In: Forgas, J. P. (2025). *The social psychology of trust.* The 27th Sydney Symposium on Social Psychology. Routledge.

Heuristic Trust

Joachim I. Krueger

Brown University, Providence, USA

Anthony M. Evans

Allstate Corporation, Chicago, USA

David J. Grüning

Max Planck Institute for Human Development, Berlin, Germany

Benjamin Aizenberg

Brown University, Providence, USA

Correspondence:

Joachim I. Krueger
Department of Cognitive & Psychological Sciences
Brown University, Box 1820
190 Thayer Street
Providence, RI 02912

+1 401 863 2503
Joachim_Krueger@Brown.edu
https://vivo.brown.edu/display/jkrueger

Keywords: interpersonal trust, trust game, free will

Note. We thank Sheila K., who entrusted us with the opening story, and Leanna Kish and Jessica Tuchin for helping with the empirical project and not shying away from difficult conversations about free will.

Abstract

Interpersonal trust requires decisions under uncertainty (not risk!) as the probability of the other person reciprocating is unknown and can only be approached with rough estimates. It is difficult, if not impossible, to optimize trust decisions in rigorous and coherent ways. A suite of social heuristics is the trustors' best means to achieve a satisfactory solution. We review the findings of a recent research program on bounded rationality in the trust game. We identify a set of social heuristics people can (or should) use when deciding whether to trust. Among these heuristics are social projection, social distance, all-or-nothing, and attention to the general normative environment. We present new empirical findings showing how people might choose whether to submit to different types of dictators in the eponymous game.

127 words

5,681 words of main text on March 20

Und die Treue, sie ist doch kein leerer Wahn; *Truth is no dream! - its power is strong.*
So nehmet auch mich zum Genossen an. *Give grace to him who owns his wrong!*
Ich sei, gewährt mir die Bitte, *'Tis mine your suppliant now to be,*
In eurem Bunde der Dritte. *Ah, let the band of love - be three!*

– F. Schiller, Die Bürgschaft – The Pledge

You must trust and believe in people or life becomes impossible.

– A. Chekhov, Ionych

IT is not for us to question the wisdom of Schiller or Chekhov, but we may wish to contextualize their message. In Schiller's famous 1799 poem, Damon is condemned to die by crucifixion for trying to knife Dionysius, the tyrant of Syracuse. His dear friend, who remains unnamed, pledges to stand in for Damon for three days so that Damon can go and give his sister away in marriage (The Greeks of Sicily appear to have done that). The tyrant is incredulous but agrees, Damon manages to return just in time, thus rewarding the friend's arguably irrational trust, and melting the tyrant's heart. Schiller's poem of rewarded trust is a high-spirited testament to German romanticism. By contrast, Chekhov's declaration in praise of trust is simple and deontological. Can we live by these high ideals?

To curb your enthusiasm, consider this true story [if this story were merely the product of fanciful imagination, it would not matter; the point stands]. A man – let's call him Max – has been texting with a woman – let's call her Maggie – for weeks. The correspondence has been going well; there is rapport. Max and Maggie have shared photos, but they have not had a video conference. Max is in his late 50s and lives in Tampa; Maggie is in her early 60s and lives in Tasmania. Max is broke, but interested in sex and adventure. Maggie is interested mainly in the former, not having been bedded in decades. She has sent Max $300 to get his passport renewed and she has promised to pay his roundtrip airfare to Tasmania (coach, we assume). To his friends' astonishment, Max is gearing up to go. Is his trust in Maggie misplaced? Does he trust himself to satisfy Maggie's needs when he meets her? Does he trust her not to cancel his return ticket once he gets there? Does he trust that

Maggie is, in fact, the woman she claims to be? As to Maggie, if she is not a man or a bot, does she trust Max to not abuse her, or disappoint her in less dramatic ways?

JIK heard this story from his friend S. Can he trust her? At any rate, some readers may shudder at this story, while Chekhov disconsolately turns in his grave. There is, of course, such a thing as excessive trust (Belfi et al., 2015; Krueger, 2011). One must remain vigilant for fear of getting exploited. This is why trusting others can never be a general social norm (as some have claimed, Dunning et al., 2014). If it were, consistent trustors would be transparently exploitable, and failures to trust would be punished by third parties – but it is not (Bicchieri et al., 2011). A weaker form of the moral-expectation hypothesis appears to hold however. Both trustors and trustees are seen as more moral inasmuch as the respectively entrust and give back more money and resources (Evans & van Calseyde, 2017; Krueger et al., 2008). Yet, some people trust too much relative to defensible standards of rational optimization (Evans & Krueger, 2016; Fetchenhauer & Dunning, 2009). By trusting others, people can build reputations of being moral, while ascriptions of competence are not affected.

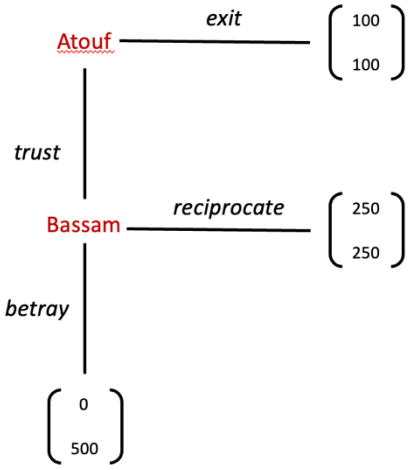
Interpersonal trust is a deep social dilemma; perhaps it is the *ur*-dilemma (Krueger & Evans, 2013). On one horn, trustors tolerate disappointment, exploitation, and regret, whereas on the other horn, they may expect to collect material, social, or emotional gains. Trust is a dilemma because it does not permit a clean quantitative risk assessment to ground rational choice (Krueger & Grüning, 2021). In its pure form, that is, under conditions of uncertainty (Knight, 1921), trustors have little guidance when trying to form expectations of the likelihood that their trust will be rewarded (Evans & Krueger, 2016). They have to replace probability with hope when accepting the vulnerability which trust demands (Luhmann, 1979; Rousseau et al., 1998; Thielmann & Hilbig, 2015).

The Game

In behavioral economics as well as in social and cognitive psychology, the Trust Game has been used as an experimental device to capture the dilemma's core (Berg et al., 1995; Evans & Krueger, 2009; Thielmann et al., 2021). The trustor, let's call him Atouf, and the trustee, let's call him

Bassam, have each received 100 Dirham from Soulaimane, a wealthy Berber Behaviorist. In the investment version of the game, Soulaimane allows Atouf to transfer any amount of his endowment to Bassam, which Soulaimane will then quadruple. Figure 1 illustrates the game in its simplified form, where the trustor either goes all in or not at all. If Atouf does nothing (i.e., exits the game), each player is left with a welcome fee of 100 Dirham. If, however, Atouf invests all into Bassam, that is, if he trusts him, Bassam has a dictator's choice between full reciprocation, which in the figure leads to the equitable outcome of 250 Dirham each, and betrayal, which leaves Bassam with the entire stock of money. In the investment version of the game, Atouf and Bassam can transfer partial amounts, a permission that allows Atouf to hedge his bet and Bassam to add nuance to his response.

