Trustworthiness: An adaptationist account

Laureon A. Merrie[1, 2*], Jaimie Arona Krems[1, 3], Daniel Sznycer,[1, 2] & Nina N. Rodriguez[1, 3]

[1] The Oklahoma Center for Evolutionary Analysis (OCEAN), Oklahoma State University

[2] Department of Psychology, Oklahoma State University

[3] Department of Psychology, University of California, Los Angeles

Accepted for publication in *Evolution and Human Behavior*

Author Note

LAM ORCID: https://orcid.org/0000-0003-3562-6946

JAK ORCID: https://orcid.org/0000-0002-2590-2241

DS ORCID: https://orcid.org/0000-0002-6510-3313

NNR ORCID: https://orcid.org/0009-0005-0715-6934

Correspondence concerning this article should be addressed to Laureon Merrie, laureon.merrie@gmail.com.

This material is based upon work supported by the National Science Foundation under Grant No. 2340942 awarded to Krems.

Abstract

The concept of TRUSTWORTHINESS plays a role in the formation, maintenance, and dissolution of friendships, marriages, and cooperative relationships from small to large scales. Here, we analyze TRUSTWORTHINESS under the assumption that such concepts evolved to guide action adaptively. Intuition and research suggest that actors trust targets who have not engaged in betrayals. However, this perspective fails to capture certain real-world behaviors (e.g., when two people cheating on their spouses enter a relationship with each other and expect mutual fidelity). Evolutionary task analysis suggests that TRUSTWORTHINESS is structured to help actors address challenges of extending trust, where actors may gain or lose from doing so. In six experiments with American adults (N=1,718), we test the hypothesis that TRUSTWORTHINESS tracks not only (i) whether targets refrain from betraying trust when given opportunities, but also (ii) the impact of betrayal on the actor. Data generally support this hypothesis across relationships (friendships, romantic, professional): Actors deem non-betrayers more trustworthy than betrayers, but also deem betrayers more trustworthy when betrayals benefit actors. TRUSTWORTHINESS may incline actors to trust to those who refrain from betraying others—a potent signal of reluctance to betray oneself—while also favoring those who betray others if it serves oneself.

Keywords: Trust, close relationships, person perception, cooperation

Trustworthiness: An adaptationist account

Trust facilitates friendships, romantic relationships, and cooperation. But misplaced trust can lead to being exploited and the attendant fitness costs (Cosmides & Tooby, 2015; Yamagishi, 2011). Lacking omniscience and time-travel, humans must decide whether to extend trust based on imperfect information—whatever information is available to the actor at the moment of decision. Here, we ask how people make these inferences. Specifically, we study the information the mind uses to judge a target as worthy of receiving one's trust.

Intuition and research alike suggest that the key driver for inferences of a target's trustworthiness is observed, reputational, or other information about that target's behavior—whether, given the opportunity, they have previously broken trust (Dasgupta, 1988; Krasnow et al., 2012; Roberts, 2020). Because past betrayal bodes future betrayal—or, in popular parlance, "once a cheater, always a cheater"—the focus on this driver implies that people do not trust those who have exploited, betrayed, or otherwise broken trust.

Here, we propose that trustworthiness judgments are calibrated not only to a target's record of betraying trust or not, but also to the expected impact on the actor of a target's betrayal. From an adaptationist perspective, the mind is, to an important extent, a constellation of neurocognitive machines specialized to solve recurrent adaptive problems (Tooby & Cosmides, 1992). Such machines exist because, on average, they promoted their own replication in ancestral environments. From this perspective, the function of TRUSTWORTHINESS is not to reflect the world objectively but to guide behavior in ways that benefit the actor (Delton & Sell, 2014).

Dissection of the concept TRUSTWORTHINESS may thus reveal features that aid actors in solving the challenges posed by trust extension, wherein actors can gain if trust is extended properly (e.g., if partners reciprocate trust), and lose if partners fail to do so (e.g., if partners

betray). For example, TRUSTWORTHINESS may track near-objective attributes such as whether a target refrains from betraying trust when given the opportunity. But, in addition, TRUSTWORTHINESS may parse the world idiosyncratically—for example, based on whether the target acts in ways that benefit *the actor* in trust-relevant events. I see Tom as trustworthy if Tom is trustworthy *to me*, similar to how people see a target as being kind if the target is kind *to them* (Krems et al., 2024; Lukaszewski & Roney, 2010; Merrie et al., 2024). In this way, TRUSTWORTHINESS may incline people to extend trust to those who refrain from betraying trust as well as those who betray *other people's* trust, especially *in one's favor*.

