Vanneste, Bart S.; Yoo, Onesun Steve

Article

Performance of trust-based governance

Journal of Organization Design

Provided in Cooperation with:
Organizational Design Community (ODC), Aarhus

Suggested Citation: Vanneste, Bart S.; Yoo, Onesun Steve (2020) : Performance of trust-based governance, Journal of Organization Design, ISSN 2245-408X, Springer, Cham, Vol. 9, Iss. 1, pp. 1-28,
https://doi.org/10.1186/s41469-020-00075-y

This Version is available at:
https://hdl.handle.net/10419/252162

Vanneste and Yoo *Journal of Organization Design* (2020) 9:14
https://doi.org/10.1186/s41469-020-00075-y

Journal of Organization Design
a SpringerOpen Journal

RESEARCH Open Access

Performance of trust-based governance

Bart S. Vanneste* and Onesun Steve Yoo

*Correspondence:
b.vanneste@ucl.ac.uk
[1]UCL School of Management, University College London, Level 38 One Canada Square, Canary Wharf, London, E14 5AA, UK

Abstract

Trust is crucial for the success of interorganizational relationships, yet we lack a clear understanding of when trust-based governance is likely to succeed or fail. This paper explores that topic via a closed-form and a computational analysis of a formal model based on the well-known trust game. We say that trust-based governance performs better in situations where it results in a willingness to be vulnerable with trustworthy others *and* an unwillingness to be vulnerable with untrustworthy others. We find that trust-based governance performs better in situations in which (a) trustworthy and untrustworthy partners exhibit markedly different behavior (high behavioral risk) or (b) the organization is willing to be vulnerable despite doubts concerning the partner's trustworthiness (low trust threshold).

Keywords: Trust-based governance, Trust game, Interorganizational relationships

Introduction

If they are to succeed, firms must work effectively with other firms (Dyer and Singh 1998). Scholars have distinguished between trust-based and contract-based governance (Granovetter 1985; Bradach and Eccles 1989; Gulati 1995; Poppo and Zenger 2002; Puranam and Vanneste 2009; Ryall and Sampson 2009; Cao and Lumineau 2015). In their relationships with suppliers, for instance, Japanese automobile assemblers rely more on trust whereas their American counterparts rely more on contracts (Sako and Helper 1998; Dyer and Chu 2003). Although we have a good understanding of when contract-based governance will succeed or fail—for example, in the presence of contracting problems such as asset specificity (see Williamson (1985); David and Han (2004); Geyskens et al. (2006))—we lack a similar understanding of trust-based governance.

We investigate the situations in which trust-based governance of interorganizational relationships is more likely to succeed or fail. Trust is the willingness to be vulnerable without the ability to monitor or control the other party, behavior that is warranted only if the other is trustworthy (Coleman 1990; Mayer et al. 1995; Rousseau et al. 1998). Governance encompasses "the initiation, termination and ongoing relationship maintenance between a set of parties" (Heide 1994, p. 72). In trust-based governance, trust is the foundation for initiating the interorganizational relationship and a lack of trust is grounds for termination.

We say that trust-based governance performs well if it leads to "optimal" trust (Wicks et al. 1999; Stevens et al. 2015): a willingness to be vulnerable with trustworthy others *and* an unwillingness to be vulnerable with untrustworthy others. Thus, we go beyond the statement that trust-based governance succeeds when the other party is trustworthy. Just as contracting theories seek to identify conditions (e.g., asset specificity, measurement difficulty, technological uncertainty) that affect contract-based governance performance regardless of the specific partners or their characteristics, we must identify analogous conditions for trust-based governance performance without relying on the trustworthiness of potential partners.

To investigate these conditions, we use the classical trust game (Camerer and Weigelt 1988; Berg et al. 1995; Attanasi et al. 2016). Thus, we conceive of trust-based governance as engaging in multiple rounds of trust game. Specifically, we consider an interorganizational relationship in which the focal organization (party A) decides whether to make itself vulnerable to actions of the other organization (party B) by staying in the relationship (Anderson and Weitz 1989; Ganesan 1994; Doney and Cannon 1997). After every period, party A receives an outcome that provides noisy feedback on the trustworthiness of party B (Lioukas and Reuer 2002). Then, A updates its "trustworthiness belief" about B and decides whether to continue or terminate the relationship. Performance of trust-based governance is high in a situation where party A maintains long relationships with trustworthy partners and also quickly breaks relationships with untrustworthy partners. We proceed in four steps. First, we use the trust game to identify potentially relevant situational features that may affect trust-based governance performance. Second, we build a formal model based on the trust game. Third, we undertake closed-form analysis to investigate how each of these situational features affects the performance of trust-based governance. Fourth, we conduct a computational analysis that illustrates and extends the closed-form analysis.

The main finding is that trust-based governance performs better in situations where behavioral risk is high and performs worse when the trust threshold is high. *Behavioral risk* represents the extent to which behavior differs between a trustworthy and an untrustworthy partner (Krishnan et al. 2006). For example, an untrustworthy partner may behave similarly to a trustworthy partner if actions are verifiable but dissimilarly if actions are non-verifiable. The *trust threshold* is the minimum level of perceived trustworthiness at which the trustor would be willing to be vulnerable to the trustee (Gambetta 1988). The trust threshold is high when an organization accepts vulnerability only if the other organization is considered trustworthy; the trust threshold is low if an organization is willing to be vulnerable even when doubting that the other party is trustworthy.

The main contribution is to the literature on the governance of interorganizational relationships—more specifically, three strands within that literature. The first strand addresses the reliance on trust as a governance mechanism. Initial work established this governance mechanism as distinct from such formal mechanisms as contracts (Granovetter 1985; Bradach and Eccles 1989; Ring and Van de 1992). More recent work has focused on the interplay between trust and formal governance (for a meta-analysis, see Cao and Lumineau (2015)). This paper extends that research as follows. We have a good understanding of when formal governance will succeed or fail, yet we lack a similar understanding for the case of trust-based governance, which we aim to address here. The second strand is concerned with the drivers of interorganizational trust (for a meta-analysis, see

Zhong et al. (2017)). This research focuses in particular on how trust evolves during an interorganizational relationship (for a meta-analysis, see Vanneste et al. (2014)). Over the course of a relationship, the other's actions and the relationship's outcomes will influence perceptions of trustworthiness and hence will affect trust. It follows that trust may increase or decrease over time. We build on that research in the following way. In our model, trust and perceived trustworthiness are not set exogenously but instead increase or decrease endogenously based on the relationship's outcomes.

The third strand in this governance literature is that on the consequences of interorganizational trust (for a meta-analysis, see Connelly et al. (2018)). Existing research has investigated the relationship between *perceived trustworthiness* and performance—and also between *trust* and performance—but not between *trust-based governance* and performance, which is the focus of this study. The first investigates how perceptions of trustworthiness affect relationship outcomes (Morgan and Hunt 1994; Dyer and Chu 2003); the second investigates the connection between those outcomes and an organization's willingness to be vulnerable (Zaheer et al. 1998; Gargiulo and Gokhan 2006; Gulati and Nickerson 2008; Stevens et al. 2015). Given that perceived trustworthiness and trust evolve during relationships, we investigate the performance of relationships whose initiation and termination are based on trust and how that performance differs across situations.

Theory

Trustworthiness, trust, and trust-based governance

Trust is viewed similarly across the domains of psychology, sociology, and economics (Rousseau et al. 1998). It is widely understood as "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, irrespective of the ability to monitor or control that other party" (Mayer et al. 1995, p. 712). Trust is not trustworthiness (Colquitt et al. 2007; McEvily and Tortoriello 2011; Özer et al. 2018). Whereas trust says something about the trusting party, trustworthiness is an attribute of the trusted party. Perceived trustworthiness involves both parties: a perception of the trusting party about the trusted party. Interorganizational trust is when the parties are organizations (Vanneste 2016).

At its core, trust-based governance implies that trust forms the basis for initiating an interorganizational relationship and that a lack of trust is the basis for terminating the relationship (cf. Heide (1994)). Two key features of trust-based governance are as follows. First, initiating an interorganizational relationship requires the acceptance of vulnerability (Mayer et al. 1995); conversely, deciding not to initiate a relationship implies rejecting vulnerability. Vulnerability occurs because a relationship with a trustworthy partner is better than no relationship, which in turn is better than a relationship with an untrustworthy partner. The second key feature is that in the event of termination, partners have no recourse. Termination occurs when trust in the partner has been lost, typically after a poor outcome (Mayer et al. 1995).

These features differ from those of formal governance mechanism such as contracts (Williamson 1975; Williamson 1985; David and Han 2004). In the first place, a goal of contracts is neither to accept nor to reject but to reduce vulnerability by establishing safeguards (Argyres et al. 2007; Ryall and Sampson 2009). The idea is that since one cannot easily distinguish trustworthy from untrustworthy partners ex ante, a contract can

provide protection should the other party turn out to be untrustworthy. Second, a contract allows for legal recourse. Even when recourse is incomplete, a contract serves as the basis for resolving disputes (Lumineau and Oxley 2012). It is for those cases of incomplete recourse that scholars have suggested alternative mechanisms, such as hierarchy and trust (Williamson 1985; Bradach and Eccles 1989).

We can illustrate the different trust constructs—trustworthiness, trust, and trust-based governance—using the so-called trust game (Camerer and Weigelt 1988; Berg et al. 1995). An example from Bohnet and Zeckhauser (2004) is illustrated by Fig. 1. The payoffs are given in parentheses, where the first number of each pair corresponds to the trustor (party A) and the second to the trustee (party B). Parties A and B each start off with, say, 10 units of currency. It is A's decision whether the game should terminate or continue. If A terminates, then A and B each keep their 10 units; this is the "(10, 10)" outcome, which reflects the parties' outside options, that is shown in the figure. If instead A continues, then the money available increases by a fixed factor (here, 1.5); now, the total amount has become 30 units. It is then B's decision how this amount should be divided: party A can be given either half (15 units in the figure's example) or less than half (8 units). Thus, party A's continuing the game is an indicator of trust because it reflects a willingness to be vulnerable, since A could end up with less than at the start. At the same time, party B's returning half is an indicator of trustworthiness because its own payoff would be maximized by instead returning as little as possible. Hence, A's belief that B would choose the high outcome for A is an indicator of perceived trustworthiness.

Because the trust game concisely captures the main elements of trust, it has proved useful for studying trust both theoretically (e.g., Dasgupta (1988); Attanasi et al. (2016)) and empirically (for reviews, see Camerer (2003); Johnson and Mislin (2011)). For the same reason, we use the trust game to study trust as a governance mechanism (Arrow 1974; Bradach and Eccles 1989; Zaheer et al. 1998; Poppo and Zenger 2002; Puranam and Vanneste 2009). In particular, we conceive of trust-based governance as engaging in multiple rounds of trust game in which the decision to continue or terminate the relationship is based on the presence or absence of trust, see Table 1.

Features of a trust situation

The goal of any model is to gain insight by reducing complexity (Lave and March 1993). We therefore anticipate that the trust game will clarify some dynamics of trust-based governance even as it (necessarily) fails to capture all of the empirical phenomenon's richness.

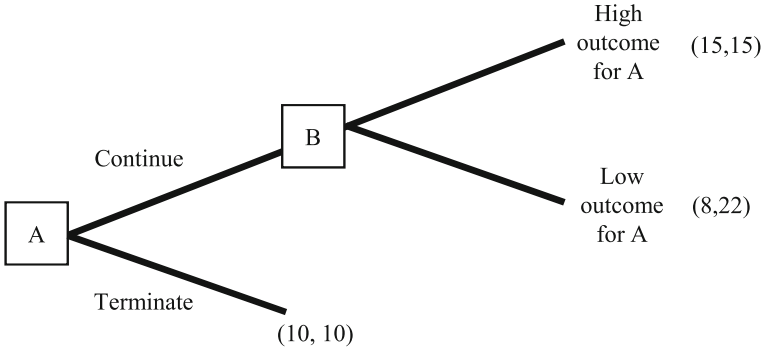

Fig. 1 Trust game: example

Table 1 Trust constructs in a trust game

Construct	Trust game
Trustworthiness	Whether B would choose the high outcome for A
Perceived trustworthiness	A's belief that B would choose the high outcome for A
Trust	Whether A would choose "continue"
Trust-based governance	Repeatedly engaging in a trust game

Thus, in line with others who have applied game structures to analyze interorganizational relationships (e.g., Heide and Miner (1992); Zeng and Chen (2003); Chatain and Zemsky (2011); Özer et al. (2011)), we follow this approach of simplifying to gain insight.