Figure 1. The extensive-form trust game. Atouf is the trustor, Bassam the trustee.



The typical game is played once between strangers who cannot communicate. The game thus eliminates key elements of many ordinary situations: repeated play and assurance giving. Whether a trustee's promise of reciprocation is an essential feature of the definition of trust is a definitional matter (Hawley, 2012; 2014). Although promises, being commitment devices, tend to increase prosocial behavior (Belot et al., 2010; Orbell et al., 1988), a betrayal that comes in the wake of a promise is more painful than a mere lack of reciprocation (but see Fetchenhauer et al., 2020, for a skeptical note on betrayal aversion).

The Heuristics

Most people would prefer to live in a high-trust society (Schilke et al., 2021). The more limited trust is the social capital of trust, the more people must rely on others to emit credible and truthful signals of trustworthiness so that the loss that comes with misplaced trust can be limited. In this uncertain world, where others have an incentive to conceal their level of trustworthiness, where nonverbal signals are unavailable or hard to read, and where others may not be fellow humans but unsympathetic bots, in such a world, socially intelligent agents can rely on a suite of heuristics that improve, though not perfect, their chances of having rewarding exchange relationships (Hoffrage & Hertwig, 2012; Rand et al., 2014; van den Berg & Wenseleers, 2018). We review some of these heuristics in contexts where they can do some good, that is, in one-shot encounters with limited or no information about the other.

It should be noted that the list below is not exhaustive and that these heuristics are not mutually exclusive. Some share significant overlap in the cues they utilize. It should also be recognized that these heuristics do not obey the obsolescent two-systems framework of mind. Instead, these computationally simple and reasonable judgment tools fit well with the fast-and-frugal heuristics paradigm (Gigerenzer & Gaissmaier, 2011; Todd & Gigerenzer, 2012). Lastly, there is no structural order to the review below. We settled for an ad-hoc-list heuristic.

[1] ***Blind trust.*** The first heuristic may be called a no-cue decision rule. Using this rule, individuals will trust without seeking or using additional information. If relevant information were accessible, they might even choose to remain ignorant. Deliberate ignorance can be attained through processes of reflection or it may be an intuitive and spontaneous stance (Krueger et al., 2020). A blind trustor may be out of options or not able to see a choice to not trust. When explaining their no-alternative move after the fact, they can, however, rationalize it as a trusting choice. Otherwise, they blindly trust and hope for the best. One way to make sense of this tactic is to accept it as a form of generalized trust (van Lange, 2015). Some individuals sometimes accept vulnerability to strangers, hoping that their behavior reflects a common cultural value and practice. If they think more deeply, or

perhaps a bit magically, they can invoke the notion of karma. Good deeds will come around, they may reason, and to trust is to do a good deed. Not being able to play tit-for-tat in a one-shot game, thoughtful but blind trustors might settle for playing with one tit. As to the unreflected variant of blind trust, it has been suggested that some forms of cooperation are so ingrained by biological and cultural evolution that they come to mind as a first response. They need not be called up with effort, but they can be inhibited when they seem unwise (Evans & Rand, 2019).

[2] ***All or nothing*****.** In experimental work, the trust game is either set up to present trustors with an all-or-nothing option or it permits graded investments. The latter may be appealing to those who wish to hedge their bets, but logic and evidence suggest that this is a bad idea. Trustees take trustors beyond their break-even point only if these trustors invest all (Pillutla et al., 2003). Likewise, intermediate investing and reciprocating is difficult to map onto well-behaved utility curves, especially *a priori* (Krueger et al., 2008; but see below for some post-hoc modeling). Hence, a serviceable heuristic is to trust fully if one is going to trust at all. In other words, theory and evidence discourage the use of a hedge heuristic. In other words, with this heuristic, trustors treat the spectral investment game as if it were a binary trust game.

[3] ***Social projection.*** The heuristic of social projection lies at the core of much of social cognition. Since Floyd Allport (1924) first described projection as a source of crowd behavior, we have learned to appreciate projection as a window into the social world, and not just as a buggy false consensus effect (Krueger, 1998; 2007). The heuristic works well because by necessity most people are similar to most other people most of the time. The logic of this heuristic extends to situations such as the trust game, where the individual player has to predict the likely behavior of another player who occupies a different role with different demands and incentives. Hence, trustors need to engage in a two-step projection process. They need to ask themselves what they would do if they were in the trustee's shoes, and they then need to project this self-projection onto the other. As a product of two projections, the result is bound to be a small but positive correlation. The projection heuristic is boundedly rational because it beats a flip-a-coin heuristic (Evans et al., 2021).

[4] ***Social distance & categorization*****.** Rates of cooperation drop the farther the partner or beneficiary is removed from the self (Krueger et al., 2016). With cooperation being costly and risky, a simple biological heuristic of using perceived genetic similarity is sufficient to explain the basic finding (Dawkins, 1977; Hamilton, 1963). The distance heuristic moderates the projection heuristic (Krueger et al., 2024 ab; Robbins & Krueger, 2005. Projecting more strongly to similar than to dissimilar others (because they *are* more similar), people come to expect more cooperation, and then come to cooperate themselves even if cooperation is not a dominating choice in the game-theoretic sense (Krueger et al., 2012 ab). Social categorization makes social distance binary by creating ingroups and outgroups (Krueger & DiDonato, 2008). We are not surprised to learn that people trust ingroup members more than they trust outgroup members (Balliet et al., 2014).

[5] ***Social norms*****.** Interpersonal trust has some normative support, although the strong implication of norm violations being punished does not appear to hold (Bicchieri et al., 2011). Whether people overtrust by bending to normative demands even after other incentives and heuristics have been controlled remains an open question. A strong normative postulate of a 'thou-shalt-trust' commandment would certainly overreach (Moses knew this). Oddly enough, a categorical imperative treating trust as a duty would not be illogical. If everyone trusted everyone all the time, society might run smoothly, albeit boringly (Kant, 1785/2012). Alas, the point can be considered moot because empirically, it will never be the case that everyone trusts.