Consider a situation wherein Anne betrays Bai to their mutual friend Cara, revealing Bai's secret crush on Cara's spouse. Anne's betrayal harms Bai, but benefits Cara (Hess & Hagen, 2002, 2023). Likewise, Anne's betrayal of Bai to Cara implies that Anne trusts Cara (e.g., not to tattle to Bai) and prefers Cara over Bai (Krems et al., 2024)—information that is potentially useful to Cara in navigating the social landscape (Basyouni & Parkinson, 2022; Bedrov et al., 2021; DeScioli & Kurzban, 2009).

There is evidence that, in addition to disposition-based inferences about targets (e.g., Are they trustworthy *in general*?), people also make relationship-specific, actor-centric inferences (e.g., Can *I* trust *them*?) (Krems et al., 2023; Lukaszewski & Roney, 2010; Shaw et al., 2017; Yamagishi, 2011). The latter inferences may more closely track the actor's outcomes if the actor were to trust that target. Real-world behavior also seems sensitive to such information. For example, people sometimes upregulate their liking and trust in targets after targets' betrayals of others to oneself (Fonseca & Peters, 2018; Peters et al., 2017). Indeed, everyday sociality, history, and fiction are all replete with examples of people trusting targets who betray *others*—as when two lovers cheat on their spouses and expect one another's everlasting fidelity.

Overview

Taken together, a target's reluctance to betray trust may not be the sole driver of judgments of that target's trustworthiness. This is not to argue that such judgments are insensitive to targets' betrayals. Rather, insofar as such betrayals are valuable in predicting a target's future behavior—and specifically augur exploitation of oneself—the mind should integrate such information into trustworthiness inferences. But not all betrayals are equal. For example, a target's history of betraying others does not *necessarily* increase that target's likelihood of betraying oneself (see Krasnow et al., 2016). Indeed, targets' acts of betrayal can sometimes benefit actors—as when a man leaves his wife for the actor, when an actor's good friend shares a mutual friend's secret to that actor, or when a target betrays one class of people (e.g., the actor's outgroup enemies) in favor of another (e.g., the actor's ingroup allies). Thus, if concepts serve to guide action adaptively, judgments of target trustworthiness may be sensitive to cues of the expected impact on the self of trusting that target (or not).

This logic predicts that a target will be deemed less trustworthy when they betray someone's trust given the opportunity, compared to when they do not. But in addition, when the target does betray, they will be deemed relatively more trustworthy if their betrayal generates benefits for the actor, rather than costs or no benefits. Because trust is important across relationship domains (Cottrell et al., 2007; Yamagishi, 2011), we test predictions across three relationship contexts: friendly, romantic, professional.

Data and code: https://osf.io/8rytz/?view_only=d905cea70de142eb90860862ec16353c.

Experiments 1-3

We test whether people deem targets more trustworthy when (1) targets eschew (versus commit) betrayal, and (2) targets' betrayal benefits (versus harms) oneself.

Method

Participants

We aimed to recruit 303 U.S.-residing participants from CloudResearch per experiment for sufficient power to detect effects of f=.18. We excluded likely bots (via ReCaptcha), participants failing an attention check, and/or those reporting different sexes at survey start and end. See Table 1.

Table 1

Sample Characteristics for Experiments 1-3

	Recruited *N*	Final *N*	Sex	M_{age}	SD_{age}
Experiment 1	329	232	56.9% female	38.73	11.49
Experiment 2	332	227	65.2% female	38.42	12.38
Experiment 3	329	181	58.3% female	37.67	12.39