A parameterization of the trust game suggests five features of a trust situation that may affect trust-based governance performance: outcome risk, behavioral risk, self-serving norm, value capture, and value creation. We denote each of these by a single parameter in the trust game; see Fig. 2. To illustrate the features involved, we use a simple and generic example. Organization A (the trustor) buys a product from organization B (the trustee), which may or may not be trustworthy. The product is either of high quality (i.e., a high outcome for A) or of low quality (a low outcome). A low-quality product is more likely with an untrustworthy than a trustworthy organization B. Here, the outside options consist of organization A buying from a party other than B and, likewise, organization B supplying a party other than A (for a summary, see Table 2).

The five features of a trust situation relate to this example as follows. First, outcome risk refers to the value difference between the high- and low-quality product. Second, behavioral risk indicates the difference in the probability of a low-quality product between an untrustworthy and a trustworthy organization B. Third, self-serving norm is the extent to which both an untrustworthy and a trustworthy organizations B deliver a low-quality product. Fourth, value capture is the extent to which organization A instead of B captures the gains from their trade. Fifth, value creation indicates how much organization A and B jointly benefit from trading with each other instead of with other parties.

Outcome risk. Trust is meaningful only in the presence of risk (Bhattacharya et al. 1998; Mayer et al. 1995). In the trust game, there is risk because the high and low outcomes are different (with the outside option lying in between).

Although outcome risk is present in many situations, its extent will differ. Outcome risk can have multiple causes. For example, a key driver of risk in transaction costs economics

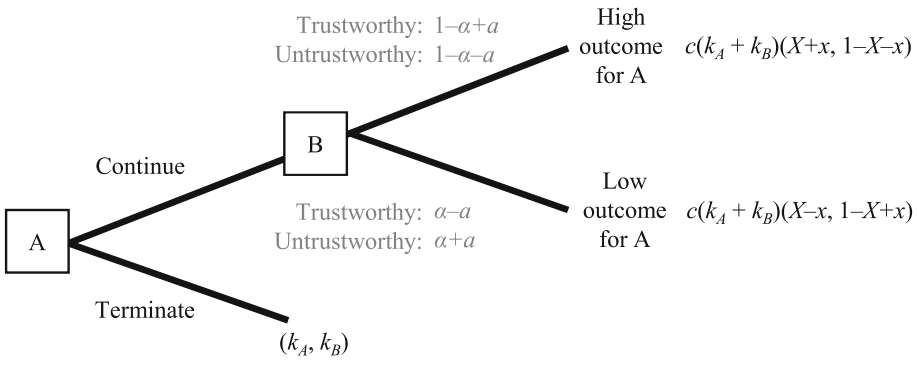

Fig. 2 Trust game: notation

Table 2 Features of a trust situation

Feature	Example	
	Theoretical	Empirical
Outcome risk (x)	Relationship-specific assets (Williamson 1985); task criticality (Kiggundu 1981)	The difference between high and low quality
Behavioral risk (a)	Binding contracts (Malhotra and Murnighan 2002)	The product's quality can be verified by an outside party
Self-serving norm (α)	Cultural context (Miller 1999)	A higher frequency of low-quality products from party B
Value capture (X)	Industry competitiveness (Porter 1980); uniqueness of relationship (Brandenburger and Stuart 1996)	Another supplier provides a similar product
Value creation (c)	Gains from trade (Jacobides and Hitt 2005)	Buyer and supplier specialize in different activities

is asset specificity (Williamson 1975; Williamson 1985; David and Han 2004). If investments are customized to the relationship, then its potential benefit is high but the other party may be able to extract value from that investment (e.g., by threatening to walk away). Outcome risk may also stem from how critical the task is (Kiggundu 1981). For example, a bank becomes more vulnerable by outsourcing its information technology systems than by outsourcing its advertising. Consider also the earlier example where organization B supplying a product to organization A. Outcome risk will be low if a low-quality product is close in value to a high-quality one, whereas outcome risk will be high if a low-quality product is much worse than a high-quality one.

We use x (with $0 < x \leq 0.5$) to signify a situation's outcome risk. In Fig. 2, outcome risk captures the difference between the high and low outcomes. In the high outcome, party A gains x (at the expense of B, which looses that x); in the low outcome, A looses x (to the benefit of B, which gains that x). Note that these "high" and "low" labels are from the perspective of party A. Because x adds to the high outcome but subtracts from the low outcome, outcome risk increases with x.

Behavioral risk. A risky situation is one with outcome risk. A trust situation is one in which that outcome risk depends on the other's actions (Bohnet and Zeckhauser 2004), i.e., behavioral risk is present (Krishnan et al. 2006). In the trust game, there is behavioral risk because a trustworthy and an untrustworthy partner may return different outcomes.

In general, a trustworthy partner will share gains more equitably than will an untrustworthy one. The extent to which this generalization holds does vary across situations, however. For example, Malhotra and Murnighan (2002) argue that binding contracts induce untrustworthy partners to behave more like trustworthy partners (because otherwise they will be penalized). In the earlier example, suppose organization B supplies a product and guarantees its quality in a binding contract; then, the interests of even an untrustworthy B are probably better served by supplying a high-quality than a low-quality product. In this case, trustworthiness matters little and behavioral risk is low. In the absence of binding contracts, it could be that low-quality products often arise with an untrustworthy B but rarely with a trustworthy B. In such case, behavioral risk is high.

We use a (with $0 < a \leq 0.5$) to denote a situation's behavioral risk. In Fig. 2, behavioral risk captures the difference in the probability of a low outcome between an untrustworthy

and a trustworthy partner. For an untrustworthy partner, the probability of a low outcome increases by a relative to a baseline (we discuss the baseline next). For a trustworthy partner, that probability decreases by a relative to the baseline. A higher probability of a low outcome must mean a lower probability of a high outcome. Thus, the probability of a high outcome decreases by a for an untrustworthy partner and increases by a for a trustworthy partner, relative to a baseline. Because a amplifies the difference between an untrustworthy and a trustworthy partner, behavioral risk increases with a.

Self-serving norm. The preceding feature indicates, for a given situation, how much the probability of a low outcome *differs* between an untrustworthy and trustworthy other. In contrast, this feature indicates, for that situation, whether that probability has a low or high *baseline*. In the trust game, a low outcome is always more likely with an untrustworthy than trustworthy other but the baseline probability of a low outcome will differ across situations.

Trust is based on an expectation that the trustee will take an action of importance to the trustor (Mayer et al. 1995; Bhattacharya et al. 1998). Yet that expectation depends on the situation's prevailing behavioral norms. In some parts of the world, for instance, a norm of self-serving might be more prevalent than elsewhere (Miller 1999). If so, then it is less likely that a trustee will return a high outcome, regardless of whether the other is trustworthy or untrustworthy. In the earlier example, a low-quality product is always more likely with an untrustworthy than trustworthy organization B. However, it can be more tempting to produce low-quality in some than other contexts, irrespective of whether B is trustworthy or untrustworthy.

We use α (with $0 \leq \alpha - a, \alpha + a \leq 1$, and $\alpha, a > 0$) to denote the self-serving norm. In Fig. 2, the self-serving norm captures how likely the low instead of the high outcome is. Ignoring the trustee, the baseline probability of a low outcome is α. Taking into account the trustee, the probability of a low outcome is $\alpha + a$ for an untrustworthy party B and $\alpha - a$ for a trustworthy B. This formulation indicates that the outcome probabilities depend on the situation's self-serving norm (α) and also on the risk pertaining to the exchange partner's behavior (a). Under this specification, the low-outcome probability can be 0 (as when a trustworthy B never produces the low outcome for A) or 1 (as when an untrustworthy B always produces that low outcome). For studying the performance of trust-based governance, the more interesting cases are those in which the outcome probabilities are neither 0 nor 1, because then a single outcome can immediately reveal the other's trustworthiness. Thus, for example, having a trustworthy partner need not result in a high outcome for party A. The reason could be that intentions translate only stochastically into outcomes—say, because of a "trembling hand" (Selten 1975) or technological uncertainty. Alternatively, it could be that even trustworthy people sometimes stray and that the frequency of straying depends on the situation. Our model focuses on the latter interpretation.

Value capture. A key goal in an interorganizational relationship is value capture (Porter 1980; Dyer and Singh 1998; Bidwell and Fernandez-Mateo 2010; Elfenbein and Zenger 2014). In the trust game, value captures refers to the split in payoffs between trustor and trustee.

The extent of value that can be captured will vary across trust situations. Thus, an organization might be able to capture most of the resulting value in one situation but only

a small part in another situation. Value capture is affected by industry competitiveness (Porter 1980) and by the uniqueness of a partner's contribution (Brandenburger and Stuart 1996). For instance, if no other potential partner can replicate the contribution, then the focal partner will probably capture most of the value. Recall that in our setup, organization B supplies a product to organization A. But suppose there is another organization that can supply a similar product; in that case, A will capture more value than if there were no competitors able to supply an equivalent product.

We use X (with $0 < X - x, X + x \leq 1$, and $X, x > 0$) to denote the proportion of value that can be captured by party A. In Fig. 2, X is included in both the high outcome and the low outcome for party A. This means that a higher X corresponds to a greater portion of the value captured by party A, irrespective of the specific outcome. Likewise, $1 - X$ is included in the payoffs for party B. If A captures a higher proportion, then B captures a lower proportion.

Value capture (X) is interpreted as the average of the low and high outcome while outcome risk (x) is interpreted as the dispersion of the low and high outcome. In the Bohnet and Zeckhauser (2004) trust game (see Fig. 1), $X = 0.385$ and $x = 0.115$; hence, A obtains either $0.385 + 0.115 = 50\%$ (high outcome) or $0.385 - 0.115 = 27\%$ (low outcome).

Value creation. Value can be captured only when it is created. The creation of value is a fundamental property of an interorganizational relationship (Zajac and Olsen 1993; Dyer and Singh 1998) and also of interorganizational trust; the latter follows because trust is relevant precisely when the other organization can create value with, but can also exploit, the focal organization (Krishnan et al. 2006; Gulati and Nickerson 2008). In the trust game, value creation is possible because the parties can jointly benefit more from interacting with each than choosing their outside options.

Value creation is necessary for trust, but the level of value creation will vary across situations; thus, one situation may create little value even as another creates a lot of value. The key driver of value creation is the potential gains from trade (Jacobides and Hitt 2005). Suppose, for instance, that organization A specializes in assembling and organization B specializes in manufacturing. In this scenario, more value is created if organization A buys products from organization B instead of from another organization that also specializes in assembling.

We use c (> 1) to denote the level of value creation; see Fig. 2. The outside options for organization A and B are denoted by, respectively, k_A and k_B (≥ 0). If A decides to engage in a relationship with B, then the amount to be shared is c multiplied by the sum of k_A and k_B. Because $0 < X - x, X + x \leq 1$, value creation is a scaling factor that ensures that the relationship can be more attractive than the outside options. In the Bohnet and Zeckhauser (2004) trust game (see Fig. 1), the outside option for each party was $k_A = k_B = 10$ and an exchange between these parties created value $c = 1.5$ times the sum of their respective outside options.

Model

We consider multiple rounds of the trust game between two organizations, A and B. We focus on A's decision, and B's decision is captured through self-serving norm (α) and behavioral risk (a). That is, we consider a multi-period decision-theoretic model involving a single player, not a repeated game-theoretic model involving two players often used

in the "shadow of the future" literature (Axelrod 1984). By using two types of partners (trustworthy or untrustworthy) instead of only one (self-interested), we can focus on situations where trustworthiness is relevant as opposed to where shadow of the future is the key mechanism.[1] At the beginning of each period of a relationship, organization A must decide whether to make itself vulnerable to the actions of organization B—that is, whether to engage in a relationship with B or instead take its outside option.

