine. The three best performers along each category are highlighted

	Estimation		
Method	Total	H-H	H-M
Subgame Perfect Equilibrium (SPE)	0.446	0.396	0.489
Inequality Aversion	0.435	0.454	0.421
Equality Reciprocity (ERC)	0.435	0.394	0.470
Charness-Rabin (CR)	0.442	0.394	0.481
Seven Strategies	0.410	0.361	0.421
Linear Reg.	0.348	0.306	0.333
Binomial Reg.	0.348	0.310	0.307
Clas. SVM	**0.079**	**0.093**	**0.095**
Clas. Tree	**0.117**	**0.134**	**0.102**
Clas. KNN Ensemble	**0**	**0**	**0**

Table 6: Performance measures: the receiver operating characteristic's area-under-the-curve (AUC), the Matthew's correlation coefficient (MCC), and the *k*-fold loss

Method	ROC-AUC	MCC	k-Fold Loss
SPE	0.51	0.085	0.446
IA	0.53	0.121	0.492
ERC	0.53	0.111	0.442
CR	0.60	0.093	0.417
Binomial Reg.*	**0.70**	**0.356**	**0.348**
Clas. SVM*	**0.66**	**0.258**	**0.387**
Clas. Tree	0.58	0.104	0.419
Clas. Tree Ensemble*	**0.70**	**0.293**	0.395
Clas. KNN Ensemble	0.62	0.205	**0.379**

The * indicates these classifiers were optimized to minimize *k*-fold loss using the estimator as a baseline. The three best performers along each category are highlighted

forming somewhat better for human–human. This pattern did not replicate for the classifiers using interdependence-only terms. All of these results are summarized in Table 5.

We calculated other performance measures, such as the receiver operating characteristic area-under-the-curve (ROC-AOC), Mathew's correlation coefficient (MCC), and k-fold loss, to assess model fit and ability to predict trust (Table 6). They indicate the same general pattern of interdependence dominating the more traditional measures; however, it is clear, though not surprising, that prediction errors are significantly higher than estimation errors. Most of the game-theoretic approaches did not do much better than random chance at prediction (estimated through k-fold loss), whereas the interdependence approaches all improved upon that. Surprisingly, binomial regression edged out the Classification Tree as the third-best performer. The binomial regression (Figure 9) and SVM were further

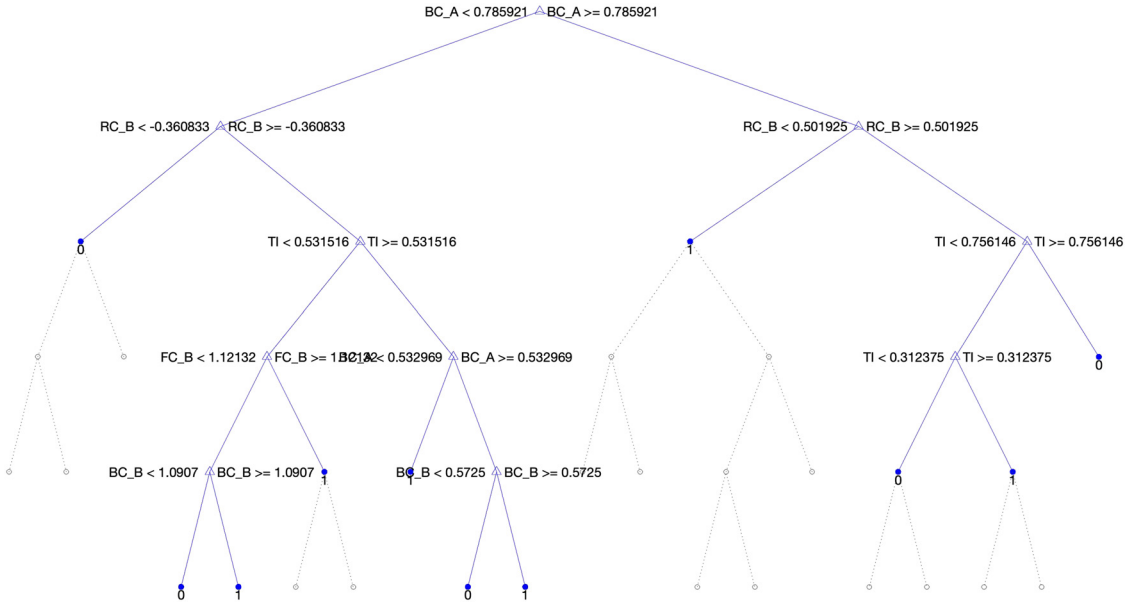

Figure 8: Full classification tree for when the trustor decides to trust from Experiment 2. Pruning was performed during optimization. Leaf nodes marked with 0's indicate lack of trust and those with 1's indicate trust.

optimized to minimize k-fold loss and the classification tree was improved upon by leveraging an ensemble learner.