Design and Procedure

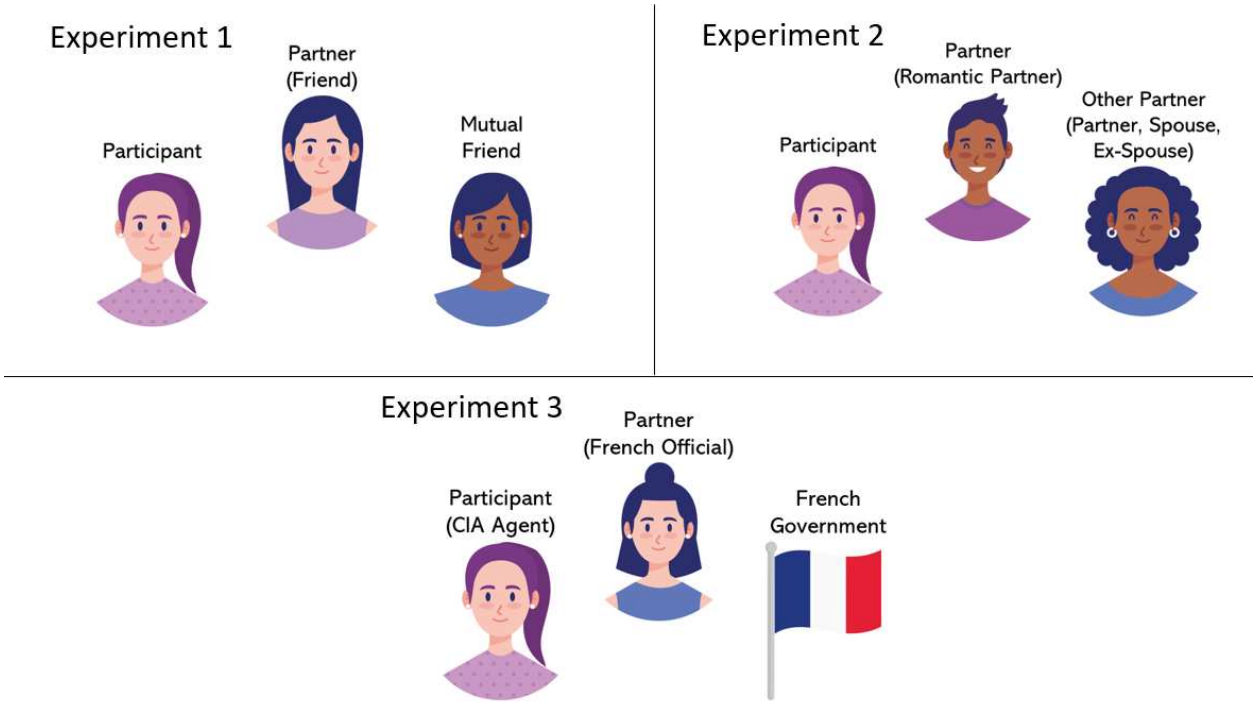
In each experiment, participants were randomly assigned to read one of three vignettes describing their interaction with a target (the Partner). Experiment 1 described sharing secrets among friends (Figure 1), Experiment 2 described romantic infidelity, and Experiment 3 described an interaction in the context of international relations, with participants acting as CIA agents attempting to cultivate a French official as a source.

Across experiments, the target (i) did not betray anyone when given the opportunity (e.g., did not share another friend's secret with participants), (ii) betrayed another person to the participant (e.g., shared another friend's secret with the participant), or (iii) betrayed the

participant to another (e.g., shared the participant's secrets with another friend). See Supplementary Materials for vignettes. Across conditions, participants learned the same information (e.g., participants learned their mutual friend's secret in every friendship condition).

Figure 1

Characters in Experiments 1-3



After reading vignettes, participants rated target trustworthiness (e.g., "I would trust [Partner] to keep my secrets") with six items on 7-point Likert scales (1=*Definitely not;* 7=*Definitely so*) (αs=.96-.98).[1]

[1] We also assessed how much participants believed targets valued them relative to the other character, generally finding that participants believe (a) non-betrayers value people they refuse to

Results

Do People Deem Non-Betrayers More Trustworthy Than Betrayers?

Yes. One-way analyses of variance (ANOVAs) compared target (i.e., Partner) trustworthiness across conditions, finding significant main effects in each experiment (ps<.001, η_p^2s≥.222). See Supplementary Materials for full analyses. Participants deemed non-betraying targets more trustworthy than targets who betrayed them or betrayed others to them. See Tables 2-3 and Figure 2.

Do People Deem Targets Whose Betrayal Benefits (vs. Costs) Them More Trustworthy?