The timing of the model is as follows. Organization A decides whether to terminate or continue the relationship. If A terminates the relationship, then both A and B receive their outside options forever after. If A continues, then organization B captures part of the value created and A receives the remainder. So with reference to the parameters shown in Fig. 2, we can say that after each period in a relationship, A receives an uncertain share ($\tilde{X}$) of the total value created ($c(k_A + k_B)$). Organization A's share is either low ($X - x$) or high ($X + x$), and it depends on organization B's trustworthiness.[2] A high share is more likely in the case of a trustworthy than of an untrustworthy B. However, if B is trustworthy, a high share is not guaranteed. Conversely, an untrustworthy B does not imply a low share to organization A. Formally, we write:

$$\text{if B is trustworthy, then } \tilde{X} = \begin{cases} X + x \text{ with probability } 1 - (\alpha - a), \\ X - x \text{ with probability } \alpha - a; \end{cases}$$

$$\text{if B is untrustworthy, then } \tilde{X} = \begin{cases} X + x \text{ with probability } 1 - (\alpha + a), \\ X - x \text{ with probability } \alpha + a. \end{cases}$$

A trust situation is one in which party A would be better-off making itself vulnerable when party B is trustworthy and would be worse-off otherwise (Coleman 1990; Berg et al. 1995). We therefore impose the following condition:

$$\text{if B is trustworthy, then } c(k_A + k_B)\mathbb{E}[\tilde{X}] > k_A;$$

$$\text{if B is untrustworthy, then } c(k_A + k_B)\mathbb{E}[\tilde{X}] < k_A.$$

During each period of a relationship, organization A learns about organization B's trustworthiness type. This acquired knowledge affects whether A will continue to make itself vulnerable in the next period or instead will terminate the relationship. Organization A's share is a noisy signal because it depends not only on B's trustworthiness but also on randomness; if the signal was not noisy, then establishing the other's trustworthiness would be straightforward. Let $p \in [0, 1]$ denote organization A's belief that firm B is trustworthy. Following rational modeling convention, we use Bayes' rule to describe how this belief changes after each period. Less rational updating approaches (e.g., reinforcement learning) would yield the same results provided that updating is in the direction of the outcome—as when a high share for A strengthens its belief that B is trustworthy. Hence, we can write:

$$\text{for share } X + x, \text{ belief increases to } \frac{p(1 - (\alpha - a))}{p(1 - (\alpha - a)) + (1 - p)(1 - (\alpha + a))};$$

[1] Observe that both α and a in reality may depend on value creation (c), value capture (X), outcome risk (x), and the outside options (k_A and k_B). The implication is that some combinations of α, a, X, x, c, k_A, and k_B are less likely to occur than others. Our approach is to provide results for all combinations regardless of their naturally occurring frequencies.

[2] For simplicity and ease of comparison with the literature on trust games, we treat outcomes as being discrete. Yet we could, without changing the model's basic structure, rewrite it to accommodate continuous outcomes as follows: a trustworthy B provides continuous outcomes with first and second moments (μ_{B1}, σ_{B1}) that are first-order stochastically dominant over the continuous outcomes of an untrustworthy B (μ_{B0}, σ_{B0}); then, in each period, the expected outcome for A is either μ_{B1} or μ_{B0}.

$$\text{for share } X - x, \text{ belief decreases to } \frac{p(\alpha - a)}{p(\alpha - a) + (1 - p)(\alpha + a)}. \quad (1)$$

Organization A's objective is to maximize the sum of discounted outcomes (V) given its belief about organization B's trustworthiness. Rewards from the next period are discounted by $\delta < 1$, and we use $\tilde{p}$ to denote the uncertain belief in the next period. Organization A's choice is to terminate or continue the relationship, as expressed formally by the following infinite-horizon Bellman equation:[3]

$$V(p) = \max \left\{ \frac{k_A}{1 - \delta}, \mathbb{E}[\, c(k_A + k_B)\tilde{X} + \delta V(\tilde{p})] \right\}. \quad (2)$$

Here, we present the scenario in which organization A is farsighted and has an infinite horizon, but our results hold also if A is shortsighted and considers only two periods.[4]

Results

Here, in the main text, we shall focus on intuitive explanations of the results using the example employed previously: organization B, which may be either trustworthy or untrustworthy, supplies organization A with a product of either low or high quality. All proofs are given in the Appendix.

Trust threshold

Organization A's key choice is whether to terminate or continue the relationship with organization B. The following lemma presents the solution structure to Eq. (2).

Lemma 1 (trust threshold) *There exists a trust threshold* (p^*) *such that organization A is willing to continue the relationship (i.e., to remain vulnerable) if and only if the trustworthiness belief* (p) *remains above that threshold* (*i.e., iff* $p > p^*$).

At any given time, A can assess whether or not to continue the relationship by checking whether its belief about B's trustworthiness is above a fixed threshold. This trustworthiness belief depends on the past relationship between organizations A and B. It is noteworthy, however, that the threshold does not, and depends only on situational features.[5] In other words, regardless of how many low- or high-quality products it has received from B in the past, A can terminate the relationship as soon as its belief about B's trustworthiness falls below a fixed level.

A high p^* reflects a situation that requires keeping a "short leash" on organization B. For example, organization A will continue the relationship with organization B only if it is perceived as highly trustworthy. In contrast, a low p^* reflects a situation that allows for more relaxed approach to organization B. Organization A will continue the relationship with organization B as long as it is perceived as moderately trustworthy.

The following proposition illustrates how the situational features influence the trust threshold.

[3] The sum of discounted outcomes with $t > 1$ periods remaining is given by $V_t(p) = \max\{k_A(1 + \delta + \cdots + \delta^{t-1}), \mathbb{E}[\, c(k_A + k_B)\tilde{X} + \delta V_{t-1}(\tilde{p})]\}$, where $V_0(p) = 0$. The term $V(p)$ is the sum's infinite-horizon extension ($t \to \infty$), which exists because the immediate rewards are bounded (Ross 1983).

[4] One could interpret organization B's decision as exhibiting either farsighted or myopic behavior. In particular, its probabilistic response can be viewed as the result of looking n periods ahead in a stationary manner. The case of $n \to \infty$ corresponds to farsighted behavior and that of $n = 0$ to myopic behavior.

[5] In the finite-horizon case, the threshold does not depend on the past relationship but does vary with the number of periods remaining: when there are fewer such periods, the threshold is higher.

Proposition 1 (**comparative statics of the trust threshold**)

(i) Outcome risk (x) *can either decrease or increase the trust threshold:*

- *If* $\alpha + a < 0.5$, *then trust threshold decreases;*
- *If* $\alpha - a > 0.5$, *then trust threshold increases;*
- *If* $\alpha - a < 0.5 < \alpha + a$, *then trust threshold decreases if and only if* $k_A > c(k_A + k_B)X$.

(ii) Behavioral risk (a) *can either decrease or increase the trust threshold:*

- *If* $k_A > c(k_A + k_B)(X + (1 - 2\alpha)x)$, *then trust threshold decreases;*
- *If* $k_A < c(k_A + k_B)(X + (1 - 2\alpha)x)$, *then trust threshold decreases or increases.*

(iii) Self-serving norm (α) *increases the trust threshold.*

(iv) Value capture (X) *decreases the trust threshold.*

(v) Value creation (c) *decreases the trust threshold.*

While the effect of self-serving norm, value capture, and value creation is intuitive (parts (iii), (iv), and (v), respectively), it shows that the effects of outcome risk and behavioral risk (parts (i) and (ii), respectively) are complicated by its interaction with other parameters.

We proceed as follows. First, we define a metric for assessing the performance of trust-based governance across situations. Second, we employ closed-form analysis to examine how the features of those situations affect that performance. Finally, we undertake a computational analysis.

Performance metric

The trust threshold just defined is "optimal" in the sense that one cannot do better in expectation in the given relationship. Yet because it is unknown whether the other is trustworthy, mistakes will be made—either continuing relationships with untrustworthy partners or terminating relationships with trustworthy ones. Some situations are more mistake prone than others. Our goal is to distinguish between situations in which such mistakes are more versus less likely to occur. The performance metric we use is as follows.

Definition 1 (performance) *Trust-based governance performs better in one versus another situation if, and only if, the expected relationship duration with a trustworthy partner increases and that with an untrustworthy partner decreases.*

In a trust situation, organization A would like ideally to stay in a relationship with a trustworthy organization B and not to begin a relationship with an untrustworthy B. When comparing two trust situations, we say that trust-based governance performs better in cases where A more nearly approaches this ideal: longer expected relationships with trustworthy partners *and* shorter expected relationships with untrustworthy partners. We say that trust-based governance performs worse in the opposite situation—that is, wherein organization A has shorter expected relationships with trustworthy partners *and* longer expected relationships with untrustworthy ones. In the event that trust-based governance results in longer expected relationships with both trustworthy and untrustworthy partners, we say that it has performed neither better nor worse.

We acknowledge that the performance measure is related to but differs from the usual economic value created or captured from an ongoing relationship. It is related because the value function $V(p)$ increases if the expected relationship duration with a trustworthy partner increases and that with an untrustworthy partner decreases. It differs because a setting with low $V(p)$ does not imply that trust-based governance does not work well. Our theoretical question is about understanding when trust-based governance performs well and when it does not. Examining this question requires a performance metric that cuts along multiple situations. Moreover, in our setting, we use value capture and value creation as input parameters rather than output of the trust-based relationship. Our definition of performance thus focuses on the decision concerning whether or not to trust.

Closed-form analysis

Here, we offer a closed-form analysis of how each of the situational features—outcome risk, behavioral risk, self-serving norm, value capture, and value creation (x, a, α, X, and c, respectively)—affects the performance of trust-based governance. The derivation of the result is complex, so we refer readers to Section B of the Appendix for detailed steps.

Proposition 2 (performance of trust-based governance)
(i) Increase in a can enhance, but not hurt, the performance of trust-based governance.
(ii) Increase in all other parameters (x, α, X, and c) neither enhances nor hurts the performance of trust-based governance.

According to our definition, trust-based governance performs better if two conditions are satisfied: the expected relationship duration with a trustworthy partner increases and that with an untrustworthy partner decreases. Most of the situational features are such that varying them implies that one, but not the other, condition holds. In that case, trust-based governance performs neither better nor worse. Specifically, changes in outcome risk, self-serving norm, value capture, or value creation have no effect on the performance of trust-based governance (this is part (ii) of the proposition).

In contrast, an increase in the behavioral risk can improve—but not hurt—the performance of trust-based governance (this is part (i)). Two underlying effects explain this result. First, greater behavioral risk induces a trustworthy B to supply high-quality products more frequently and induces an untrustworthy B to supply them less frequently. As a result, A learns about B more quickly. Hence, the expected relationship duration lengthens with a trustworthy B and shortens with an untrustworthy B. In isolation, this effect enhances the performance of trust-based governance. Changes in outcome risk, value capture, or value creation have no effect on duration while a greater self-serving norm always lengthens the duration, the extent to which depends on behavioral risk (see Lemma A-6 in Appendix).

Second, greater behavioral risk can either increase or decrease the trust threshold. In particular, a lower (resp. higher) trust threshold increases (resp. decreases) the expected relationship duration for both a trustworthy B and an untrustworthy B. So in isolation, this effect neither enhances nor diminishes the performance of trust-based governance (Proposition 1). Nonetheless, that performance may improve (but not decline) in response to the combination of these two effects, depending on their relative strengths.

In sum of all the situational features, behavioral risk is the most influential in determining the success or failure of trust-based governance.

Computational analysis

The closed-form analysis of the performance of trust-based governance is both ceteris paribus (changing one variable at a time) and marginal (changing that one variable minimally). To derive results for simultaneous and substantial changes, we employ a simulation and construct a "classification tree." When combined with the closed-form analysis, this approach yields a more complete account of when trust-based governance performs better.