4.3 Discussion

4.3.1 When to trust: Part II

Once again the interdependence variables were retained and dominated across the board throughout the machine learning methods, especially the KNN ensemble and SVM. As in Experiment 1, the SVM did not allow us to "see inside" and understand which variables mattered most. However, we can turn to the classification tree to understand not only the significant variables but also the underlying logic (Figure 8). Except for the root, this tree resembles that of Experiment 1, with a slightly negative RC_B ranking above TI. While the split values differ somewhat they approximate the pattern we saw before. The differences here are that we have now disentangled FC_A and BC_A, as well as RC_B and BC_B. While before BC_A was subsumed into TI through its forced equivalence to RC_A, now we can see that BC_A is important enough to become the root of this tree. BC_B and FC_B also serve as interesting additions as they were seemingly absent from the tree in Experiment 1 despite being indicated as playing a role in the linear regression there.

The power of the classification tree is the explainability it enables. Working through it, we see that the trustor's primary consideration is BC_A or what they stand to gain from cooperation, followed by the trustee's commitment (RC_B), followed by the TI. This alone gets us to 70% accuracy, well above the other methods. Further refinements also account for FC_B and BC_B, while the whole tree taking on a finer and finer grain. FC_A by itself is by far is the least important variable of the interdependence terms for determining when the trustor trusts. However, it has a pivotal role in TI and it is clear that the main roles of FC_A (and even RC_A) is through its interaction in TI.

```
Generalized linear regression model:
    logit(trust) ~ 1 + BC_B + SPE*RC_B + FC_A*RC_B + RC_B*FC_B + BC_A*TI
    Distribution = Binomial

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)    -0.92514     0.7972    -1.1605       0.24585
    SPE            -0.52673    0.44234    -1.1908       0.23374
    FC_A            -1.5168    0.70698    -2.1454       0.03192
    BC_A             2.9979    0.87599     3.4222    0.00062107
    RC_B            -2.3936     1.2375    -1.9341        0.0531
    FC_B           -0.23589    0.22535    -1.0468       0.29521
    BC_B            0.82598    0.30035     2.7501     0.0059583
    TI               3.3946     1.1323     2.9981     0.0027165
    SPE:RC_B         -2.134    0.80131    -2.6631      0.007742
    FC_A:RC_B        4.9535     1.6077      3.081     0.0020627
    RC_B:FC_B        1.5508    0.55551     2.7917     0.0052428
    BC_A:TI         -3.5877     1.5166    -2.3656      0.017999

480 observations, 468 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 59.4, p-value = 1.17e-08
```

Figure 9: Step-optimized binomial regression. Optimization led to all non-interdependence variables to be dropped except the SPE.

4.3.2 The limits of realism

In this experiment, subjects were presented with a text-based scenario as well as a payoff table and asked if they would choose to trust or not, whereas when the first data set was built and tested only payoff tables were provided without much grounding in reality. However, a number of valid concerns arose in our more realistic approach to trust problems. The amount of time and focus people placed on the table vs the text was a serious concern as was the legibility of the table. The first experiment used participants in a university game theory class who were used to reading payoff tables, whereas the participants in the second experiment were drawn from a general pool, over a wide age range (18–50+) and with varying degrees of education. After the experiment, participants were asked to assess their experience. All said that the amount of training was sufficient, but some felt it was overly wordy. Others indicated that the table could be confusing, as it simply presented so much information, and that a redesign would help. We expected this issue and thought it would lead many to focus on the text-based scenarios. However, the majority of respondents said that they focused more on the table, as it provided and summarized key information, especially the more scenarios they went through. This indicates that a longer training period may be appropriate.

The need for a longer training period, the split focus brought by increasing realism, and the shift from a regression problem to a classification problem all increased noise in the data. This, in turn, may help shed light on why the accuracy of all approaches fell significantly between Experiment 1 and Experiment 2. Even so, the overall pattern of the findings remained the same.

Given the level of understanding we aimed to achieve and the novelty of the scenario format, no time limit was instantiated, as mentioned earlier. Thus, while many types of risk were tested, this experiment did not test trust under time pressure. Time pressure acts not only as another category of risk but also that directly impacts cognitive workload. This is left for future research,

though whether it can be realistically tested within this framework is an open challenge.