A more modest normative claim can be made on heuristic grounds. As the norm of reciprocity is strong enough to help many trustees overcome the temptation to defect, it can also inspire them to trust in turn, an effect that can reveal itself in repeated interactions among the same partners or even in a generalized sense – when the tit follows the tat. If Henry trusts James, James may then be willing to trust Charles (Nowak & Sigmund, 2005). In the organizational-psychology literature it is curious to see a myopic concern with the question of how employees' trust in their leaders can be strengthened. Why should we not ask how leaders can be brought to trust their employees more? Rabbinical sages and Taoist masters taught that to trust is to treat others with respect (Murnighan, 2012; reviewed in Krueger, 2013).

Respect is easily reciprocated, perhaps with a view toward karma. Indeed, people who are known to believe in karma are trusted more than those who do not (Ong et al., 2022). Of course, even in such a benign atmosphere there will be those who see and seize an opportunity to defect (Krueger, 2011). Hence the dilemma.

Before social norms are extolled too enthusiastically, it is well to remember that most norms are unabashedly parochial or interpreted as such (see heuristic [4]; Balliet et al., 2014). Risky and costly behavior, from kindness to cooperation, is less likely to be extended to outgroup than to ingroup members. It is harder to get people to adopt a *sub specie humanitatis* perspective than it is to get them stoked about their own tribe, class, or academic department. Peter Singer (1981) has tried, though.

[6] ***Reputational concerns.*** Humans care about how they are perceived, especially in their parochial world (Krueger et al., 2020). If they deny this they are probably lying. As fragile and interdependent creatures, humans are well advised to be vigilant and sensitive to how they come across. In social psychology, the joint study of person perception and impression management is as old as the discipline itself (Allport & Allport, 1921). The heuristic advice is to act in such a manner that one's reputation be enhanced. *Ceteris paribus*, trusting bestows more reputational benefits than distrusting, but only in the absence of strong cues discouraging trust. A cheated trustor appears a fool. In short, heuristics provide guidance under conditions of uncertainty, but if they are used in the face of countervailing signals, they can do more harm than good.

If a principled decision to trust opens the agent up to being exploited, a simple adjunct heuristic can help. If there is more than one trust game and if people are watching, trustors can partially randomize their behavior. This will keep trustees guessing and the audience from hastily assuming *naivité* (Grüning & Krueger, 2021; 2024). If all uncertainty were lost, so would trust.

[7] ***Social heuristics.*** When people are unsure about what to do, they can look to others for guidance. They can observe what most people do, what successful people do, what their friends do, and what experts and people with authority do (Hertwig & Herzog, 2009). These social-imitation heuristics can be expected to work more often than not. There is a price to pay, however. All these non-egocentric

heuristics cast the person in the role of the follower, leaving initiative and leadership to others. When everyone wants to be a heuristic follower, the few who resist this urge can call the shots, amass power, and abuse it. Much of the history of social psychology has consisted of writing the script against these heuristics (Krueger, 2012). From Sherif to Milgram via Asch, the moral tenor of social psychology has been a cry for independence. Do not follow the herd! Question authority! Make up your own mind and rationally so!

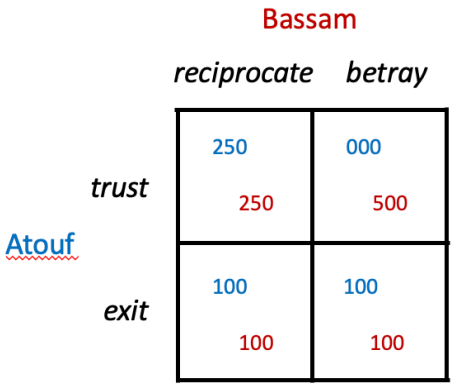
True progress can only be found if the proper balance of heuristic and analytic thinking is understood. We emphasize that social heuristics can mitigate the trust dilemma when more exhaustive evidence-methods fail or are not available. Returning to our opening example, it should be plain that Max has sufficient prima-facie evidence to distrust Maggie. If the desire for exotic travel keeps him from seeing this, heuristic consultations with his friends should tilt his mental balance away from this foolish adventure. When the trustworthiness of another person is dubious, asking others if they would trust that person goes a long way. The advice-taking literature suggests that it takes a strong majority opinion to overturn an agent's inclination, but with enough external input the wisdom of the crowd can assert itself (Kämmer et al., 2023). The inverse of Max's dilemma follows the same logic, but we suspect it is more rarely realized, that is, in situations where the agent is inclined to distrust, while social advisors and role models counsel trust.

[8] ***Trusting to learn*****.** In an earlier Sydney Symposium, we (Krueger & Grüning, 2023) explored an outcome asymmetry of trust and distrust. Heuristic trust is defensible on the expectation that whatever the outcome – reciprocity or betrayal – trustors gain information about their social world (Fiedler et al., 2023). Their expectations about the trustworthiness of others become more accurate and more calibrated than the expectations of those who rarely trust. Suspicious people are prone to remain suspicious because their expectations can never be falsified. High trustors accept the occasional painful disappointment, but low trustors rarely experience the rewards of social reciprocation. As Hoca Camide once said, "Trust is sustained by the occasional betrayal." The same logic explains religion. If God never betrayed man, we would all be atheists.

[9] *Egocentrism*. The fundamental egocentrism of the social mind as seen in social projection and in the underweighting of well-intentioned advice asserts itself in the trust game itself (Evans & Krueger, 2011; 2014) and other social dilemmas (Krueger et al., 2018). Trustors pay inordinate attention to their own potential payoffs, while neglecting the trustees' payoffs. This is a heuristic strategy because it uses some information while neglecting other. Egocentric selectivity performs better than blind trust or random trust, but it fails to attain optimal outcomes. In a trust game, where all payoffs are common knowledge, egocentrism is problematic because the trustee's incentive to act selfishly is in plain view, and it is critical for an accurate assessment of the likelihood of reciprocity. In this sense, the neglect of the trustee's incentives is an *a fortiori* finding. In the wild, trustees' incentives are often invisible, not in the least because the trustees themselves have incentives to conceal them (*caveat emptor,* Max!).

Egocentrism is a sensible shortcut when the trustees' interests are "encapsulated" in the trustors' interests (Hardin, 2006; Krueger, 2006; see also Balliet & Lindström, 2023). Children trust their parents because it is in their parents' interest not to kill or abandon them (exceptions exist, and not just at the criminal fringe, Kimbrough et al., 2021). In the one-shot trust game among strangers, however, interests are not encapsulated aside perhaps from shared reputational concerns. To see the misalignment, consider the game in matrix form. Assume that Bassam the trustee has pre-committed to reciprocation or betrayal before knowing Atouf the trustor's choice. The two players choose simultaneously. Correlating the two sets of outcomes, we find that r = -.42. With ranked outcomes the correlation drops to -.58.