The simulation follows the setup of our formal model, in which c, X, x, α, and a signify (respectively) value creation, value capture, outcome risk, the self-serving norm, and behavioral risk. We analyze a wide spectrum of situations. More specifically, $X, \alpha \in [0.1, 0.9]$ vary with a step size of 0.1, $x, a \in [0.05, 0.45]$ vary with a step size of 0.05, and $c \in [1.1, 2]$ vary with a step size of 0.1. The outside option for party B is set to $k_B = 1$ while the outside option for party A ranges between one tenth and ten times that value ($k_A \in [0.1, 0.2, \ldots, 1, 2, \ldots, 10]$). The discount factor is set to $\delta = 0.9$. We consider only those combinations for which trust matters—that is, cases in which A would continue the relationship if it knew that organization B were trustworthy yet would exit the relationship if it knew that B were untrustworthy.[6]

These combinations produce a total of 35,664 different trust situations. Whether organization A indeed starts a relationship depends on its belief about organization B. We analyze an A that initially perceives B as being trustworthy with probability $p = 0.5$, yielding 28,085 situations in which A would prefer to begin a relationship.[7] For each of these situations, we compute the trust threshold p^* by solving the Bellman equation (2) using the standard method of value iteration (Ross 1983). We simulate 200 relationships—100 in which B is untrustworthy and 100 in which B is trustworthy—for a maximum of 1,000,000 periods per relationship.

Figure 3 plots the performance of trust-based governance for a randomly selected 3% of the simulated data. Each dot represents one situation. The average number of periods with an untrustworthy (resp. trustworthy) B is marked on the horizontal (resp. vertical) axis. The median for all data is about 4 periods with an untrustworthy B and 16,176 periods with a trustworthy B, as indicated by the vertical and horizontal line (respectively). For each situation, it would be best for A to always preserve its relationship with a trustworthy B and to terminate its relationship with an untrustworthy B. When benchmarked against this ideal, the performance of trust-based governance varies markedly. It performs better in situations in the upper left quadrant (Q_{I}) and worse in the lower right quadrant (Q_{IV}) while performing at levels in between in the other two quadrants (Q_{II} and Q_{III}). Thus, the performance of trust-based governance depends strongly on the situation.

To identify the type of situations in which trust-based governance performs better, we use a classification tree. This method divides observations (here, situations) into mutually exclusive and collectively exhaustive groups and assigns each group to a categorical

[6] We also ensured that $0 \leq X - x < X + x \leq 1$, that $0 \leq \alpha - a < \alpha + a \leq 1$, and that either of B's outcomes from the relationship is better than its outside option.

[7] We also ran simulations using $p = 0.9$. Doing so led to more situations (34,742) in which organization A wants to enter a relationship. The other results are similar to those reported here.

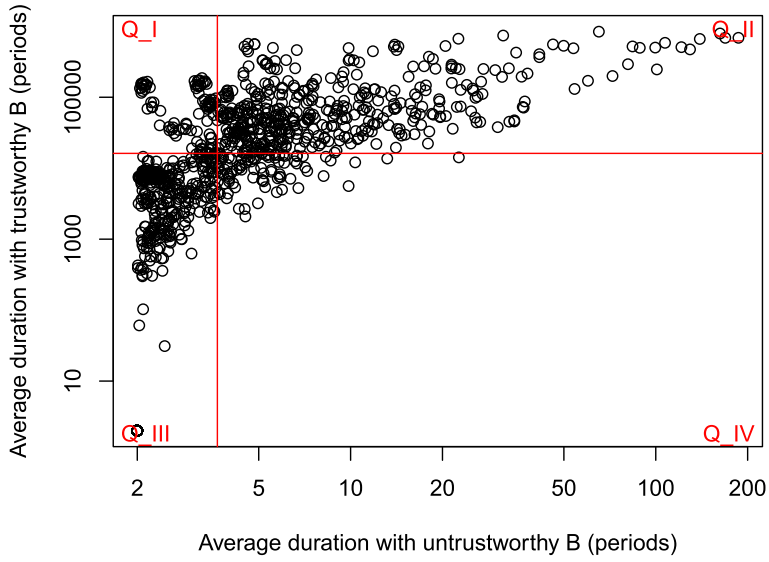

Fig. 3 Performance of trust-based governance for a randomly selected 3% of simulated data

outcome (here, one of the performance quadrants in Fig. 3). The structure of a classification tree is similar to that of a decision tree. It has multiple internal nodes (see Fig. 4). At each internal node, the observations are split into two groups according to a simple rule involving one feature and one cutoff point (e.g., $a < 0.325$). For details on how the splits are chosen, see Friedman et al. (2001). In essence, with this method, the splits are chosen sequentially. For each split, all possible features and cutoff points are considered. The chosen split is the one that most reduces "impurity"—in other words, the one that leads to groups consisting mainly of observations with the same outcome. Branches connect the internal nodes leading to terminal nodes or leaves. For a given leaf, the most

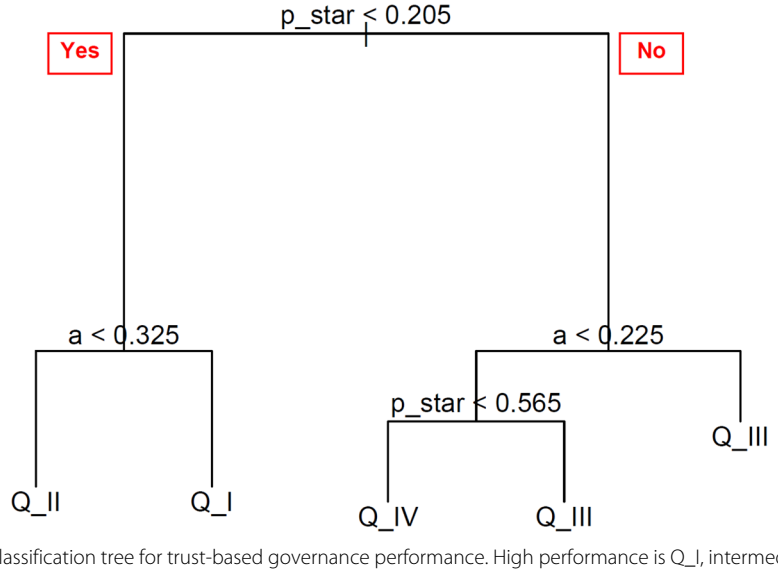

Fig. 4 Classification tree for trust-based governance performance. High performance is Q_I, intermediate performance is Q_II and Q_III, and low performance is Q_IV; *a* is the behavioral risk and *p*_star is the trust threshold

commonly occurring performance quadrant is taken to be indicative of the resulting group.

For our purposes, the benefits of a classification tree are fourfold. First, and most importantly, a classification tree enables us to compare groups of situations that differ substantially (i.e., not marginal changes) and on multiple features (i.e., not ceteris paribus). Second, it is a nonparametric method that can handle complex relationships (e.g., nonlinearities and interactions)—as our closed-form analysis suggests is the case here. Third, a classification tree automatically selects features, i.e., it includes the variables that matter strongly and excludes those that matter little. For example, recall from the closed-form analysis that we expect behavioral risk to matter more and some of the other situational features less. Fourth, the resulting classification is easy to interpret.

The dependent variable is the focal performance quadrant in Fig. 3. The following independent variables are included for consideration: all situational features (c, X, x, α, and a), A's outside option k_A (B's outside option k_B is fixed), and the trust threshold (p^*). The last measure is a nonlinear combination of the situational variables. The method determines whether that measure has predictive power beyond that of the individual situational variables.

Training of the classification tree is based on a random selection of 70% of the situations. The most common outcome (Q_{II} or intermediate performance) occurs more than four times as frequently as the least common outcome (Q_I or high performance). To give each outcome equal importance during the estimation, we randomly select a subset of the observations such that all other outcomes occur as often as the least common outcome; thus, the training sample contains $4 \times 2404 = 9616$ situations.

Figure 4 depicts the trained classification tree. Left branches indicate "yes" (the inequality is satisfied) and right branches indicate "no" (the inequality is not satisfied). This classification tree indicates that trust-based governance performs best (Q_I or high performance) in situations where behavioral risk is high ($a \geq 0.325$) yet trust threshold is low ($p^* < 0.205$); trust-based governance performs worst (Q_{IV} or low performance) in opposite situations—that is, when $a < 0.225$ and $0.205 \leq p^* < 0.565$ (note that $p^* \geq 0.565$ for only 11% of all simulated data).

Although the tree has a limited number of branches and only five leaves,[8] this simple model predicts well: it correctly classifies 79% of the situations from the test sample (which consists of the 30% of situations *not* in the initial selection for the training sample). Furthermore, we observe that even a tree extended to 15 internal nodes uses the trust threshold and behavioral risk for determining all but one of the splits. These results lead us to conclude that behavioral risk and the trust threshold are the most important features.

So in line with the closed-form analysis, the computational analysis finds that increased behavioral risk a leads to trust-based governance performing better. The simulation reveals that also a lower trust threshold is associated with higher performance of trust-based governance. Combining the closed-form results with this computational results yields a more complete picture of when trust-based governance performs better.

We highlight two additional surprising elements of these findings. First, recall that relationship duration depends on the trust threshold's level and on how soon it is breached.

[8] This number of leaves was selected based on $K = 10$ fold cross-validation using the training sample.

So all else equal, lower trust thresholds should lead to longer relationships and quicker updating should lead to shorter relationships (i.e., high behavioral risk). In line with this thinking, one might well suppose that the performance of trust-based governance is similar in these two situations: (i) low behavioral risk combined with a high trust threshold and (ii) high behavioral risk combined with a low trust threshold. To the contrary, we find that the performance of trust-based governance is low in case (i) and high in case (ii). Second, where self-serving is the norm, it seems reasonable to expect low performance of trust-based governance. However, we do not find self-serving norms to matter much for the trust-based governance's performance.

Discussion

We investigate in which situations trust-based governance of interorganizational relationships performs well and in which it does not. Our basic premise is that some situations lend themselves better for sustaining relationships with trustworthy partners than other situations do. One implication is that across trust situations, we should observe substantial variation in the average duration of interoganizational relationships. To explore this implication, we compiled data from all samples included in a trust meta-analysis (Vanneste et al. 2014) that reported an average duration for interorganizational relationships ($k = 19$ samples reported in 16 papers, $N = 3,201$). The data include relationships from different industries (e.g., automotive, textile, and electronics) and countries (e.g., USA, Korea, and Turkey). We comment on two aspects of these data (see Fig. 5). First, the sample size-weighted average duration of an interorganizational relationship is substantial (14.7 years). Second, the weighted standard deviation is also substantial (9.7 years). If researchers picked settings in which trust was important (e.g., where the decision to continue a relationship depended in part on the level of trust between exchange partners), then this finding indicates that the performance of trust-based governance may well depend on the situation.

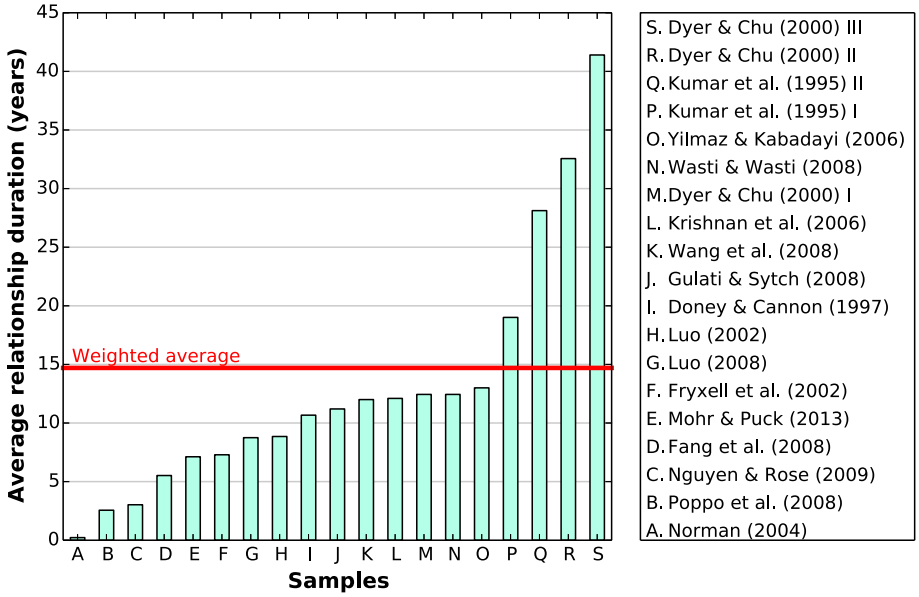

Fig. 5 Average duration of interorganizational relationships in trust studies

Our investigation aims to distinguish between situations where trust-based governance performs well and where it does not. We use the trust game to identify five features of a situation—value creation, value capture, outcome risk, self-serving norm, and behavioral risk—and then construct a formal model to investigate how these features affect the performance of trust-based governance. Using both closed-form and computational analysis, we find that out of these five, behavioral risk is especially important: trust-based governance performs better in situations characterized by high levels of behavioral risk. We also find that the trust threshold, which is a nonlinear combination of the five situational features, matters: more specifically, trust-based governance performs worse in situations where the trust threshold is high.