Another key takeaway from increasing realism revealed itself when we were designing the scenarios. Some scenarios indicated implicit payoff constraints that we had to account for when generating the payoff matrices. The most common of these were $A_{22} > A_{21}$ (11/34 scenarios), $B_{22} > B_{21}$ (10/34 scenarios), and $B_{11} > B_{12}$ (7/34 scenarios). Generally, this meant that often in trust interactions, satisfaction for not trusting the untrustworthy outweighed the regret for not trusting the trustworthy (as discussed in Section 3.5), that suspicion hurts potential trustees, and that the Temptation condition does not always even apply in real life (see Section 1.3).

A final concern regarding realism is while machine type was specified to some degree in each scenario (*autonomous vehicle, emergency guidance robot*), much was left to the participants' imagination including the extent of anthropomorphism. While this avoided anchoring bias that showing pictures or more detailed descriptions may have introduced, it leaves open questions regarding the influence of design on perceived risk and situation normality.

5 General discussion

5.1 A declaration of interdependence

Together, our two studies revealed the power of the interdependent approach to understanding what defines trust games and when trust actually occurs. Game theory played a crucial role in helping define and refine this approach. Both interdependence theory and game theory converged on how they defined what constitutes a trust game. The requirements of exposure and improvement are accepted across the board, and interdependence theory allowed us to understand those requirements more deeply, as a set of constraints on commitment, cooperation, and coercion.

However, once it came to how people actually play trust games, the game theory strategies proved insufficient, especially once applied to real-life scenarios. It remains clear that people depart significantly from "rational" gameplay of the SPE, a fact that both HRI and psychological game theory have long struggled to explain. Part of the problem may be in the more narrow definition of a trust game in game theory, specifically the Temptation condition ($B_{12} > B_{11}$). In fact, in the second experiment, we tested Temptation explicitly against our other variables and found that it has a near-perfect negative correlation (−0.96) with the SPE, clearly demonstrating that the standard game theory set up has a deep internal contradiction, requiring Temptation while hypothesizing the importance of SPE simultaneously. It is little wonder that "rationally" one should never trust.

Thus, while it is tempting to understand Temptation at what makes trust games actual "games," it is crucial that here we confirmed that it is neither a necessary nor sufficient condition as such. Once we broke the $B_{21} = B_{22}$ equality, Temptation could be re-framed as either $RC_B < 0$ or $BC_B < 0$, that is the relative power/lack of commitment of the trustee or their gains from being competitive. While trust is harder when $RC_B < 0$ or $BC_B < 0$, it is not impossible, nor do $RC_B > 0$, $BC_B > 0$ guarantee that trust will be given or reciprocated. The game is still afoot.

Beyond justifying the dropping of this constraint, the interdependence approach provides a better explanation of trust, in terms of both accuracy and parsimony. Crucially, interdependence allows us to have a conversation about HRI trust without having to resort to reciprocity, altruism, and fairness backed by convoluted explanations of anthropomorphism. By using commitment, coercion, and cooperation to explain trust our model allows us to bridge the divide between human–human and human–machine trust. Furthermore, this approach couches trust in familiar terms, those that we use regularly to describe when and why humans trust.

5.2 Implications for HRI

If trusting is about perceived commitment and cooperative gains as opposed to strict rationality, fairness, equality, or respect, then very different conclusions may be drawn regarding trust as it relates to humans vs robots. Robots are already perceived by humans as being more fair, just, and even reliable [51,52], though this effect is moderated by anthropomorphization. Thus, per Experiment 1, robots conform nicely to the notion of the rational trustee, following the SPE, even if they do not have exhibit extra marginal gains from reciprocity or fairness (mm2). On the other hand, it then falls to the trustor (or modeler) to capture how much benefit the robot can bring the trusting human either unilaterally (FC_A) or through cooperation (BC_A), as well as how much they should commit (RC_A) to the interaction or consider alternatives (CL_{alt}). Furthermore, it appears that the "over-trust" of robots and humans may really come down to perceived gains, power, and need. In the motivating example of the human and the self-driving car, the relative assessments of commitment, safety (via coercion), and reputation (via cooperation)

seemed to explain the interaction more effectively than kindness or fairness.