Yes. Participants deemed targets more trustworthy when benefitting from the betrayal (Targets betrayed another in participants' favor) versus not (Targets betrayed participants in another's favor). See Tables 2-3, Figure 2.

---

betray more than people they could have betrayed *to*, and (b) betrayers value those *to whom* they betray more than those they betrayed (Supplementary Materials). Other exploratory measures not analyzed are available on OSF.

Table 2

Means (SEs) for Target (i.e., Partner) Trustworthiness Across Conditions in Experiments 1-3

	Non-betrayal (Partner Does Not Betray)	**Partner Betrays Other Party**	**Partner Betrays Participant**
Experiment 1	4.78(.16)	3.18(.15)	2.35(.15)
Experiment 2	5.55(.15)	2.80(.16)	1.79(.28)
Experiment 3	4.27(.19)	2.96(.23)	2.26(.22)

Table 3

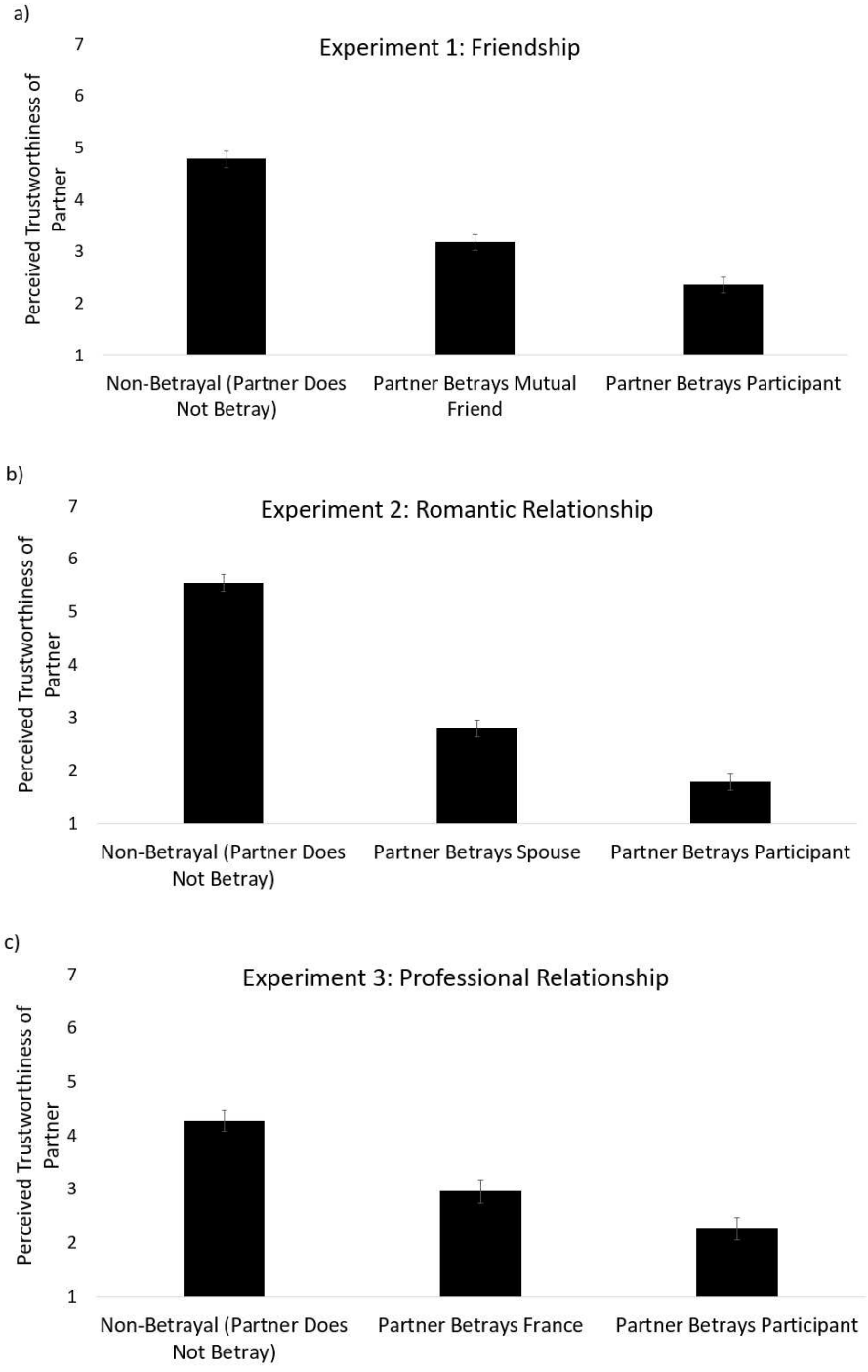
Inferential Statistics for Tests of Target (i.e., Partner) Trustworthiness in Experiments 1-3