These key results bear implications for the governance of interorganizational relationships. Dyer and Singh (1998) note that because competitive advantage relies in large part on such relationships, it is critical that their governance be effective. It has been more than three decades since trust was identified as a form of governance distinct from contracts (Granovetter 1985; Bradach and Eccles 1989). A vast literature has since investigated the nature and role of interorganizational trust (for reviews, see McEvily and Zaheer (2006); Zaheer and Harris (2006)). At the same time, the literature on contracts has come to be dominated by Williamson's (1991) "discriminating alignment" hypothesis, whereby contracts work better in some situations than in others (Geyskens et al. 2006). We propose a similar discriminating alignment hypothesis for trust-based governance: trust works better in some situations than in others.

Both the closed-form and the computational analyses suggest empirical research testing this hypothesis to prioritize, among the many features that could be relevant, behavioral risk and the trust threshold. Figure 6 illustrates the effects of these two features on performance. Recall that if behavioral risk is high, then trustworthy and untrustworthy partners exhibit substantially different behavior. For example, such partners are easily distinguished if one tends to supply a low-quality product while the other tends to supply a high-quality product. If they behave similarly, then behavioral risk is low, making it harder to distinguish between them.

The left panel of Fig. 6 shows that behavioral risk determines how quickly a party changes its trustworthiness belief about the other party. The top example illustrates the

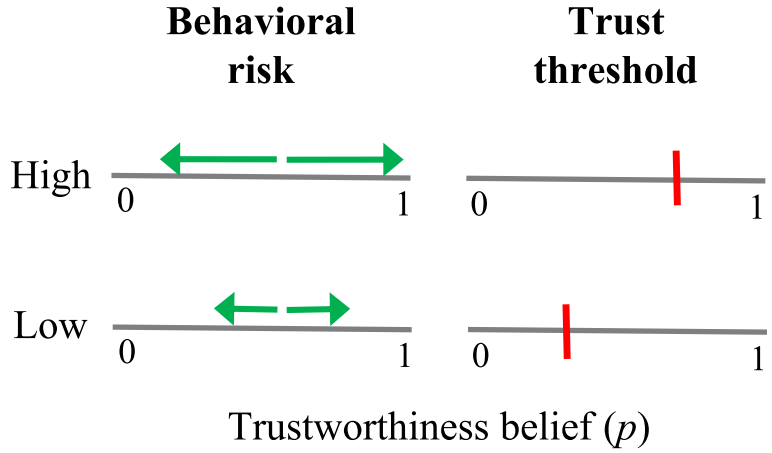

Fig. 6 Behavioral risk and trust threshold are key drivers of trust-based governance performance

situation with high behavioral risk. A low- or high-quality product leads to large changes in perceptions of trustworthiness. In the left panel's bottom example, a new product leads only to small changes because the two partner types take similar actions.

One factor that affects behavioral risk is the verifiability of outcomes—in other words, by whether an outsider can objectively establish the product's quality (Baker et al. 2002). A product's non-verifiability may lead an untrustworthy partner to act differently toward a trustworthy partner. In the case of verifiability, the behavioral risk could be low in the sense that untrustworthy and trustworthy partners then act more alike.

Non-verifiability hinders contract-based governance (Levin 2003), but our results suggest it could improve the performance of trust-based governance. The reason is that trust-based governance performs well when it results not only in a willingness to be vulnerable with trustworthy partners but also in an unwillingness to be vulnerable with untrustworthy partners. Situations involving non-verifiability, and hence behavioral risk, yield better information about the other's trustworthiness. What follows is a quicker dismissal of untrustworthy partners and longer relationships with trustworthy ones.

The study's other key component, the trust threshold, captures how much one's willingness to be vulnerable depends on how trustworthy the other is perceived to be. In the right panel of Fig. 6, we can see that in the top example, one organization is willing to be vulnerable to the other's actions only if the former perceives the latter as trustworthy; in the bottom example, vulnerability is accepted even if the other is suspected of being untrustworthy. In this case, then, the trust threshold is low. Thus, trust is a situation where the trust threshold is between, not equal to, 0 and 1. The trust threshold indicates the willingness of a party to be vulnerable even when there is a suspicion of a lack of trustworthiness. Thus, we predict that trust-based governance performance is best when behavioral risk is high and also the trust threshold is low. An alternative formulation is that high performance follows when one learns quickly about the other's trustworthiness and when trustworthiness belief is only mildly relevant.

Because multiple features affect the trust threshold in a nonlinear way (and these features also affect the trust-based governance performance), the most promising avenue is to measure the trust threshold directly. A survey-based approach is suitable with the interorganizational relationship as the unit of analysis.

Using our model, we can describe the difference between our study on trust-based governance and previous studies on the performance implications of perceived trustworthiness and trust as follows. Consider two situations, I and II, in which perceived trustworthiness (p) is 0.5 and the trust threshold (p^*) is 0.2; however, behavioral risk (a) is 0.3 in situation I and 0.1 in situation II. Research examining perceived trustworthiness seeks to understand the relationship between p and relational outcomes ($V(p)$). As in our model, existing work reports a positive association between p and $V(p)$. Studies that focus on trust aim to understand what the "optimal" level of trust should be (e.g., should p^* equal 0.2 or instead some other value?). We are interested in understanding, given p and p^*, whether trust-based governance performs better in situation I than in II. Thus, we highlight the situation contingent nature of trust-based governance.

The model focuses on the performance of trust-based governance across situations. A next step is to understand, for a given situation, when does trust-based governance outperform alternative governance arrangements. One prominent alternative is contracts (Williamson 1991; Mayer and Argyres 2004; Puranam and Vanneste 2009). An

interpretation of the trust game is that it concerns payoffs beyond those specified in a binding contract. But to understand their relative strength, both need to be explicitly modeled. One approach is to incorporate contracting cost so that the extent of contracting becomes a choice rather than a baseline. We see the comparison of contract- and trust-based governance as an important area for future research.

Conclusion

Trust has been described as an "important lubricant of the social system" (Arrow 1974, p. 23). Although many have heralded the advantages of relying on trust in interorganizational relationships, doubts remain about when interorganizational trust works best. Those doubts motivate our research question: When does trust allow for maximal value to be realized with trustworthy partners while limiting vulnerability to untrustworthy partners? We employ a systematic analysis of trust situations and find that trust-based governance performs better when behavioral risk is high or when the trust threshold is low.

Appendix

The dynamic programming model of this paper is an optimal stopping problem. Hence, our proofs rely on analysis of the following expression, which formalizes the difference in value between continuing and terminating a relationship when t periods remain:

$$B_t(p) \triangleq \mathbb{E}[\, c(k_A + k_B)\tilde{X} + \delta V_{t-1}(\tilde{p}(p))] - k_A(1 + \delta + \cdots + \delta^{t-1}),$$

where

$$\tilde{p}(p) \triangleq \begin{cases} \frac{p(1-(\alpha-a))}{p(1-(\alpha-a))+(1-p)(1-(\alpha+a))}, & P(\tilde{X} = X + x) = p(1-(\alpha-a)) + (1-p)(1-(\alpha+a)); \\ \frac{p(\alpha-a)}{p(\alpha-a)+(1-p)(\alpha+a)}, & P(\tilde{X} = X - x) = p(\alpha-a) + (1-p)(\alpha+a). \end{cases}$$

We define the special case of $t = 1$ by denoting the benefit of continuing the relationship for exactly one more period as:

$$\begin{aligned} M(p) &\triangleq \mathbb{E}[\, c(k_A + k_B)\tilde{X}] - k_A \\ &= [\, p(\alpha - a) + (1-p)(\alpha + a)]\, c(k_A + k_B)(X - x) \\ &\quad + [\, 1 - (p(\alpha - a) + (1-p)(\alpha + a))]\, c(k_A + k_B)(X + x) - k_A \\ &= c(k_A + k_B)X - k_A + c(k_A + k_B)(1 - 2\alpha - 2a + 4ap)x. \end{aligned}$$

A. Proof of Lemma 1 (trust threshold) and Proposition 1 (comparative statics)

The next three lemmas describe the properties of $B_t(p)$ and will facilitate the proofs of Lemma 1 and Proposition 1.

Lemma A-1 *$B_t(p)$ is increasing in p.*

Lemma A-2 *Let $B_t(p, x)$ denote $B_t(p)$ with its dependence on x made explicit. Then, $B_t(p, 0) > 0$ if and only if $c(k_A + k_B)X - k_A > 0$.*

Lemma A-3 *Let $B_t(p, a)$ denote $B_t(p)$ with its dependence on a made explicit. Then, $B_t(p, 0) > 0$ if and only if $c(k_A + k_B)(X + (1 - 2\alpha)x) > k_A$.*

Proof of Lemma A-1 We will show by induction that $B_t(p)$ is increasing in p for all t. We start with the following recursive expression of that function:

$$\begin{aligned} B_t(p) &= \mathbb{E}[\, c(k_A+k_B)\tilde{X} + \delta V_{t-1}(\tilde{p}(p))] - k_A\left(1+\delta+\cdots+\delta^{t-1}\right) \\ &= \mathbb{E}[\, c(k_A+k_B)\tilde{X}] + \delta\mathbb{E}\left[\max\left\{k_A\left(1+\delta+\cdots+\delta^{t-2}\right),\right.\right. \\ &\quad \left.\left.\mathbb{E}\left[c(k_A+k_B)\tilde{X} + \delta V_{t-2}(\tilde{p}(\tilde{p}(p)))\right]\right\}\right] \\ &\quad - k_A(1+\delta+\cdots+\delta^{t-1}) \\ &= \mathbb{E}[\, c(k_A+k_B)\tilde{X}] - k_A + \delta\mathbb{E}[\, \max\{0, B_{t-1}(\tilde{p}(p))\}]\,. \end{aligned}$$

If $t = 1$, then $B_1(p) = M(p)$ and is increasing in p. Now, suppose that $B_{t-1}(p)$ is also increasing in p. Then, $B_t(p) = M(p) + \delta\mathbb{E}[\, \max\{0, B_{t-1}(\tilde{p}(p))\}]$ is increasing in p because (a) $M(p)$ is increasing and (b) $\delta\mathbb{E}[\, \max\{0, B_{t-1}(\tilde{p}(p))\}]$, as a composition of nondecreasing functions, is itself nondecreasing. □

Proof of Lemma A-2 We will show by induction that, for all t,

$$B_t(p,0) = \begin{cases} (1+\delta+\cdots+\delta^{t-1})(c(k_A+k_B)X - k_A) & \text{if } c(k_A+k_B)X - k_A > 0, \\ c(k_A+k_B)X - k_A & \text{otherwise.} \end{cases}$$

If $t = 1$, then $B_1(p,0) = M(p,0) = c(k_A+k_B)X - k_A$ and so the expression is satisfied. Now, suppose that:

$$B_{t-1}(p,0) = \begin{cases} (1+\delta+\cdots+\delta^{t-2})M(p,0) & \text{if } M(p,0) > 0, \\ M(p,0) & \text{otherwise.} \end{cases}$$

Then, we have:

$$\begin{aligned} B_t(p,0) &= M(p,0) + \delta\mathbb{E}[\, B_{t-1}(\tilde{p}(p),0)]^+ = M(p,0) + \delta[\, B_{t-1}(p,0)]^+ \\ &= \begin{cases} M(p,0) + (1+\delta+\cdots+\delta^{t-2})M(p,0) & \text{if } M(p,0) > 0, \\ M(p,0) + 0 & \text{otherwise;} \end{cases} \\ &= \begin{cases} (1+\delta+\cdots+\delta^{t-1})M(p,0) & \text{if } M(p,0) > 0, \\ M(p,0) & \text{otherwise.} \end{cases} \end{aligned}$$