Figure 2: The simultaneous-move trust game



[10] ***Defaults*****.** Conventionally, trustors make investments by moving some of their money to the trustee. This arrangement has ecological validity, but it has no claim to exclusivity. The money may as well initially rest with the trustee so that the trustor needs to pull back whatever they wish not to invest. Evans et al. (2011) found that a change in the default did not matter when trustors were rested and attentive. If they were mentally fatigued by a concurrent cognitive task, however, they moved less money away from the default. They trusted less in the standard game and they trusted more in the inverted game. Building on this work, we used two versions of the dictator game to see if respondents would rather trust a dictator to transfer money to them or a thief to take money away from them. We submit that respondents should trust dictators *less* because dictators can give money away only by overcoming their own endowment bias. We predicted, however, that respondents would trust dictators *more* due to an egocentric focus on their own potential gains and losses. Paired with a dictator, trustors can only gain; paired with a thief, they can only lose.

To close this brief review of trust heuristics, we need to ask if they are, one the whole, rather self-focused. Would better results be obtained if trustors tried harder to understand and predict the trustee's perspective? This is certainly so (Evans & Krueger, 2011), but perspective taking tends to be a more reflective and cognitively expensive enterprise. Still, trustors can make some progress in this direction by – paradoxically – using the social projection heuristic described above. By asking themselves whether they themselves would reciprocate trust, they need to think about what it would be like to be trustee, and look at the game from that vantage point.

A Simulation Experiment

A trustee is effectively a dictator (Hoffman et al., 1996; Krueger et al., 2017). After the trustor has made a move, the trustee can reflect on the trustor's intentions, frame of mind, and expectations. Trustees are not insensitive to the trustors having accepted vulnerability while benefitting the trustee. Average returns, although they let the trustor break even without enriching them, are larger than the average share of the money transferred by dictators in the eponymous game (Evans et al., 2013). Most subjects in the dictator game fare better than most trustors because they never have to invest any

endowment; they cannot incur a loss, only a lack of a gain. Conversely, dictators end up worse off than trustees. Any transfer a dictator makes is a loss. This is true for the trustee as well, but because most trustors make investments, which are multiplied in value (e.g., by Soulaimane the Munificent), and because most trustees return only the trustors' investment without sharing the capital gains, most trustees leave with more money than they start with.

These inequalities suggest that if players had to choose which game to play, they should be advised to prefer to play with a dictator game than with a trustee. At the same time, players should prefer to be trustees rather than dictators. This preference is supported by a power difference. A trustee's power is conditional; trustees can send a signal to the trustors based on what the trustors have done – with the caveat that if trustors invest nothing, the trustees' power is void. A dictator may seem to have absolute power, but this power is hollow because it cannot be construed as a reward or a punishment for the subject's antecedent behavior – there is none. In short, we submit that respondents will prefer to assume opposite roles in the two games. This is a hypothesis yet to be tested.

Subjects in a dictator game are reduced to a sort of impotent form of trust. They must accept dependence on a dictator's benevolence without having invested anything to make the dictator better off and to trigger the norm of reciprocity. If they receive a fair amount, they cannot reciprocate; if they receive offensively little, they cannot punish the dictator by refusing the deal, as a player in the ultimatum game can (Güth et al., 1982). Dictators' subjects are forced into a state of abject dependence; they cannot choose it. Yet, these subjects do not face the issue of loss aversion; investing nothing, they can lose nothing. We consider it likely that respondents would rather depend on a dictator than on a trustee. The former's decision may be disappointing, but only the latter's decision packs an emotional punch in addition to fixing the monetary outcome.

Like the trust game, the dictator game has a built-in default (Evans et al., 2011). The initial endowment lies with the dictator, who is free to share money with the subject. Any gain to the subject is a loss to the dictator. This is arbitrary, although it might reflect many situations in real social ecologies. The logic of the game does not change when the default is flipped. Consider the "thief

game." Here, the endowment lies with the subject, and the dictator (now "thief") can take whatever they wish, much like a taxman unplugged. Logic offers no reason why the final outcome should be any different between the two variants of this basic distribution game. Likewise, social preference theories or theories of social value orientation do not anticipate differential treatment. These theories interpret the amount transferred as a stable and trait-like expression of the player's benevolence or prosociality (Krueger et al., 2017), a point to which we will return later.

Some prominent approaches to social-cognitive psychology and behavioral economics suggest that players face different motivational pressures in the dictator and the thief game. Consider first the player in charge of the final distribution. As a dictator, Bassam (see Figure 1) faces a conflict between self-regard and benevolence. Having received 100 Dirham from Soulaimane, he will promptly view this money as his endowment; anything handed over to Atouf creates a sense of a loss, an experience only mitigated by the warm glow of giving and perhaps a self-congratulatory lift in his reputation as a generous person, at least in his own mind (Krueger et al., 2020). Self-regard tends to win, especially in one-shot games with strangers. The Bassams of the research world give between 25 and 30% on average (Doñate-Buendía et al., 2022). A subject (like Atouf) who knows his psychology should not expect more. In his role as a thief, however, Bassam has no endowment with which to part. Taking money from Atouf, Bassam can realize a gain, while at the same time achieving a more equitable distribution. It seems that a thief, who is not hampered by the pull of an endowment effect, will leave more money to the subject than a dictator will give. Stated differently, a thief will take less money away than a dictator would hold onto for himself. If this is so, and if players in the role of the subject know this, they should prefer to enter the distribution game with a preference to play with the thief.

To see if respondents express such a differential preference, we collected data from 224 respondents in an online survey conducted at Brown University in February of 2025. We asked them to consider being the subject, that is, the receiving player, in a dictator game and in a thief game. For each game, respondents made a prediction of how much money a random other person would transfer

to them or leave with them out of an endowment of $10, respectively as a dictator or as a thief. Then, they rated how willing they would be to play such a game, with ratings ranging from 1 (unwilling) to 7 (very willing). Notice that strict rationality assumes the subject to expect to receive no money at all (Binmore, 2007). Willingness ratings would then be indeterminate and could assume any value simply for the sake of any response being made. More realistically, respondents may assume receipts of $0 to be unlikely. If so, rationality would dictate invariant willingness ratings of 7 because the expected value of the game is likely positive and cannot be negative. For uniform maximum ratings, the correlations between predicted moneys and willingness ratings would also be indeterminate. Realistically, however, monetary expectations should be strong predictors of willingness to play, reflected in a positive correlation between these two measures.

The experimental design comprised two further prediction variables. All respondents were presented with one dictator/thief who was introduced as someone who believes in free will and rejects determinism and another player who was said (putatively on the basis of prior surveys) to reject free will and embrace determinism. Lastly, respondents themselves completed a short version of a two-facet scale tapping into free will belief and determinism (Nadelhoffer et al., 2014; see also Krueger & Grüning, 2023).