	F	*p*	η_p^2	95% CI
Experiment 1				
Main effect of Partner Behavior	60.71	<.001	.347	--
Non-betrayal vs. Partner Betrays Mutual Friend	45.20	<.001	.235	1.13, 2.07
Non-betrayal vs. Partner Betrays Participant	113.56	<.001	.431	1.98, 2.88
Partner Betrays Mutual Friend vs. Partner Betrays Participant	17.92	<.001	.100	0.44, 1.21
Experiment 2				
Main effect of Partner Behavior	151.78	<.001	.575	--
Non-betrayal vs. Partner Betrays Spouse	160.74	<.001	.531	2.33, 3.19
Non-betrayal vs. Partner Betrays Participant	182.05	<.001	.657	3.22, 4.32
Partner Betrays Spouse vs. Partner Betrays Participant	8.25	.005	.087	0.31, 1.71
Experiment 3				
Main effect of Partner Behavior	25.54	<.001	.223	--
Non-betrayal vs. Partner Betrays France	18.83	<.001	.135	0.71, 1.91

Non-betrayal vs. Partner Betrays Participant	44.40	<.001	.261	1.41, 2.61
Partner Betrays France vs. Partner Betrays Participant	5.70	.019	.051	0.12, 1.28

Figure 2

Perceptions of Target (i.e., Partner) Trustworthiness Across Conditions in Experiments 1-3



Note. Error bars represent *SE*s.

Experiments 4-6

Targets (a) either betray someone or not; if they betray, the betrayal (b) either affects the participant (i.e., benefits them) or has no effect on them (because participants are merely third-party observers). We predict that participants will deem targets more trustworthy when targets (1) eschew betrayal versus betray, and (2) when participants benefit from versus are unaffected by the betrayal.

Method

Participants

We aimed to recruit 341 U.S.-residing participants from CloudResearch per experiment for sufficient power to detect effects f=.18. Participants were excluded from analyses as above. See Table 4.

Table 4

Sample Characteristics for Experiments 4-6

	Recruited *N*	Final *N*	Sex	M_{age}	SD_{age}
Experiment 4	480	346	62.1% female	39.30	12.31
Experiment 5	480	366	62.6% female	39.04	12.64
Experiment 6	481	366	57.7% female	41.27	13.57

Design and Procedure

Participants were randomly assigned to read one of four vignettes, in which the target—a friend (Experiment 4), romantic partner (Experiment 5), or fellow diplomat (Experiment 6)—either betrays or does not betray another person, and the participant is either a first-party interacting with the target or a third-party observer unaffected by the (non-)betrayal.

In Experiment 4's first-person conditions, the target knows the (absent) mutual friend's secret and could betray it to participants. The target either withholds the secret from the participant (non-betrayal, first-person) or shares it with the participant (betrayal, first-person). The third-person conditions mirror the above, but with participants reading about the target betraying (or not) the mutual friend's secret to *another person*. For each experiment, the information that participants learned was held constant across conditions (e.g., participants learned the mutual friend's secret in every condition), regardless of whether the target withheld or shared the secret (e.g., the mutual friend's secret is conveyed via a text message to the target that participants see on the target's phone).

In Experiment 5's first-person conditions, the participant is in a romantic relationship with the target, who is either not cheating on anyone (non-betrayal, first-person) or cheating on their spouse with the participant (betrayal, first-person). In Experiment 6's first-person conditions, the target is a high-level French government official who has sensitive French intelligence and could betray it to the participant. The target chooses either to protect the intelligence (non-betrayal, first-person) or to leak it to the participant (betrayal, first-person). The third-person conditions mirrored the first-person conditions (replacing participants with a same-sex other person). See Supplementary Materials for vignettes.