□

Proof of Lemma A-3 We will show by induction that, for all t,

$$B_t(p,0) = \begin{cases} (1+\delta+\cdots+\delta^{t-1})(c(k_A+k_B)(X+(1-2\alpha)x) - k_A) & \text{if } \gamma > k_A, \\ c(k_A+k_B)(X+(1-2\alpha)x) - k_A & \text{otherwise;} \end{cases}$$

to save space, we have put $\gamma \triangleq c(k_A+k_B)(X+(1-2\alpha)x)$. If $t = 1$, then $B_1(p,0) = M(p,0) = \gamma - k_A$, which satisfies the expression. Next, suppose that:

$$B_{t-1}(p,0) = \begin{cases} (1+\delta+\cdots+\delta^{t-2})M(p,0) & \text{if } M(p,0) > 0, \\ M(p,0) & \text{otherwise.} \end{cases}$$

Then,

$$B_t(p,0) = M(p,0) + \delta\mathbb{E}[\,B_{t-1}(\tilde{p}(p),0)]^+ = M(p,0) + \delta[\,B_{t-1}(p,0)]^+$$
$$= \begin{cases} M(p,0) + (1+\delta+\cdots+\delta^{t-2})M(p,0) & \text{if } M(p,0) > 0, \\ M(p,0) + 0 & \text{otherwise;} \end{cases}$$
$$= \begin{cases} (1+\delta+\cdots+\delta^{t-1})M(p,0) & \text{if } M(p,0) > 0, \\ M(p,0) & \text{otherwise.} \end{cases}$$

□

Proof of Lemma 1 It follows from Lemma A-1 that the benefit of continuing over terminating, $B_t(p)$, is increasing in p for any t. Hence, for any t, there exists a unique p_t^* such that $B_t(p_t^*) = 0$, where it is optimal to terminate if and only if the trustworthiness belief $p_t < p_t^*$. Since a unique threshold exists for each t, there must exist a threshold as $t \to \infty$. □

Proof of Proposition 1 For all parts (i)–(v), we consider the implicit function defined by $B_t(p^*, y) = 0$ for $y \in \{c, X, x, \alpha, a\}$; this function establishes the trust threshold. By the implicit function theorem, $\frac{\partial p^*}{\partial y} = -\frac{\partial B(p,y)}{\partial y} \Big/ \frac{\partial B(p,y)}{\partial p}$. Since $B_t(p,y)$ is increasing in p (by Lemma A-1), it follows that:

$$\frac{\partial p^*}{\partial y} > 0 \iff \frac{\partial B(p,y)}{\partial y} < 0.$$

To prove parts (i) and (ii), we will show by induction that $B_t(p)$ increases with c (and with X) for all t and p. We already know that $B_1(p) = M(p)$ increases with c (and X) for all p. Next, suppose that $B_{t-1}(p)$ increases with c (and X). Since $\tilde{p}(p)$ is independent of c (and X), it follows that also $\mathbb{E}[\max\{0, B_{t-1}(\tilde{p})\}]$ increases with c (and X). Therefore, $B_t(p) = M(p) + \delta\mathbb{E}[\max\{0, B_{t-1}(\tilde{p})\}]$ also increases with c (and X).

(iii) Increased outcome risk can either increase or decrease the trust threshold. We will first show by induction that for all t and p, the term $B_t(p)$ increases with x if $\alpha + a < 0.5$ but decreases with x if $\alpha - a > 0.5$. We can also write $M(p)$ as follows:

$$M(p) = (c(k_A + k_B)X - k_A) + c(k_A + k_B)(1 - 2(\alpha - a))x - 4c(k_A + k_B)a(1-p)x.$$

It is clear from this expression that $B_1(p) = M(p)$ increases with x if $\alpha + a < 0.5$ (first equality) and decreases with x if $\alpha - a > 0.5$ (second equality). Now, we suppose that these properties hold for $B_{t-1}(p)$; then, because $\tilde{p}(p)$ is independent of x, they must hold also for $B_t(p) = M(p) + \delta\mathbb{E}[\max\{0, B_{t-1}(\tilde{p}(p))\}]$.

Next, let $\alpha - a < 0.5 < \alpha + a$. In this intermediate case, the following argument holds provided that $B_t(p,x)$ is monotonic in x (which is the only case we observe). Suppose first that $c(k_A + k_B)X - k_A > 0$. Then, $B_t(p,0) > 0$ by Lemma A-2 and so, in order for there to exist an x such that $B_t(p^*, x) = 0$, the value of $B_t(p^*, x)$ must decrease with x; that is, we must have $\frac{\partial p^*}{\partial x} > 0$. Now, suppose $c(k_A + k_B)X - k_A < 0$. Then, $B_t(p,0) < 0$ (again by Lemma A-2) and so, if there exists an x such that $B_t(p^*, x) = 0$, then $B_t(p^*, x)$ must increase in x—that is, $\frac{\partial p^*}{\partial x} < 0$.

(iv) We will show by induction that $B_t(p)$ decreases with α for all t and p. Recall that $B_1(p) = M(p)$ decreases with α for all p. Note that $\tilde{p}(p)$ decreases with α, and suppose that $B_{t-1}(p)$ also decreases with α. Then, $\mathbb{E}[\max\{0, B_{t-1}(\tilde{p}(p))\}]$ decreases with α, from which it follows that $B_t(p) = M(p) + \mathbb{E}[\max\{0, B_{t-1}(\tilde{p})\}]$ also decreases with α.

(v) An increase in behavioral risk can either increase or reduce the trust threshold. We illustrate this statement for a $B_t(p,a)$ that is quasi-convex in a (which is the only shape we observe). On the one hand, if $k_A > c(k_A + k_B)(X + (1-2\alpha))$, then $B_t(p,0) < 0$ (by Lemma A-3) and so $B_t(p,a)$ must increase with a when $B_t(p^*,a) = 0$; that is, $\frac{\partial p^*}{\partial a} < 0$. On the other hand, if $k_A < c(k_A + k_B)(X + (1-2\alpha))$, then $B_t(p,0) > 0$ (again by Lemma A-3). Provided that $B_t(p,a)$ is quasi-convex in a, it can decrease and then increase with a. If there is a unique a such that $B_t(p,a) = 0$, then $B_t(p,a)$ must decrease with a for $B_t(p^*,a) = 0$; that is, $\frac{\partial p^*}{\partial a} > 0$. Suppose there exist $a_1 < a_2$ such that $B_t(p_1^*,a_1) = B_t(p_2^*,a_2) = 0$. In that case, $B_t(p,a)$ decreases with a at $B_t(p_1^*,a_1) = 0$ (i.e., $\partial p_1^*/\partial a > 0$ at $a \le a_1$) and increases with a at $B_t(p_2^*,a_2) = 0$ (i.e., $\partial p_2^*/\partial a < 0$ at $a \ge a_2$). □

B. Proof of Proposition 2 (performance of trust-based governance)

The following three lemmas characterize the updating process of p_t and will help us to prove Proposition 2.

Lemma A-4 *Suppose that p_0, the initial belief at time 0, is equal to p. Then, for $p_t \le p^*$, organization A must receive at least $\overline{k}$ low outcomes in t periods, where:*

$$\overline{k} = \left\lceil \frac{\log\left[\left(\frac{1}{p^*}-1\right)/\left(\frac{1}{p}-1\right)\right]}{\log\left[\left(\frac{\alpha+a}{\alpha-a}\right)/\left(\frac{1-\alpha-a}{1-\alpha+a}\right)\right]} - \frac{\log\left(\frac{1-\alpha-a}{1-\alpha+a}\right)}{\log\left[\left(\frac{\alpha+a}{\alpha-a}\right)/\left(\frac{1-\alpha-a}{1-\alpha+a}\right)\right]} t \right\rceil.$$

Lemma A-5 *Let $T \ge 0$ denote the random duration of the relationship. Then, for all t, the following statements hold:*

(i) *$P(T > t \mid \text{trustworthy})$ increases with a, decreases with α, and is independent of c, X, and x;*

(ii) *$P(T > t \mid \text{untrustworthy})$ decreases with a, decreases with α, and is independent of c, X, and x.*

Lemma A-6 (adjustment of trustworthiness belief *p*)

(i) *Value creation (c), value capture (X), and risk (x) do not affect the upward or downward adjustment of trustworthiness belief p.*

(ii) *Self-serving norm (α) increases the magnitude of upward adjustment (after a high outcome) and decreases the magnitude of downward adjustment (after a low outcome) in trustworthiness belief p.*

(iii) *Behavioral risk (a) increases the magnitude of both upward and downward adjustments in trustworthiness belief p.*

Proof of Lemma A-4 We start by observing that according to Bayes' rule, the updated belief after m low outcomes and n high outcomes is:

$$p_t = \frac{p(\alpha-a)^m(1-\alpha+a)^n}{p(\alpha-a)^m(1-\alpha+a)^n + (1-p)(\alpha+a)^m(1-\alpha-a)^n}. \tag{A.1}$$

We therefore seek the minimum k such that $p_t = \frac{p(\alpha-a)^k(1-\alpha+a)^{t-k}}{p(\alpha-a)^k(1-\alpha+a)^{t-k}+(1-p)(\alpha+a)^k(1-\alpha-a)^{t-k}} \le p^*$. Taking the equality, we have:

$$
\begin{aligned}
p^* &= \frac{p}{p + (1-p)\left(\frac{\alpha+a}{\alpha-a}\right)^k \left(\frac{1-\alpha-a}{1-\alpha+a}\right)^{t-k}} \\
&\iff \left(\frac{1}{p^*} - 1\right)\frac{p}{1-p} = \left(\frac{\alpha+a}{\alpha-a}\right)^k \left(\frac{1-\alpha-a}{1-\alpha+a}\right)^{t-k} \\
&\iff \log\left[\left(\frac{1}{p^*} - 1\right)\Big/\left(\frac{1}{p} - 1\right)\right] = k \log\left[\left(\frac{\alpha+a}{\alpha-a}\right)\Big/\left(\frac{1-\alpha-a}{1-\alpha+a}\right)\right] \\
&\quad + t \log\left(\frac{1-\alpha-a}{1-\alpha+a}\right) \\
&\iff k = \frac{\log\left[\left(\frac{1}{p^*} - 1\right)\Big/\left(\frac{1}{p} - 1\right)\right]}{\log\left[\left(\frac{\alpha+a}{\alpha-a}\right)\Big/\left(\frac{1-\alpha-a}{1-\alpha+a}\right)\right]} - \frac{\log\left(\frac{1-\alpha-a}{1-\alpha+a}\right)}{\log\left[\left(\frac{\alpha+a}{\alpha-a}\right)\Big/\left(\frac{1-\alpha-a}{1-\alpha+a}\right)\right]} t.
\end{aligned}
$$

The minimum number k is thus the first integer greater than this expression. □

Proof of Lemma A-5 (i) We shall proceed by way of induction. As a base case, let s denote the shortest possible duration from the start of the relationship for belief p to fall below the threshold p^*, which corresponds to receiving s consecutive low outcomes. The probability of this occurring with a trustworthy partner is $P(T = s \mid \text{trustworthy}) = (\alpha - a)^s$. It follows that $P(T > s \mid \text{trustworthy}) = 1 - (\alpha - a)^s$ increases with a, decreases with α, and is independent of c, X, and x.

Next suppose that $P(T > t - 1 \mid \text{trustworthy})$ increases with a, decreases with α, and is independent of c, X, and x. Since $P(T > t \mid \text{trustworthy}) = P(T > t - 1 \mid \text{trustworthy}) - P(T = t \mid \text{trustworthy})$, it follows that (a) $P(T > t \mid \text{trustworthy})$ increases with a, decreases with α, and is independent of c, X, and x, and (b) $P(T = t \mid \text{trustworthy})$ decreases with a, increases with α, yet is also independent of c, X, and x.