To review the main findings, we begin by noting that neither the respondents' own free will belief nor the other players' purported belief had any effect on the outcome variables. These null findings may seem surprising because some research suggests that free will believers are more cooperative than determinists, and might therefore also be more trusting. Our null findings are rather aligned, however, with the skeptical work on this matter (Buttrick et al., 2020; Krueger & Grüning, 2025; Sapolsky, 2023; reviewed in Krueger & Grüning, 2024).

Critically, respondents expected to take home more money in the dictator game than in the thief game, and they were more willing to enter the former than the latter game. Figures 3 and 4 show the means and the statistical evaluation for these results, and they also show that the other players' free-will beliefs made no discernible difference.

Figure 3: Money expected from dictators and thieves

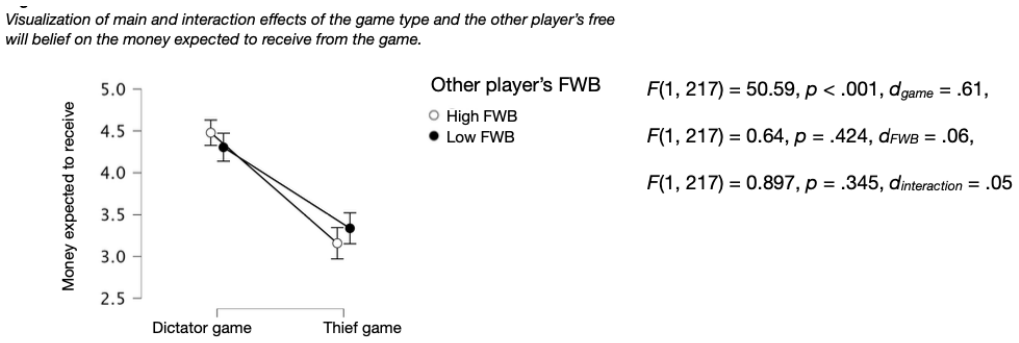
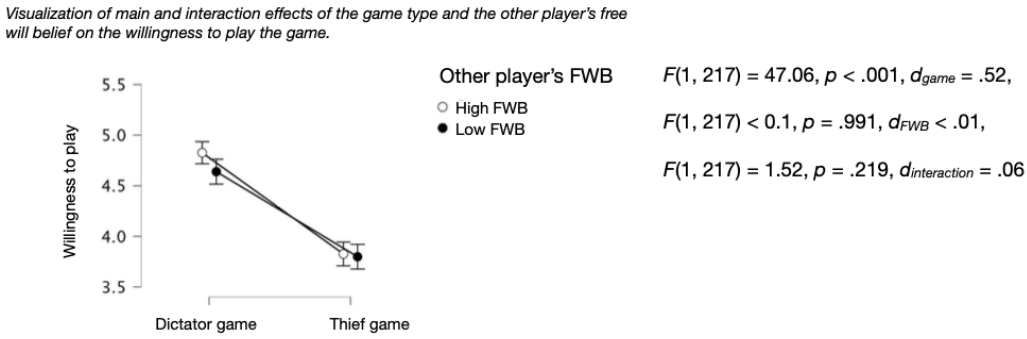
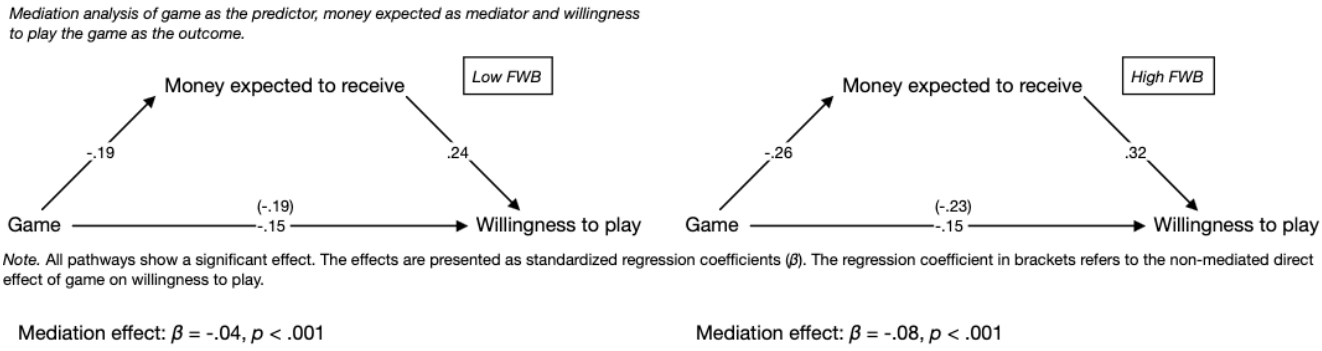


Figure 4: Willingness to play with dictators and thieves



The correlations between expected returns and willingness to play were positive in each of the four conditions (dictator vs thief, high vs low other's belief in free will) but they were not large, ranging from $r = .20$ to .34. Regression analyses, displayed in Figure 5, yielded evidence for partial but strong mediation regardless of whether the other player was said to be a free will believer or a determinist. Still, respondents were more willing to engage with a dictator than a thief even when their own monetary expectations were controlled. We suspect that this differential reflects loss aversion. In the thief game, respondents are the initial holders of the endowment. When thieves take money away, respondents experience a loss. By contrast, when dictators transfer some of their endowment, respondents experience a gain. Loss aversion, which is a stylized though not uncontested fact in behavioral economics, says that a monetary transfer causes more delight and more distress respectively in the dictator and in the thief game (Brown et al., 2024).

Figure 5: Willingness to play depend on the role of the other and expected returns



Note. All pathways show a significant effect. The effects are presented as standardized regression coefficients (β). The regression coefficient in brackets refers to the non-mediated direct effect of game on willingness to play.

Mediation effect: $\beta = -.04$, $p < .001$

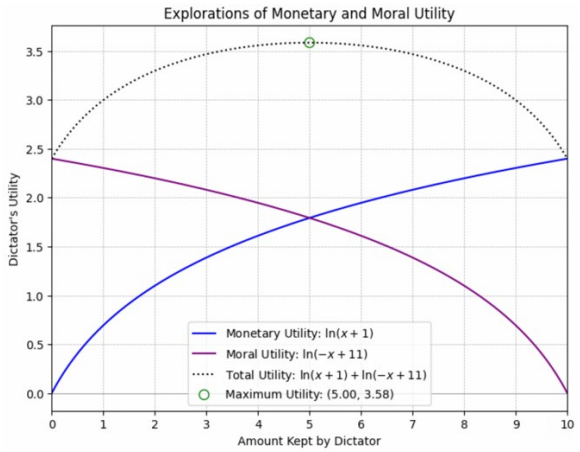
Mediation effect: $\beta = -.08$, $p < .001$

We consider it likely – although this has not yet been shown – that dictators are less generous than thieves. Dictators have to overcome their own endowment effect when making a prosocial transfer. Thieves, by contrast, face no endowment effect; they have to pull money from the subjects to enrich themselves. Thieves may see a 50-50 split as a line difficult cross. By contrast, dictators may find this same line difficult to reach. If so, it would be all the more interesting that respondents put more trust in a dictator. Such differential trust would be irrational in the sense of violating subjects' own interests, though it would be explicable in light of the general difficulty of taking the perspectives of a player whose interests are not aligned with one's own (Evans & Krueger, 2011). A comparatively greater trust in dictators and in thieves is socially myopic.