We assessed target trustworthiness as in Experiments 1-3 (αs=.95-.98).[2]

[2] We also assessed perceptions of the targets' relative valuation of their partners, generally finding that participants believe non-betrayers value those they refuse to betray over those they could have betrayed *to*, while betrayers value those *to whom* they betray more than those they

Results

Do People Deem Non-Betrayers More Trustworthy Than Betrayers?

Yes. Separate 2(Target behavior: betrayal vs. non-betrayal) X 2(Participant role: affected first-person vs. unaffected third-person) ANOVAs found significant main effects of target behavior on trustworthiness across experiments (ps<.001, η_p^2s≥.115). Participants deemed targets more trustworthy when targets did not betray others. See Tables 5-6 and Figure 3.

Do People Deem Targets Whose Betrayal Benefits Them More Trustworthy?

In two of three cases, yes. In Experiments 4-5, participants dealing with targets who betrayed friends or spouses *to oneself* deemed targets more trustworthy than those reading about a similar betrayal as an unaffected observer (e.g., the target shared a mutual friend's secret with another friend, but participants were not involved with the target or mutual friend). See Tables 5-6, Figure 3.

Notably, participants' role also impacted target trustworthiness when targets *did not* betray: targets were deemed more trustworthy by participants cast as targets' friends or partners versus observers. However, the overall pattern of results also suggests that the effect of targets' betrayal on the actor (i.e., actor benefits vs. not) plays a role in judgments of target trustworthiness (Experiment 4: η_p^2's ≥.013; Experiment 5: η_p^2's ≥.011). See Table 6.

In Experiment 6, participants' role did not significantly impact target trustworthiness. Both when the target betrayed their native government and did not, participants perceived target

---

betrayed (Supplementary Materials). Other exploratory measures not analyzed are available on OSF.

trustworthiness similarly, and regardless of whether participants were cast as colleagues or as unaffected observers ($ps \geq .055$).

Table 5

Means (SEs) for Target's Trustworthiness Across Conditions in Experiments 4-6

	Non-betrayal	**Betrayal**
Experiment 4	***M*(*SE*)**	***M*(*SE*)**
First-Person (Participant is Affected by Target Behavior)	4.89(0.14)	3.96(0.14)
Third-Person (Participant is Unaffected by Target's Behavior)	4.45(0.15)	3.43(0.15)
Experiment 5		
First-Person (Participant is Affected by Target Behavior)	5.56(0.13)	2.54(0.13)
Third-Person (Participant is Unaffected by Target Behavior)	5.20(0.12)	2.10(0.12)
Experiment 6		
First-Person (Participant is Affected by Target Behavior)	4.11(0.17)	2.99(0.16)
Third-Person (Participant is Unaffected by Target Behavior)	4.56(0.16)	2.72(0.16)