From the expression for p_t in (A.1), the relationship duration T is equal to t if

$$
\frac{p(\alpha-a)^k(1-\alpha+a)^{t-k}}{p(\alpha-a)^k(1-\alpha+a)^{t-k} + (1-p)(\alpha+a)^k(1-\alpha-a)^{t-k}} \leq p^*
$$

and $p_\tau > p^*$ for all $\tau < t$. By Lemma A-4, these conditions will be satisfied only if there are exactly $\overline{k}$ low outcomes in t periods. There are $\frac{t!}{\overline{k}!\,(t-\overline{k})!}$ combinations of high–low outcome sequences that will lead to $p_t \leq p^*$ after t periods. Of these, there may be combinations that include sequences where $p_\tau < p^*$ for some $\tau < t$, so the number of combinations $\ell \leq \frac{t!}{\overline{k}!\,(t-\overline{k})!}$. Therefore,

$$
P(T = t \mid \text{trustworthy}) = \ell \cdot (\alpha - a)^{\overline{k}}(1-\alpha+a)^{t-\overline{k}}.
$$

This equality is independent of c, X, and x. Moreover, since ℓ is an integer, it follows that an incremental change in a or in α would not affect its value. We can therefore write:

$$
\begin{aligned}
\frac{\partial}{\partial a} P(T = t \mid \text{trustworthy}) &= \ell \cdot \Big[-\overline{k}(\alpha-a)^{\overline{k}-1}(1-\alpha+a)^{t-\overline{k}} \\
&\quad + (\alpha-a)^{\overline{k}}(1-\alpha+a)^{t-\overline{k}-1}(t-\overline{k})\Big] \\
&= \ell \cdot (\alpha-a)^{\overline{k}}(1-\alpha+a)^{t-\overline{k}}\left[\frac{-\overline{k}}{\alpha-a} + \frac{t-\overline{k}}{1-\alpha+a}\right] < 0;
\end{aligned}
$$

$$\begin{aligned}\frac{\partial}{\partial\alpha}P(T=t \mid \text{trustworthy}) &= \ell\cdot\Big[\overline{k}(\alpha-a)^{\overline{k}-1}(1-\alpha+a)^{t-k}\\ &\quad -(\alpha-a)^{\overline{k}}(1-\alpha+a)^{t-\overline{k}-1}(t-\overline{k})\Big]\\ &= \ell\cdot(\alpha-a)^{\overline{k}}(1-\alpha+a)^{t-\overline{k}}\left[\frac{\overline{k}}{\alpha-a}-\frac{t-\overline{k}}{1-\alpha+a}\right]>0.\end{aligned}$$

Here, the second inequality follows because $\frac{\overline{k}}{\alpha+a} < \frac{t-\overline{k}}{1-\alpha-a}$ if and only if $\frac{\overline{k}}{t} < \alpha+a$; the first inequality follows because $\frac{\overline{k}}{t} > -\frac{\log\left(\frac{1-\alpha-a}{1-\alpha+a}\right)}{\log\left[\left(\frac{\alpha+a}{\alpha-a}\right)/\left(\frac{1-\alpha-a}{1-\alpha+a}\right)\right]} > \alpha-a$ (from Lemma A-4) implies that $\frac{t-\overline{k}}{1-\alpha+a} < \frac{\overline{k}}{\alpha-a}$.

(ii) For the untrustworthy case, we examine regions $t > \frac{\overline{k}}{\alpha+a}$ and $t \le \frac{\overline{k}}{\alpha+a}$ separately.

(a) For any $t > \frac{\overline{k}}{\alpha+a}$, we have $P(T=t \mid \text{untrustworthy}) = \ell\cdot(\alpha+a)^{\overline{k}}(1-\alpha-a)^{t-\overline{k}}$, where this equality is independent of c, X, and x. Once again, that ℓ is an integer implies that an incremental change in a or in α will not affect its value. Therefore:

$$\begin{aligned}\frac{\partial}{\partial a}P(T=t \mid \text{untrustworthy}) &= \ell\cdot\Big[\overline{k}(\alpha+a)^{\overline{k}-1}(1-\alpha-a)^{t-\overline{k}}\\ &\quad -(\alpha+a)^{\overline{k}}(1-\alpha-a)^{t-\overline{k}-1}(t-\overline{k})\Big]\\ &= \ell\cdot(\alpha+a)^{\overline{k}}(1-\alpha-a)^{t-\overline{k}}\left[\frac{\overline{k}}{\alpha+a}-\frac{t-\overline{k}}{1-\alpha-a}\right]<0;\\ \frac{\partial}{\partial\alpha}P(T=t \mid \text{untrustworthy}) &= \ell\cdot\Big[k(\alpha+a)^{\overline{k}-1}(1-\alpha-a)^{t-\overline{k}}\\ &\quad -(\alpha+a)^{\overline{k}}(1-\alpha-a)^{t-\overline{k}-1}(t-\overline{k})\Big]\\ &= \ell\cdot(\alpha+a)^{\overline{k}}(1-\alpha-a)^{t-\overline{k}}\left[\frac{\overline{k}}{\alpha+a}-\frac{t-\overline{k}}{1-\alpha-a}\right]<0.\end{aligned}$$

Here, the two inequalities both follow because $\frac{\overline{k}}{\alpha+a} < \frac{t-\overline{k}}{1-\alpha-a}$ if and only if $\frac{\overline{k}}{t} < \alpha+a$. So not only is $t > \frac{\overline{k}}{\alpha+a}$, we also have that $P(T > t \mid \text{untrustworthy}) = \sum_{t>\overline{k}/(\alpha+a)}^{\infty} P(T=t \mid \text{untrustworthy})$ decreases with a, decreases with α, and is independent of c, X, and x.

(b) For any $t \le \frac{\overline{k}}{\alpha+a}$, we will show the result by induction. As a base case, the probability of receiving consecutive s bad outcomes from an untrustworthy partner is $P(T=s \mid \text{untrustworthy}) = (\alpha+a)^s$. Hence, $P(T > s \mid \text{untrustworthy}) = 1-(\alpha+a)^s$ decreases with a, decreases with α, and is independent of c, X, and x.

Now, suppose that $P(T > t-1 \mid \text{untrustworthy})$ decreases with a, decreases with α, and is independent of c, X, and x. Since $P(T > t \mid \text{untrustworthy}) = P(T > t-1 \mid \text{untrustworthy}) - P(T=t \mid \text{untrustworthy})$, it follows that $P(T > t \mid \text{untrustworthy})$ decreases with a, decreases with α, and is independent of c, X, and x *provided that* $P(T = t \mid \text{untrustworthy})$ increases with a, increases with α, and is independent of c, X, and x. Therefore:

$$\begin{aligned}\frac{\partial}{\partial a}P(T=t \mid \text{untrustworthy}) &= \ell \cdot \Big[k(\alpha+a)^{\bar{k}-1}(1-\alpha-a)^{t-\bar{k}} \\ &\quad -(\alpha+a)^{\bar{k}}(1-\alpha-a)^{t-\bar{k}-1}(t-\bar{k})\Big] \\ &= \ell \cdot (\alpha+a)^{\bar{k}}(1-\alpha-a)^{t-\bar{k}}\left[\frac{\bar{k}}{\alpha+a}-\frac{t-\bar{k}}{1-\alpha-a}\right] > 0; \\ \frac{\partial}{\partial \alpha}P(T=t \mid \text{untrustworthy}) &= \ell \cdot \Big[\bar{k}(\alpha+a)^{\bar{k}-1}(1-\alpha-a)^{t-\bar{k}} \\ &\quad -(\alpha+a)^{\bar{k}}(1-\alpha-a)^{t-\bar{k}-1}(t-\bar{k})\Big] \\ &= \ell \cdot (\alpha+a)^{\bar{k}}(1-\alpha-a)^{t-\bar{k}}\left[\frac{\bar{k}}{\alpha+a}-\frac{t-\bar{k}}{1-\alpha-a}\right] > 0.\end{aligned}$$

Here, the two inequalities both follow because $\frac{\bar{k}}{\alpha+a} > \frac{t-\bar{k}}{1-\alpha-a}$ if and only if $\frac{\bar{k}}{t} > \alpha + a$. Hence, $P(T > t \mid \text{untrustworthy})$ decreases with a, decreases with α, and is independent of c, X, and x. □

Proof of Lemma A-6 We start by observing that for any trustworthiness belief p, a high outcome increases p to $\frac{p(1-\alpha+a)}{p(1-\alpha+a)+(1-p)(1-\alpha-a)} = \frac{p\left(\frac{1-\alpha+a}{1-\alpha-a}\right)}{1-p\left(1-\frac{1-\alpha+a}{1-\alpha-a}\right)}$ and an low outcome reduces p to $\frac{p(\alpha-a)}{p(\alpha-a)+(1-p)(\alpha+a)} = \frac{p\left(\frac{\alpha-a}{\alpha+a}\right)}{1-p\left(1-\frac{\alpha-a}{\alpha+a}\right)}$. These expressions show that both upward and downward adjustments are independent of c, X, and x (part (i)). For any p, the updated beliefs (after both high and low outcomes) increase with α. The implication is that with higher α, the extent of adjustment after a high outcome increases and that extent after a low outcome decreases (part (ii)). Finally, for all p, we know that the updated belief after a high (resp. low) outcome increases (resp. decreases) with a—in other words, the magnitudes of both upward and downward adjustments are increasing in a (part (iii)). □

Proof of Proposition 2 For $z \in \{c, X, x, \alpha, a\}$, we can use the chain rule to write:

$$\frac{d\mathbb{E}[\,T]}{dz} = \frac{\partial \mathbb{E}[\,T]}{\partial z} + \frac{\partial \mathbb{E}[\,T]}{\partial p^*} \cdot \frac{\partial p^*}{\partial z}. \tag{A.2}$$

The first term on the right-hand side represents the change in the expected duration due to the adjustment in belief p while holding the threshold p^* constant, and the second term represents the change in the expected duration due to the change in the threshold p^*. For each type of partner (trustworthy or untrustworthy), we have:

$$\mathbb{E}[\,T \mid \text{type}] = \sum_{t=1}^{\infty} tP(T=t \mid \text{type}) = \sum_{t=1}^{\infty} P(T > t \mid \text{type}).$$

We conclude that whether $\frac{\partial \mathbb{E}[T|\text{type}]}{\partial z}$ is positive, negative, or zero depends on whether $\frac{\partial P(T>t|\text{type})}{\partial z}$ is (respectively) positive, negative, or zero—which in turn is given by Lemma A-5. Also, $\frac{\partial \mathbb{E}[T]}{\partial p^*} < 0$ because the duration must decrease with an increased termination threshold. And finally, it follows from Proposition 1 that $\frac{\partial p^*}{\partial z}$ can be positive, negative, or zero depending on the situational features. Note that for two of these features, a and X, the sign changes depend on the *other* features' values. For each situational feature (including expected duration), Table 3 reports the possible signs for each component of (A.2). We can see that performance of trust-based governance increases if and only if the expected duration with a trustworthy partner increases *and* that with an untrustworthy partner decreases. □

Table 3 Comparative statics for trust-based governance performance

Part	Feature		Partner	$\frac{\partial \mathbb{E}[T]}{\partial z}$	$\frac{\partial \mathbb{E}[T]}{\partial p^*}$	$\frac{\partial p^*}{\partial z}$	$\frac{d\mathbb{E}[T]}{dz}$	Performance
(i)	Behavioral risk (a)	*Either*	Trustworthy	+	−	+	+/−	+/0
			Untrustworthy	−	−	+	−	
		Or	Trustworthy	+	−	−	+	
			Untrustworthy	−	−	−	−/+	
(ii)	Value creation (c)		Trustworthy	0	−	+	−	0
			Untrustworthy	0	−	+	−	
	Value capture (X)		Trustworthy	0	−	+	−	0
			Untrustworthy	0	−	+	−	
	Outcome risk (x)	*Either*	Trustworthy	0	−	+	−	0
			Untrustworthy	0	−	+	−	
		Or	Trustworthy	0	−	−	+	
			Untrustworthy	0	−	−	+	
	Self-serving norm (α)		Trustworthy	−	−	+	−	0
			Untrustworthy	−	−	+	−	

Acknowledgements
We thank the Editor and Reviewers for their valuable comments. We thank Bilal Gokpinar for many discussions that spurred us to undertake this project. We are grateful also for the helpful comments of Kenan Arifoglu, Isabel Fernandez-Mateo, and members of the School of Management reading group at University College London. Finally, we are indebted to the excellent research assistance of Caroline Koekkoek.