We now return to the question of how social preferences might explain dictators' or trustees' allocation decisions. It has been shown mathematically that with linear scaling, dictators or trustees who consider and weigh their own and the subject's outcomes can maximize their expected values only by either transferring no money at all or by transferring the amount that makes both players equally well off (Krueger et al., 2008). An exception is a dictator/trustee who cares exactly as much about the other's welfare as about their own. This person would be indifferent to the final distribution. It was also shown that when a dictator/trustee's money outcome is logarithmically scaled to reflect diminishing utilities for larger amounts, and if transfers are inversely scaled so that increased giving (their own losses) is more strongly felt at small than at large amounts, the net utility function, that is,

the summed utility of having and giving, is an inverted U, peaking at an intermediate compromise point. Figure 6 displays this arrangement.

Figure 6: A dictator/trustee's net utility function (dotted line) as a composite of the utility of the money retained (blue line) and the money transferred (purple line).



Note: Both component functions assume diminishing marginal utility. When little is transferred the experienced gain to the other is greater than the loss to the self (right half of the graph) with the reverse being true when much is transferred (left half). In this display, the averaging weights for self and other are both 1.0.

The maximum composite utility for the dictator assuming they care as much positively about their own outcome as they care negatively about the subject's outcome, and assuming a natural log transformation, is half of the endowment, as proven here:

$$Total\ Utility = \ln(x+1) + \ln(-x+11)$$

Taking the derivative and setting it equal to zero to find the maximum:

$$\frac{d}{dx}\left(\ln(x+1) + \ln(-x+11)\right) = \frac{1}{x+1} - \frac{1}{-x+11} = 0$$

$$\frac{1}{x+1} = \frac{1}{-x+11} \quad (1)$$

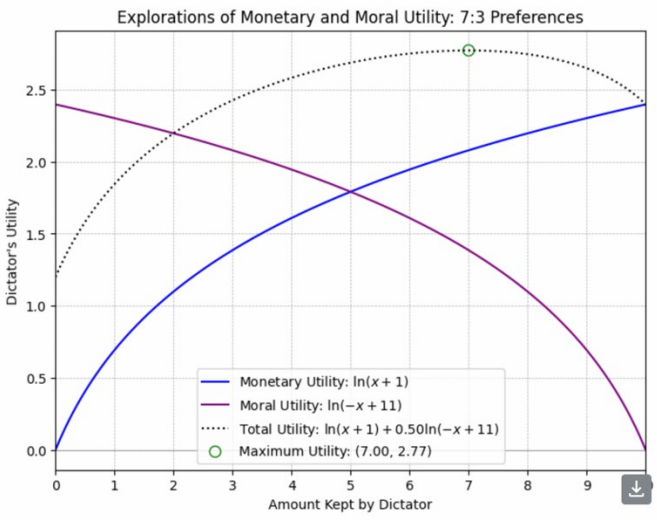
$$-x+11 = x+1 \quad (2)$$

$$2x = 10 \quad (3)$$

$$x = 5 \quad (4)$$

Therefore, under these conditions, a dictator's maximum utility occurs under an equal split

Figure 7: A dictator/trustee's net utility function (dotted line) peaking at a transfer of $3 to the subject.



The utility function for the dictator's own outcome is the same as in Figure 6, but the weight of benevolence is half of the egocentric weight. The function for the subject's outcome (i.e., the dictator's morality function) now drops off more sharply for small transfers to obtain this result, as proven here:

Taking the derivative and setting it equal to zero to find the maximum:

$$\frac{d}{dx}\left(\ln(x+1)+b\ln(-x+11)\right)=\frac{1}{x+1}-\frac{b}{-x+11}=0$$

$$\frac{1}{x+1}=\frac{b}{-x+11} \quad (1)$$

$$-x+11=b(x+1) \quad (2)$$

$$-x+11=bx+b \quad (3)$$

$$bx+x=11-b \quad (4)$$

$$x(b+1)=11-b \quad (5)$$

$$x=\frac{11-b}{b+1} \quad (6)$$

If we want maximum x to be 7:

$$7=\frac{11-b}{b+1} \quad (7)$$

$$7(b+1)=11-b \quad (8)$$

$$7b+7=11-b \quad (9)$$

$$8b=4 \quad (10)$$

$$b=\frac{4}{8}=0.50 \quad (11)$$

Therefore, for a person who derives maximum utility from a 7:3 split, they will continue to weigh their monetary utility as they did before, while decreasing the weight of their moral utility by a half:

$$Total\ Utilitu=\ln(x+1)+0.50\ln(-x+11)$$

Outlook

As the trait of trustworthiness ranks near the top in terms of social desirability (Britz et al., 2023), the wish and the need to trust come into focus. If trust were a given, however, it would eliminate itself. With all uncertainty gone, there would be no desire, no regret, no fear. Trust becomes an issue of importance because its attainment is – and will always remain – uncertain. The fragility of trust and trustworthiness add spice to life on a good day, while adding to civilization's discontent on most others.

To those who were hoping to receive a recipe allowing them to decoct social harmony by burning away all uncertainty, our analysis must be a disappointment. Perhaps some readers feel that their trust in us has been betrayed. "Weren't you," they might sigh, "supposed to show us when and how and whom to trust? What good is your science if you can't do that?" To these disconsolates we say that much like quantum physics, the study of human behavior does not yield certainties, but it provides a frame of reference and tools for managing and perhaps reducing uncertainty. And it is not for nothing that the stubborn persistence of the trust dilemma protects us from getting bored.

References

Allport, F. H. (1924). *Social psychology*. Houghton Mifflin.

Allport, F. H., & Allport, G. W. (1921). Personality traits: Their classification and measurement. *Journal of Abnormal Psychology and Social Psychology, 16*, 6–40.

Balliet, D., & Lindström, B. (2023). Inferences about interdependence shape cooperation. *Trends in Cognitive Sciences,27,* 583–595.

Balliet, D., Wu, D., & de Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140*, 1556-1581.

Belfi, M., Koscik, T. R., & Tranel, D. (2015). Damage to the insula is associated with abnormal interpersonal trust. *Neuropsychologia, 71*, 165-172.