Several major dimensions of HRI trust can be understood through the lens of interdependence. First and foremost, there are close parallels between reflexive, fate, and bilateral control and the recently proposed and aptly named autonomy dimensions of Commitment, Specification, and Control of ref. [30]. More specifically, interdependence can be seen as another set of insights into the antecedents, correlates, and underlying dimensions of human–robot trust. Affective trust (often called benevolence), assessing whether the other agent is competitive or wishes to cooperate (as in refs [3,18,21]), is foundational to determining whether a trust game even exists [42] and underlies bilateral control (BC). Social [3,53] and structural trust [18,20] are keys to determining levels of commitment (RC) and the norms at play (e.g., equity or kindness). Familiarity [18,19] helps establish thresholds and refines calibrations of anticipated payoffs. While anthropomorphism may shift strategy choice (especially for trust repair) by triggering psychological norms [22,54], it is likely to also play a key role in establishing familiarity and situation normality [18,55,56], and thus feeds into trust calibration. This effect, however, may be confounded by the uncanny valley at some limit [52,57]. While more mechanical robots may be seen as fairer and more efficient, more humanoid ones may be accorded more respect and forgiveness during trust repair. In the middle of the "uncanny trust valley," robots may be seen as having qualities of both ends, either for better [52] or for worse [57].

However, if ref. [2] is correct in that humans are motivated to trust via norm fulfillment out of respect, a completely independent theory of trust would be necessary for humans vs robot trustees. This possibility is made all the more interesting for HRI if that respect is predicated on individual moral autonomy vs a personal autonomy based on agency [58,59]. On the other hand, previous work [53,60,61] on trust in social psychology and HRI has suggested that the underlying dimensions, antecedents, and correlates of trust for human–human and HRI heavily overlap and function in similar ways. Our work strongly comes out on the side of the latter and leaves a major testable contention for future work.

Finally, we have primarily focused on the human being the trustor and the robot being the trustee. The modeling approach we have taken above, though, further opens the door to allowing robots to decide whether to trust the humans with whom they interact. Perhaps, more importantly, such models would allow robots to be more self-aware of higher-order self-reflection, being able to assess the likelihood that they will be trusted by humans and whether this trust is well calibrated, an assessment they can use to give feedback in aiding the human to calibrate their own trust even further.

6 Conclusion

HRI and game theory have each been slowly working towards more complete theories, models, and metrics of trust for the last 35 years. Both have gone beyond capability and pure rationality and started to incorporate psychological and social factors. However, these two fields have yet to fully recognize each other's potential for cross-calibration. Interdependence theory, with its focus on cooperation, control, and commitment, is a key to bridging this gap. Crucially, this work further validated previous research on interdependence theory from social psychology by testing it on a wide range of games and a large subject pool. These variables, especially as they relate to trusting, are shown to be powerfully predictive and are equally amenable to being integrated with previous game-theoretic trust work, as well as expand on an emerging holistic approach to trust in HRI and beyond. Interdependence-based approaches, unlike previous game theory strategies for assessing trust, are equally understandable for human and non-human agents and imply a strong general neuro-psychological model of trust, furthering our goal of illustrating a more complete theory of interactional trust for humans and automation.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The data that support the findings of this study are available from the corresponding author, YSR, upon reasonable request.

References

[1] M. Bacharach, G. Guerra, and D. J. Zizzo, "The self-fulfilling property of trust: An experimental study," *Theory Decision*, vol. 63, no. 4, pp. 349–388, 2007, DOI: https://doi.org/10.1007/s11238-007-9043-5.

[2] D. Dunning, J. E. Anderson, T. Schlösser, D. Ehlebracht, and D. Fetchenhauer, "Trust at zero acquaintance: More a matter of respect than expectation of reward," *J. Pers. Soc. Psychol.*, vol. 107, no. 1, pp. 122–141, 2014, DOI: https://doi.org/10.1037/a0036673.