Authors' contributions
BV and OY conceived of the presented idea. BV, with support of OY, developed the theory. OY, with support of BV, developed the model. BV and OY analyzed the results and wrote the paper. All authors read and approved the final manuscript.

Funding
The authors declare no external funding.

Availability of data and materials
Only simulated data used.

Competing interests
The authors declare that they have no competing interests.

Received: 21 September 2018 Accepted: 30 April 2020
Published online: 17 June 2020

References

Anderson E, Weitz B (1989) Determinants of continuity in conventional industrial channel dyads. Mark Sci 8(4):310–323
Argyres NS, Bercovitz J, Mayer KJ (2007) Complementarity and evolution of contractual provisions: an empirical study of IT services contracts. Organ Sci 18(1):3–19
Arrow KJ (1974) The limits of organization. WW Norton & Company, New York, NY
Attanasi G, Battigalli P, Manzoni E (2016) Incomplete-information models of guilt aversion in the trust game. Manag Sci 62(3):648–667
Axelrod R (1984) The evolution of cooperation. Basic Books, New York, NY
Baker G, Gibbons R, Murphy KJ (2002) Relational contracts and the theory of the firm. Q J Econ 117(1):39–84
Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. Games Econ Behav 10(1):122–142
Bhattacharya R, Devinney TM, Pillutla MM (1998) A formal model of trust based on outcomes. Acad Manag Rev 23(3):459–472
Bidwell M, Fernandez-Mateo I (2010) Relationship duration and returns to brokerage in the staffing sector. Organ Sci 21(6):1141–1158
Bohnet I, Zeckhauser R (2004) Trust, risk and betrayal. J Econ Behav Organ 55(4):467–484
Bradach JL, Eccles RG (1989) Price, authority, and trust—from ideal types to plural forms. Ann Rev Soc 15:97–118
Brandenburger AM, Stuart HW (1996) Value based business strategy. J Econ Manag Strat 5(1):5–24
Camerer C (2003) Behavioural game theory. Russell Sage Foundation, Princeton, NJ
Camerer C, Weigelt K (1988) Experimental tests of a sequential equilibrium reputation model. Econometrica 56(1):1–36
Cao Z, Lumineau F (2015) Revisiting the interplay between contractual and relational governance: a qualitative and meta-analytic investigation. J Oper Manag 33–34:15–42
Chatain O, Zemsky P (2011) Value creation and value capture with frictions. Strat Manag J 32(11):1206–1231
Coleman J (1990) Foundations of social theory. Harvard University Press, Cambridge, MA
Colquitt JA, Scott BA, LePine JA (2007) Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. J Appl Psychol 92(4):909–927

Connelly BL, Crook TR, Combs JG, Ketchen DJ, Anguinis H (2018) Competence- and integrity-based trust in interorganizational relationships: which matters more? J Manag 44(3):919–945
Dasgupta P (1988) Trust as a commodity. In: Gambetta DG (ed). Trust. Basil Blackwell, New York, NY. pp 49–72
David RJ, Han SK (2004) A systematic assessment of the empirical support for transaction cost economics. Strat Manag J 25(1):39–58
Doney PM, Cannon JP (1997) An examination of the nature of trust in buyer-seller relationships. J Mark 2(2):35–51
Dyer JH, Chu W (2000) The determinants of trust in supplier-automaker relationships in the U.S., Japan, and Korea. J Int Bus Stud 31(2):259–285
Dyer J, Chu W (2003) The role of trustworthiness in reducing transaction costs and improving performance: empirical evidence from the United States, Japan, and Korea. Organ Sci 14(1):57–68
Dyer JH, Singh H (1998) The relational view: cooperative strategy and sources of interorganizational competitive advantage. Acad Manag Rev 23(4):660–679
Elfenbein DW, Zenger TR (2014) What is a relationship worth? Repeated exchange and the development and deployment of relational capital. Organ Sci 25(1):222–244
Fang E, Palmatier RW, Scheer LK, Li N (2008) Trust at different organizational levels. J Mark 72(2):80–98
Fryxell GE, Dooley RS, Vryza M (2002) After the ink dries: the interaction of trust and control in US-based international joint ventures. J Manag Stud 39(6):865–886
Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning. Springer, Berlin
Gambetta D (1988) Can we trust? In: Gambetta D (ed). Trust: making and breaking cooperative relations. Basil Blackwell, Cambridge, MA. pp 213–237
Ganesan S (1994) Determinants of long-term orientation in buyer-seller relationships. J Mark 58(2):1–19
Gargiulo M, Gokhan E (2006) The dark side of trust. In: Bachmann R, Zaheer A (eds). Handbook of trust research. Edward Elgar, Northampton, MA. pp 165–184
Geyskens I, Steenkamp JBEM, Kumar N (2006) Make, buy, or ally: a transaction cost theory meta-analysis. Acad Manag J 49(3):519–543
Granovetter M (1985) Economic action and social structure: the problem of embeddedness. Am J Sociol 91(3):481–510
Gulati R (1995) Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. Acad Manag J 38(1):85–112
Gulati R, Nickerson JA (2008) Interorganizational trust, governance choice, and exchange performance. Organ Sci 19(5):688–708
Gulati R, Sytch M (2008) Does familiarity breed trust? Revisiting the antecedents of trust. Manag Decis Econ 29(2-3):165–190
Heide JB (1994) Interorganizational governance in marketing channels. J Mark 58(1):71–85
Heide JB, Miner AS (1992) The shadow of the future: effects of anticipated interaction and frequency of contact on buyer-seller cooperation. Acad Manag J 35(2):265–291
Jacobides MG, Hitt LM (2005) Losing sight of the forest for the trees? Productive capabilities and gains from trade as drivers of vertical scope. Strat Manag J 26(13):1209–1227
Johnson ND, Mislin AA (2011) Trust games: a meta-analysis. J Econ Psychol 32(5):865–889
Kiggundu MN (1981) Task interdependence and the theory of job design. Acad Manag Rev 6(3):499–508
Krishnan R, Martin X, Noorderhaven NG (2006) When does trust matter to alliance performance? Acad Manag J 49(5):894–917
Kumar N, Scheer LK, Steenkamp JBEM (1995) The effects of supplier fairness on vulnerable resellers. J Mark Res 32(1):54–65
Lave CA, March JG (1993) An introduction to models in the social sciences. University Press of America, Lanham, MD
Levin J (2003) Relational incentive contracts. Am Econ Rev 93(3):835–857
Lioukas CS, Reuer JJ (2002) Isolating trust outcomes from exchange relationships: social exchange and learning benefits of prior ties in alliances. Acad Manag J 58(6):1826–1847
Luo Y (2002) Building trust in cross-cultural collaborations: toward a contingency perspective. J Manag 28(5):669–694
Luo Y (2008) Structuring interorganizational cooperation: the role of economic integration in strategic alliances. Strat Manag J 29(6):617–637
Lumineau F, Oxley JE (2012) Let's work it out (or we'll see you in court): litigation and private dispute resolution in vertical exchange relationships. Organ Sci 23(3):820–834
Malhotra D, Murnighan JK (2002) The effects of contracts on interpersonal trust. Admin Sci Q 47(3):534–559
Mayer KJ, Argyres NS (2004) Learning to contract: evidence from the personal computer industry. Organ Sci 15(4):394–410
Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. Acad Manag Rev 20(3):709–734
McEvily B, Tortoriello M (2011) Measuring trust in organisational research: review and recommendations. J Trust Res 1(1):23–63
McEvily B, Zaheer A (2006) Does trust still matter? Research on the role of trust in interorganizational exchange. In: Bachmann R, Zaheer A (eds). Handbook of Trust Research. Edward Elgar, Cheltenham, U.K. pp 280–300
Miller DT (1999) The norm of self-interest. Am Psychol 54(12):1053–1060
Mohr AT, Puck J (2013) Revisiting the trust-performance link in strategic alliances. Manag Int Rev 53(2):269–289
Morgan RM, Hunt SD (1994) The commitment-trust theory of relationship marketing. J Mark 58(3):20–38
Nguyen TV, Rose J (2009) Building trust—evidence from Vietnamese entrepreneurs. J Bus Ventur 24(2):165–182
Norman PM (2004) Knowledge acquisition, knowledge loss, and satisfaction in high technology alliances. J Bus Res 57(6):610–619
Özer Ö, Subramanian U, Wang Y (2018) Information sharing, advice provision, or delegation: what leads to higher trust and trustworthiness? Manag Sci 64(1):474–493
Özer Ö, Zheng Y, Chen KY (2011) Trust in forecast information sharing. Manag Sci 57(6):1111–1137
Poppo L, Zenger T (2002) Do formal contracts and relational governance function as substitutes or complements? Strat Manag J 23(8):707–725
Poppo L, Zhou KZ, Ryu S (2008) Alternative origins to interorganizational trust: an interdependence perspective on the shadow of the past and the shadow of the future. Organ Sci 19(1):39–55

Porter ME (1980) Competitive strategy. The Free Press, New York, NY
Puranam P, Vanneste BS (2009) Trust and governance: untangling a tangled web. Acad Manag Rev 34(1):11–31
Ring PS, Van de Ven AH (1992) Structuring cooperative relationships between organizations. Strat Manag J 13(7):483–498
Ross SM (1983) Introduction to stochastic dynamic programming. Academic Press, London, U.K.
Rousseau DM, Sitkin SB, Burt RS, Camerer C (1998) Not so different after all: a cross-discipline view of trust. Acad Manag Rev 23(3):393–404
Ryall MD, Sampson RC (2009) Formal contracts in the presence of relational enforcement mechanisms: evidence from technology development projects. Manag Sci 55(6):906–925
Sako M, Helper S (1998) Determinants of trust in supplier relations: Evidence from the automotive industry in Japan and the United States. J Econ Behav Organ 34(3):387–417
Selten R (1975) Reexamination of the perfectness concept for equilibrium points in extensive games. Int J Game Theory 4(1):25–55
Stevens M, MacDuffie JP, Helper S (2015) Reorienting and recalibrating inter-organizational relationships: strategies for achieving optimal trust. Organ Stud 36(9):1237–1264
Vanneste BS (2016) From interpersonal to interorganisational trust: The role of indirect reciprocity. J Trust Res 6(1):7–36
Vanneste BS, Puranam P, Kretschmer T (2014) Trust over time in exchange relationships: meta-analysis and theory. Strat Manag J 35(12):1891–1902
Wang Q, Bradford K, Xu J, Weitz B (2008) Creativity in buyer-seller relationships: the role of governance. Int J Res Market 25(2):109–118
Wasti SN, Wasti SA (2008) Trust in buyer and supplier relations: the case of the Turkish automotive industry. J Int Bus Stud 39(1):118–131
Wicks AC, Berman SL, Jones TM (1999) The structure of optimal trust: moral and strategic implications. Acad Manag Rev 24(1):99–116
Williamson OE (1975) Markets and hierarchies—analysis and antitrust implications. The Free Press, New York, NY
Williamson OE (1985) The economic institutions of capitalism. Free Press, New York, NY
Williamson OE (1991) Comparative economic organization: the analysis of discrete structural alternatives. Admin Sci Q 36(2):269–296
Yilmaz C, Kabadayi ET (2006) The role of monitoring in interfirm exchange: effects on partner unilateral cooperation. J Bus Res 59(12):1231–1238
Zaheer A, Harris J (2006) Interorganizational trust. In: Shenkar O, Reuer JJ (eds). Handbook of strategic alliances. Sage Publications, Thousand Oaks, CA. pp 169–197
Zaheer A, McEvily B, Perrone V (1998) Does trust matter? Exploring the effects of interorganizational and interpersonal trust on performance. Organ Sci 9(2):141–159
Zajac EJ, Olsen CP (1993) From transaction cost to transacional value analysis: implications for the study of interorganizational strategies. J Manag Stud 30(1):131–145
Zeng M, Chen X-P (2003) Achieving cooperation in multiparty alliances: a social dilemma approach to partnership management. Acad Manag Rev 28(4):587–605
Zhong W, Su C, Peng J, Yang Z (2017) Trust in interorganizational relationships: a meta-analytic integration. J Manag 43(4):1050–1075

Publisher's Note