Belot, M., Bhaskar V., & van de Ven, J. (2010). Promises and cooperation: Evidence from a TV game show. *Journal of Economic Behavior & Organiation*, *73*, 396–405.

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

Bicchieri, C., Xiao, E., & Muldoon, R. (2011). Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy & Economics*, *10*, 170–187.

Binmore, K. (2007). *Game theory: A very brief introduction*. Oxford University Press.

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior and Organization, 55*, 467-484.

Britz, S., Rader, L., Gauggel, S., & Mainz, V. (2023). An English list of trait words including valence, social desirability, and observability ratings. *Behavior Research Methods, 55,* 2669-2686.

Brown, A. L., Imai, T., Vieider, F. M., & Camerer, C. F. (2024). Meta-analysis of empirical estimates of loss aversion. *Journal of Economic Literature 62*, 485–516.

Buttrick, N. R., Aczel, B., Aeschbach, L. F., Bakos, B. E., Brühlmann, F., Claypool, H. M. *et al.* (2020). Many labs 5: Registered replication of Vohs and Schooler (2008), experiment 1 *Advances in Methods and Practices in Psychological Science, 3,* 429-438.

Dawkins, R. (1977). *The selfish gene*. Oxford University Press.

Doñate-Buendía, A., García-Gallego, A., & Petrović, M. (2022). Gender and other moderators of giving in the dictator game: A meta-analysis. *Journal of Economic Behavior & Organization, 198,* 280-301.

Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology, 107*, 122-141.

Evans, A. M., & Krueger, J. I. (2009). The psychology (and economics) of trust. *Social and Personality Psychology Compass: Intrapersonal Processes, 3*, 1003-1017. doi:10.1111/j.1751-9004.2009.00232.x

Evans, A. M., & Krueger, J. I. (2011). Elements of trust: Risk taking and expectation of reciprocity. *Journal of Experimental Social Psychology*, *47*, 171-177. doi:10.1016/j.jesp.2010.08.007

Evans, A. M., & Krueger, J. I. (2014). Outcomes and expectations in dilemmas of trust. *Judgment & Decision Making, 9*, 90-103. http://journal.sjdm.org/13/13502/jdm13502.pdf

Evans, A. M., & Krueger, J. I. (2015). The edge of trust: An introduction. *Social Cognition, 33,* 359-367. doi: 10.1521/soco.2015.33.5.359 [introduction to special issue]

Evans, A. M., & Krueger, J. I. (2016). Bounded prospection in dilemmas of trust and reciprocity. *Review of General Psychology, 20,* 17-28. doi.org/10.1037/gpr0000063

Evans, A. M., & Krueger, J. K. (2017). Ambiguity and expectation-neglect in dilemmas of trust. *Judgment & Decision Making, 12,* 584-595. http://journal.sjdm.org/17/17131/jdm17131.pdf

Evans, A. M., Athenstaedt, U., & Krueger, J. I. (2013). The development of trust and altruism during childhood. *Journal of Economic Psychology, 36,* 82-95. doi.org/10.1016/j.joep.2013.02.010

Evans, A. M., Dillon, K. D., Goldin, G., & Krueger, J. I. (2011). Trust and self-control: The moderating role of the default. *Judgment and Decision Making, 6*, 697-705.

Evans, A. M., Ong, H. H., & Krueger, J. I. (2021). Social proximity and respect for norms in trust dilemmas. *Journal of Behavioral Decision Making, 35,* 657-668

Evans, A. M., & Rand, D. G. (2019). Cooperation and decision time. *Current Opinion in Psychology, 26*, 67–71.

Evans, A. M., & van de Calseyde, P. P. F. M. (2017). The reputational consequences of generalized trust. *Personality and Social Psychology Bulletin*, *44*, 492-507.

Fetchenhauer, D., & Dunning, D. A. (2009). Do people trust too much or too little? *Journal of Economic Psychology, 30*, 263–276.

Fetchenhauer, D., Lang, A. S., Ehlebracht, D., Schlösser, T., & Dunning, D. (2020). Does betrayal aversion really guide trust decisions towards strangers? *Journal of Behavioral Decision Making, 33*, 556-566.

Fiedler, K., Juslin, P., & Denrell, J. (2023). *Sampling in judgment and decision making*. Cambridge University Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482.

Grüning, D. J., & Krueger, J. I. (2021). Strategic thinking: A random walk into the rabbit hole. *Collabra Psychology*, *7*(1), 24921. https://doi.org/10.1525/collabra.24921

Grüning, D. J., & Krueger, J. I. (2024). Strategic reasoning in the shadow of self-enhancement: Benefits and costs. *British Journal of Social Psychology, 63,* 1725-1742

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*, 367-388.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *American Naturalist, 97*, 354-56.

Hardin, R. (2006). *Trust*. Polity Press.

Hawley, K. (2012). *Trust: A very short introduction*. Oxford, UK: Oxford University Press.

Hawley, K. (2014). Trust, distrust and commitment. *Noûs*, 48, 1-20.

Hertwig, R., & Herzog, S. M. (2009). Fast and frugal heuristics: Tools of social rationality. *Social Cognition, 27*, 661–698.

Hoffman, E., McCabe, K., & Smith, V. (1996). Social distance and other-regarding behavior in dictator games. The American Economic Review, 86, 653-660.

Hoffrage, U., & Hertwig, R. (2012). Simple heuristics in a complex social world. In J. I. Krueger (ed.) *Social Judgment and Decision Making*, pp. 135–150. Psychology Press.

Kämmer, J., Choshen-Hillel, S., Müller-Trede, J., Black, S., & Weibler, J. (2023). A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision, 10*, 107–137.

Kant, I. (1785/2012). *Groundwork of the metaphysic of morals*. Edited and translated by M. Gregor & J. Timmermann. Cambridge University Press.

Kimbrough, E. O., Myers, G. M., & Robson, A. J. (2021). Infanticide and human self-domestication. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.667334

Knight, F. H. (1921). *Risk, uncertainty and profit*. Houghton Mifflin.

Krueger, J. (1998). On the perception of social consensus. *Advances in experimental social psychology, 30*, 163-240. Academic Press.

Krueger, J. I., (2006). Trusting Calvin and Hobbes. Review of 'Cooperation without trust?' by K. S. Cook, R. Hardin, & M. Levi. *PsycCRITIQUES- Contemporary Psychology: APA Review of Books, 51*(8), article 17.

Krueger, J. I. (2007). From social projection to social behaviour. *European Review of Social Psychology, 18,* 1-35.

Krueger, J. I. (2011). Altruism gone mad. In B. Oakley, A. Knafo, G. Madhavan, & D. S. Wilson (Eds.), *Pathological altruism* (pp. 392-402). Oxford University Press.