[3] K. Schaefer, *"The perception and measurement of human-robot trust,"* Ph.D. dissertation, University of Central Florida, 2013.
[4] M. Madsen and S. Gregor, "Measuring human-computer trust, in *Proceedings of 11th Australasian Conference on Information Systems*, 2000, pp. 6–8.
[5] G. Harrison and J. List, "Field experiments," *J. Econ. Lit.*, vol. 42, no. 4, pp. 1009–1055, 2004, DOI: https://doi.org/10.1257/0022051043004577.
[6] M. Deutsch, "Trust and suspicion," *J. Conflict Resolut.*, vol. 2, no. 4, pp. 265–279, 1958.
[7] M. Deutsch, "The effect of motivational orientation upon trust and suspicion," *Human Relat.*, vol. 13, no. 2, pp. 123–139, 1960.
[8] M. Deutsch, *The Resolution of Conflict: Constructive and Destructive Processes*, New Haven, USA: Yale University Press, 1977.
[9] J. Rotter, "A new scale for the measurement of interpersonal trust," *J. Pers.*, vol. 35, no. 4, pp. 651–655, 1967.
[10] J. Rotter, "Interpersonal trust, trustworthiness, and gullibility," *Amer. Psychol.*, vol. 35, no. 1, pp. 1–7, 1980, DOI: https://doi.org/10.1037/0003-066X.35.1.1.
[11] J. Rempel, J. Holmes, and M. Zanna, "Trust in close relationships scale," *J. Pers. Soc. Psych.*, vol. 49, no. 1, pp. 95–112, 1985.
[12] N. Luhmann, *Trust and Power*, Chichester, UK: John Wiley & Sons, 1979.
[13] B. Barber, *The Logic and Limits of Trust*, New Brunswick, NJ: Rutgers University Press, 1983.
[14] J. Jalava, "From norms to trust: The luhmannian connections between trust and system," *Europ. J. Soc. Theory*, vol. 6, no. 2, pp. 173–190, 2003, DOI: https://doi.org/10.1177/1368431003006002002.
[15] B. M. Muir, "Trust between humans and machines, and the design of decision aids, *Int. J. Man–Machine Stud.*, vol. 27, no. 5–6. pp. 527–539, 1987, DOI: https://doi.org/10.1016/S0020-7373(87)80013-5.
[16] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Human–Computer Stud.*, vol. 40, pp. 153–184, 1994, DOI: https://doi.org/10.1006/ijhc.1994.1007.
[17] R. C. Mayer, J. H. Davis, and D. F. Schoorman, "An integrative model of organizational trust," *Acad. Manag. Rev.*, vol. 20, no. 3, pp. 709–734, 1995, DOI: https://doi.org/10.2307/258792.
[18] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in online shopping: An integrated model," *MIS Quarter.*, vol. 27, no. 1, pp. 51–90, 2003, DOI: https://doi.org/10.2307/30036519.
[19] M. Körber, "Theoretical considerations and development of a questionnaire to measure trust in automation," in *Proceedings of the 20th Congress of the International Ergonomics Association* (*IEA 2018*), Vol. VI, 2018, pp. 13–30.
[20] D. H. McKnight, M. Carter, J. B. Thatcher, and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Trans. Manag. Inform. Syst.* (*TMIS*), vol. 2, no. 2, pp. 1–25, 2011, DOI: https://doi.org/10.1145/1985347.1985353.
[21] S. Chien, M. Lewis, Z. Semnani-Azad, and K. Sycara, "An empirical model of cultural factors on trust in automation," *Proc. Human Factors Ergonom. Soc.*, vol. 58, no. 1, pp. 859–863, 2014, DOI: https://doi.org/10.1177/1541931214581181.
[22] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015, DOI: https://doi.org/10.1177/0018720814547570.
[23] C. Castelfranchi and R. Falcone, *Trust Theory: A Socio-Cognitive and Computational Model*, Chichester, UK: John Wiley & Sons, 2010.
[24] J. D. Lee and K. A. See, "Trust in automation: designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004, DOI: https://doi.org/10.1518/hfes.46.1.50.30392.
[25] T. Law and M. Scheutz, "Trust: Recent concepts and evaluations in human–robot interaction," in *Trust in Human-Robot Interaction*, C. S. Nam and J. B. Lyons (Eds.), London: Academic Press/Elsevier, 2020, pp. 27–57, DOI: https://doi.org/10.1016/B978-0-12-819472-0.00002-2.
[26] B. F. Malle and D. Ullman, "A multi-dimensional conception and measure of human-robot trust," in *Trust in Human–Robot Interaction*, C. S. Nam and J. B. Lyons (Eds.), London: Academic Press/Elsevier, 2020, pp. 3–25, DOI: https://doi.org/10.1016/B978-0-12-819472-0.00001-0.
[27] Y. S. Razin and K. M. Feigh, "Hitting the road: Exploring human-robot trust for self-driving vehicles," in *2020 IEEE International Conference on Human-Machine Systems* (*ICHMS*), Rome: IEEE, 2020, pp. 1–6, DOI: https://doi.org/10.1109/ICHMS49158.2020.9209525.
[28] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Towards safe and trustworthy social robots: ethical challenges and practical issues, in *International Conference on Social Robotics*, Paris: Springer, 2015, pp. 584–593.
[29] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *2015 10th ACM/IEEE International Conference on Human–Robot Interaction* (*HRI*), IEEE, 2015, pp. 1–8.
[30] D. J. Atkinson, W. J. Clancey, and M. H. Clark, "Shared awareness, autonomy and trust in human-robot teamwork," in *2014 AAAI Fall Symposium Series*, 2014.
[31] J. Meyer, R. Wiczorek, and T. Günzler, "Measures of reliance and compliance in aided visual scanning," *Human Factors*, vol. 56, no. 5, pp. 840–849, 2014, DOI: https://doi.org/10.1177/0018720813512865.
[32] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, 2010, DOI: https://doi.org/10.1177/0018720810376055.
[33] J. W. Thibaut and H. H. Kelley, *The Social Psychology of Groups*, New York: John Wiley & Sons, 1959.
[34] A. R. Wagner, *The Role of Trust and Relationships in Human–Robot Social Interaction*, Ph.D. dissertation, Atlanta, GA, USA: Georgia Institute of Technology, 2009.
[35] P. Robinette, *Developing Robots that Impact Human–Robot Trust in Emergency Evacuations*, Ph.D. Dissertation, Atlanta, GA, USA: Georgia Institute of Technology, 2015.
[36] P. A. M. Van Lange and C. E. Rusbult, "Interdependence theory," P. A. M. Van Lange, A. W. Kruglanski, and E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology*, 2012,

pp. 251–272, DOI: https://doi.org/10.4135/9781446249222.n39.
[37] J. Geanakoplos, D. Pearce, and E. Stacchetti, "Psychological games and sequential rationality," *Games Econom. Behav.*, vol. 1, no. 1, pp. 60–79, 1989, DOI: https://doi.org/10.1016/0899-8256(89)90005-5.
[38] J. Ermisch and D. Gambetta, "People's trust: The design of a survey-based experiment," in *ISER Working Paper Series*, no. 2006–34. University of Essex, Institute for Social and Economic Research (ISER), Colchester, 2006. https://www.econstor.eu/bitstream/10419/91938/1/2006-34.pdf.
[39] D. Balliet and P. A. Van Lange, "Trust, conflict, and cooperation: A meta-analysis," *Psych. Bulletin*, vol. 139, no. 5, pp. 1090–1112, 2013, DOI: https://doi.org/10.1037/a0030939.
[40] E. Ert, I. Erev, and A. E. Roth, "A choice prediction competition for social preferences in simple extensive form games: An introduction," *Games*, vol. 2, no. 3, pp. 257–276, 2011, DOI: https://doi.org/10.3390/g2030257.
[41] P. Battigalli and M. Dufwenberg, "Dynamic psychological games," *J. Econ. Theory*, vol. 144, no. 1, pp. 1–35, 2009, DOI: https://doi.org/10.1016/j.jet.2008.01.004.
[42] Y. Razin and K. Feigh, "Toward interactional trust for humans and automation: Extending interdependence," in *2019 IEEE SmartWorld: Advanced Trusted Computing*, 2019, pp. 1348–1355, DOI: https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00247.
[43] J. F. Nash, "Equilibrium points in n-person games," *Proc. Nat. Acad. Sci.*, vol. 36, no. 1, pp. 48–49, 1950.
[44] J. M. Gottman, *The Science of Trust: Emotional Attunement for Couples*, New York: W.W. Norton & Company, 2011.
[45] C. E. Rusbult and B. P. Buunk, "Commitment processes in close relationships: an interdependence analysis," *J. Soc. Pers. Relationships*, vol. 10, no. 2, pp. 175–204, 1993, DOI: ttps://doi.org/10.1177/026540759301000202.
[46] H. H. Kelley and J. W. Thibaut, *Interpersonal Relations: A Theory of Interdependence*, New York, NY: John Wiley & Sons, 1978.
[47] G. E. Bolton and A. Ockenfels, "ERC: A theory of equity, reciprocity, and competition," *Am. Econ. Rev.*, vol. 90, no. 1, pp. 166–193, 2000.
[48] M. Rabin, "Incorporating fairness into game theory and economics," *Am. Econom. Rev.*, Vol. LXXXIII, pp. 1281–1302, 1993.
[49] R. E. Stuck, *"Perceived relational risk and perceived situational risk: Scale development*, Ph.D. Dissertation, Atlanta, GA, USA: Georgia Institute of Technology, 2020.
[50] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behav. Res. Meth.*, vol. 52, no. 1, pp. 388–407, 2020, DOI: https://doi.org/10.1101/438242.
[51] E. de Visser, S. Monfort, R. Mckendrick, M. Smith, P. Mcknight, et al., "Almost human: Anthropomorphism increases trust resilience in cognitive agents," *J. Exp. Psych. Appl.*, vol. 22, pp. 331–349, 2016, DOI: https://doi.org/10.1037/xap0000092.
[52] R. Häuslschmid, M. von Buelow, B. Pfleging, and A. Butz, "Supporting trust in autonomous driving," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 2017, pp. 319–329.
[53] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cog. Ergonom.*, vol. 4, no. 1, pp. 53–71, 2000, DOI: https://doi.org/10.1207/S15327566IJCE0401_04.
[54] C. Nass, Y. Moon, and P. Carney, "Are people polite to computers? Responses to computer-based interviewing systems," *J. Appl. Soc. Psych.*, vol. 29, no. 5, pp. 1093–1109, 1999, DOI: https://doi.org/10.1111/j.1559-1816.1999.tb00142.x.
[55] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: A three-factor theory of anthropomorphism," *Psych. Rev.*, vol. 114, no. 4, pp. 864–886, 2007, DOI: https://doi.org/10.1037/0033-295x.114.4.864.
[56] S. Park, "Multifaceted trust in tourism service robots," *Annals Tourism Res.*, vol. 81, art. 102888, 2020, DOI: ttps://doi.org/10.1016/j.annals.2020.102888.
[57] C. B. Nordheim, *Trust in Chatbots for Customer Service-findings from a Questionnaire Study*, Master's Thesis, Oslo, Norway: University of Oslo, 2018.
[58] J. Zhu, *Intentional Systems and the Artificial Intelligence (AI) Hermeneutic Network: Agency and Intentionality in Expressive Computational Systems*, Ph.D. dissertation, Atlanta, GA, USA: Georgia Institute of Technology, 2009.
[59] F. Alaieri and A. Vellino, "Ethical decision making in robots: Autonomy, trust and responsibility," in *International Conference on Social Robotics*, Kansas City: Springer, 2016, pp. 159–168, DOI: https://doi.org/10.1007/978-3-319-47437-3_16.
[60] J. B. Lyons and C. K. Stokes, "Human–human reliance in the context of automation," *Human Factors*, vol. 54, no. 1, pp. 112–121, 2012, DOI: https://doi.org/10.1177/0018720811427034.
[61] F. M. Verberne, J. Ham, and C. J. Midden, "Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars," *Human Factors*, vol. 54, no. 5, pp. 799–810, 2012, DOI: https://doi.org/10.1177/0018720812443825.