Krueger, J. I. (2012). The (ir)rationality project in social psychology: A review and assessment. In J. I. Krueger (Ed.). *Social judgment and decision-making* (pp. 59-75). Psychology Press.

Krueger, J. I. (2013). The new Tao of leadership. Review of 'Do nothing! How to stop overmanaging and become a great leader' by J. K. Murnighan. *Journal of Economic Psychology, 35*, 108-109.

Krueger, J. I., & DiDonato, T. E. (2008). Social categorization and the perception of groups and group differences. *Social and Personality Psychology Compass: Group Processes, 2*, 733-750.

Krueger, J. I., DiDonato, T. E., & Freestone, D. (2012). Social projection can solve social dilemmas. *Psychological Inquiry, 23,* 1-27.

Krueger, J. I., & Evans, A. M. (2013). Fiducia: Il dilemma sociale essenziale / Trust: The essential social dilemma. *In-Mind: Italy, 5*, 13-18. http://www.tonymevans.com/wp-content/uploads/2015/07/krueger-evans-2013.pdf

Krueger, J. I., Evans, A. M., & Heck, P. R. (2017). Let me help you help me: Trust between profit and prosociality. In P. A. M. Van Lange, B. Rockenbach, & T. Yamagishi (Eds.), *Social dilemmas: New perspectives on trust* (pp. 121-138). Oxford University Press.

Krueger, J. I., Freestone, D., & DiDonato, T. E. (2012). Twilight of a dilemma: A réplique. *Psychological Inquiry, 23,* 85-100.

Krueger, J. I., & Grüning, D. J. (2021). Psychological perversities and populism. In J. P. Forgas, W. D. Crano, & K. Fiedler (eds.). *The social psychology of populism: The tribal challenge to liberal democracy. The Sydney Symposium on Social Psychology, 22,* 125-142. Taylor & Francis.

Krueger, J. I., & Grüning, D. J. (2023). Strategy, trust, and freedom in an uncertain world. In J. P. Forgas, W. D. Crano, & K. Fiedler (eds.), *The psychology of insecurity*. *The Sydney Symposium on Social Psychology, 24,* 150-169. Routledge.

Krueger, J. I., & Grüning, D. J. (2024). The unceremonious death of free will. Review of 'Determined: A science of life without free will' by Robert M. Sapolsky. *American Journal of Psychology, 137,* 93-97.

Krueger, J. I., & Grüning, D. J. (2025). The false belief in free will. In J. P. Forgas (ed.), *The psychology of false beliefs. The Sydney Symposium of Social Psychology, 26,* 101-118.

Krueger, J. I., Grüning, D. J., Heck, P. R., & Freestone, D. (2024a). Inductive reasoning model. *Psychological Inquiry*, *resume*, 11-25.

Krueger, J. I., Grüning, D. J., Heck, P. R., & Freestone, D. (2024b). Inductive reasoning renewed: A reply to commentators. *Psychological Inquiry*, *35*, 69–79.

Krueger, J. I., Hahn, U., Ellerbrock, D., Gächter, S., Hertwig, R., Kornhauser, L. A., Leuker, C., Szech, N., & Waldmann, M. R. (2020). Normative implications of deliberate ignorance. In R. Hertwig & C. Engel (Eds.) Deliberate ignorance: Choosing not to know. *Strüngmann Forum Reports*, *29*, 257-287. MIT Press

Krueger, J. I., Heck, P. R., Evans, A. M., & DiDonato, T. E. (2020). Social game theory: Preferences, perceptions, and choices. *European Review of Social Psychology, 31*, 322-353.

Krueger, J. I., Heck, P. R., & Wagner, D. (2018). Egocentrism in the volunteer's dilemma. *American Journal of Psychology, 131*, 403-415.

Krueger, J. I., Massey, A. L., & DiDonato, T. E. (2008). A matter of trust: From social preferences to the strategic adherence to social norms. *Negotiation & Conflict Management Research, 1*, 31-52. doi:10.1111/j.1750-4716.2007.00003.x

Krueger, J. I., Ullrich, J., & Chen, L. J. (2016). Expectations and decisions in the volunteer's dilemma: effects of social distance and social projection. *Frontiers in Psychology: Cognition, 7,* article 1909. doi: 10.3389/fpsyg.2016.01909

Krueger, J. I., Vohs, K. D., & Baumeister, R. F. (2008). Is the allure of self-esteem a mirage after all? *American Psychologist, 63*, 64-65.

Luhmann, N. (1979). *Trust and power*. Wiley.

Misztal, B. A. (2001) Normality and trust in Goffman's theory of interaction order. *Sociological Theory, 19,* 312–324.

Murnighan, J. K. (2012). *Do nothing! How to stop overmanaging and become a great leader*. Penguin.

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. T. (2014). The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and cognition*, *25*, 27-41.

Nowak, M., Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*, 1291–1298.

Ong, H. H., Evans, A. M., Nelissen, R. M., & van Beest, I. (2022). Belief in karma is associated with perceived (but not actual) trustworthiness. *Judgment and Decision Making*, *17*(2), 362-377.

Orbell, J. M., Van de Kragt, A. J., & Dawes, R. M. (1988). Explaining discussion-induced cooperation. *Journal of Personality and Social Psychology, 54*, 811–819.

Pillutla, M. M., Malhotra, D., & Murnighan, J. K. (2003). Attributions of trust and the calculus of reciprocity. *Journal of Experimental Social Psychology*, *39*, 448-455.

Rand, D., Peysakhovich, A., Kraft-Todd, G. *et al.* (2014). Social heuristics shape intuitive cooperation. *Nature Communications,* 5, 3677. https://doi.org/10.1038/ncomms4677

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review, 9*, 32-47.

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

Sapolsky, R. M. (2023). *Determined: A science of life without free will*. Penguin.

Schilke, O., Reimann, M., & Cook, K. S. (2021). Trust in social relations. *Annual Review of Sociology, 47,* 239-259.

Singer P. (1981). *The expanding circle*. Clarendon Press.

Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021). Economic games: An introduction and guide for research. *Collabra: Psychology,7* (1): 19004. https://doi.org/10.1525/collabra.19004

Thielmann, I., & Hilbig, B. E. (2015). Trust: An integrative review from a person-situation perspective. *Review of General Psychology*, *19*, 249–277.

Todd, P. M. & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. Oxford University Press.

van den Berg, & P., Wenseleers, T. (2018). Uncertainty about social interactions leads to the evolution of social heuristics. *Nature Communications, 9*, 2151. https://doi.org/10.1038/s41467-018-04493-1

van Lange, P. A. M. (2015). Generalized trust: Four lessons from genetics and culture. *Current Directions in Psychological Science*, *24*, 71-76.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.