Appendix

A.1 Glossary of terms

A_{11}	Trustor's payoff for successfully placed trust
A_{12}	Trustor's payoff (cost) if betrayed
A_{21}	Trustor's payoff (cost) for not trusting and regret
A_{22}	Trustor's payoff if they do not trust an untrustworthy player
B_{11}	Trustee's payoff for successfully returned trustworthiness
B_{12}	Trustee's payoff if they betray the trustor
B_{21}	Trustee's payoff (cost) if they are not trusted when they are trustworthy
B_{22}	Trustee's payoff (cost) if they are not trusted when they are not trustworthy
RC	*Reflexive control*: How much unilateral power each actor has over their own outcomes
FC	*Fate/Partner control*: How much unilateral power each actor has over the other player's outcomes
BC	*Bilateral control*: How much one actor's choice further facilitates or inhibits the other's outcomes, i.e. the cooperative gain
τ_B	Probability of the trustee acting trustworthy
TI	Gottman's TI
CL_{alt}	Comparison level for the alternative
SPE	Sub-perfect equilibrium

A.2 Derivations of interdependent trust conditions

Theorem 1. *If Improvement and Exposure are true* (equations (2) *and* (4)), $BC_A > 0$.

Given $A_{11} > A_{21}$ *and* $A_{22} > A_{12}$

$$A_{11} > A_{21} \cap A_{22} > A_{12}$$
$$A_{11} + A_{22} > A_{21} + A_{12}$$
$$A_{11} + A_{22} - A_{21} - A_{12} > 0$$
$$\therefore \ BC_A > 0$$

Theorem 2. *If Improvement and Exposure are true*, $FC_A > 0$.

Given $A_{11} > A_{22}$ *and* $A_{21} > A_{12}$

$$A_{11} > A_{22} \cap A_{21} > A_{12}$$
$$A_{11} + A_{21} > A_{22} + A_{12}$$
$$A_{11} + A_{21} - A_{22} - A_{12} > 0$$
$$\therefore \ FC_A > 0$$

Theorem 3. *If Improvement and Exposure are true*, $FC_A \geq |RC_A|$.

Given $A_{21} > A_{12}$

$$A_{21} > A_{12}$$
$$2A_{21} > 2A_{12}$$
$$A_{21} - A_{12} > A_{12} - A_{21}$$
$$A_{21} - A_{12} + A_{11} - A_{22} > A_{12} - A_{21} + A_{11} - A_{22}$$
$$2FC_A > 2RC_A$$
$$\therefore \ FC_A > RC_A$$

$$A_{11} > A_{22}$$
$$2A_{11} > 2A_{22}$$
$$A_{11} - A_{22} > -A_{11} + A_{22}$$
$$A_{11} - A_{22} + A_{21} - A_{12} > -A_{11} + A_{22} + A_{21} - A_{12}$$
$$2FC_A > -2RC_A$$
$$\therefore \ FC_A > -RC_A$$

$$A_{11} > A_{22}$$
$$2A_{11} > 2A_{22}$$
$$A_{11} - A_{22} > -A_{11} + A_{22}$$
$$FC_A > RC_A \cap FC_A > -RC_A$$
$$\therefore \ FC_A > |RC_A|$$

Theorem 4. *If Improvement and Exposure are true*, $BC_A \geq |RC_A|$.

Given $A_{21} > A_{12}$

$$A_{22} > A_{12}$$
$$2A_{22} > 2A_{12}$$
$$A_{22} - A_{12} > A_{12} - A_{22}$$
$$A_{22} - A_{12} + A_{11} - A_{21} > A_{12} - A_{22} + A_{11} - A_{21}$$
$$2BC_A > 2RC_A$$
$$\therefore \ BC_A > RC_A$$

$$A_{11} > A_{21}$$
$$2A_{11} > 2A_{21}$$
$$A_{11} - A_{21} > -A_{11} + A_{21}$$
$$A_{11} - A_{21} + A_{22} - A_{12} > -A_{11} + A_{21} + A_{22} - A_{12}$$
$$2BC_A > -2RC_A$$
$$\therefore \ BC_A > -RC_A$$

$$BC_A > RC_A \cap BC_A > -RC_A$$
$$\therefore \ BC_A > |RC_A|$$

A.3 Derivations of trust measures

Theorem 1. *The Nash equilibrium can be expressed in Interdependence terms as* $\tau_B = \frac{1}{2} - \frac{RC_A}{2BC_A}$.

Given $\mathrm{BC}_A = 0.5(A_{11} + A_{22} - A_{12} - A_{21})$ *and* $\mathrm{RC}_A = 0.5(A_{11} + A_{12} - A_{12} - A_{22})$.

$$
\begin{aligned}
\tau_B &= \frac{A_{22} - A_{12}}{A_{11} + A_{22} - A_{12} - A_{21}} \\
&= \frac{A_{22} - A_{12}}{2\mathrm{BC}_A} \\
&= \frac{A_{22} - A_{12}}{2\mathrm{BC}_A} \times \frac{2}{2} \\
&= \frac{2A_{22} - 2A_{12}}{4\mathrm{BC}_A} \\
&= \frac{A_{11} - A_{11} + A_{21} - A_{21} + 2A_{22} - 2A_{12}}{4\mathrm{BC}_A} \\
&= \frac{(A_{11} + A_{22} - A_{12} - A_{21}) - (A_{11} + A_{12} - A_{12} - A_{22})}{4\mathrm{BC}_A} \\
&= \frac{2\mathrm{BC}_A - 2\mathrm{RC}_A}{4\mathrm{BC}_A} \\
\therefore\ \tau_B &= \frac{1}{2} - \frac{\mathrm{RC}_A}{2\mathrm{BC}_A}
\end{aligned}
$$

Theorem 2. *Gottman's TI can be expressed in Interdependence terms as* $\mathrm{TI} = \frac{1}{2} - \frac{\mathrm{RC}_A}{2\mathrm{FC}_A}$.

Given $\mathrm{FC}_A = 0.5(A_{11} + A_{21} - A_{12} - A_{22})$ *and* $\mathrm{RC}_A = 0.5(A_{11} + A_{12} - A_{12} - A_{22})$.

$$
\begin{aligned}
\mathrm{TI} &= \frac{A_{11} - A_{22}}{A_{11} + A_{21} - A_{12} - A_{22}} \\
&= \frac{A_{11} - A_{22}}{2\mathrm{FC}_A} \\
&= \frac{A_{11} - A_{22}}{2\mathrm{FC}_A} \times \frac{2}{2} \\
&= \frac{2A_{11} - 2A_{22}}{4\mathrm{FC}_A} \\
&= \frac{A_{12} - A_{12} + A_{21} - A_{21} + A_{11} + A_{11} - A_{22} - A_{22}}{4\mathrm{FC}_A} \\
&= \frac{(A_{11} + A_{21} - A_{12} - A_{22}) + (A_{11} + A_{12} - A_{12} - A_{22})}{4\mathrm{FC}_A} \\
&= \frac{2\mathrm{FC}_A + 2\mathrm{RC}_A}{4\mathrm{FC}_A} \\
\therefore\ \mathrm{TI} &= \frac{1}{2} + \frac{\mathrm{RC}_A}{2\mathrm{FC}_A}
\end{aligned}
$$