Table 6

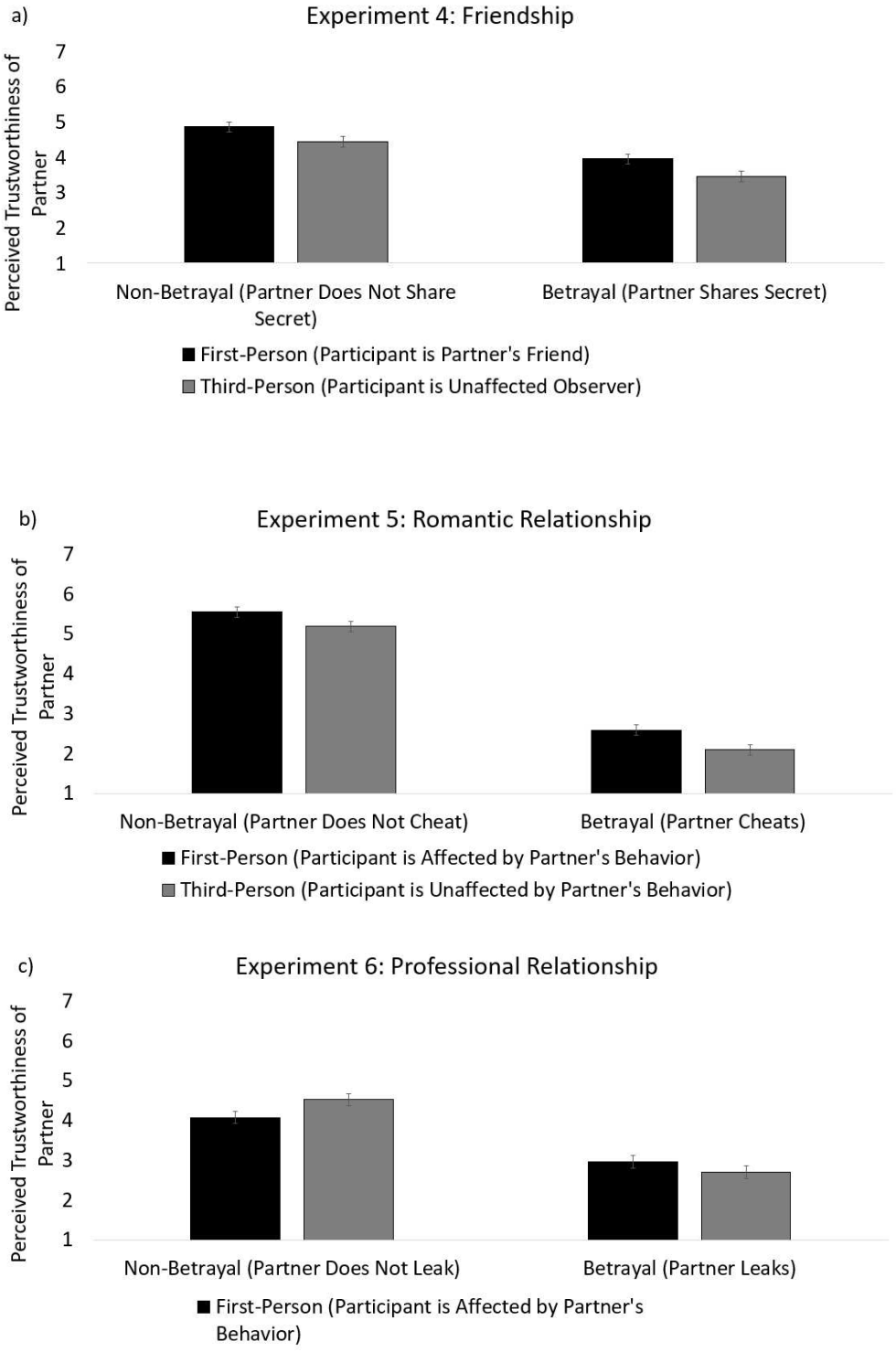
Inferential Statistics for Tests of Target's Trustworthiness Experiments 4-6

	F	*p*	η_p^2	95% CI
Experiment 4				
Main effect of Target Behavior	44.42	<.001	.115	--
Main effect of Participant Role	11.06	<.001	.031	--
Interaction of Target Behavior X Participant Role	0.09	.770	.000	--
Non-betrayal, First-Person vs. Betrayal, First-Person	21.43	<.001	.059	0.54, 1.33
Non-betrayal, Third-Person vs. Betrayal, Third-Person	22.99	<.001	.063	0.60, 1.43
Betrayal, First-Person vs. Betrayal, Third-Person	6.68	.010	.019	0.13, 0.93
Non-betrayal, First-Person vs. Non-betrayal, Third-Person	4.51	.034	.013	0.03, 0.86
Experiment 5				
Main effect of Target Behavior	584.79	<.001	.618	--
Main effect of Participant Role	10.04	<.001	.027	--
Interaction of Target Behavior X Participant Role	0.11	.745	.000	--
Non-betrayal, First-Person vs. Betrayal, First-Person	269.85	<.001	.427	2.66, 3.39

Non-betrayal, Third-Person vs. Betrayal, Third-Person	317.65	<.001	.467	2.77, 3.45
Betrayal, First-Person vs. Betrayal, Third-Person	6.17	.013	.017	0.09, 0.79
Non-betrayal, First-Person vs. Non-betrayal, Third-Person	3.98	.046	.011	0.01, 0.72
Experiment 6				
Main effect of Target Behavior	83.96	<.001	.188	--
Main effect of Participant Role	0.30	.583	.001	--
Interaction of Target Behavior X Participant Role	4.95	.027	.013	--
Non-betrayal, First-Person vs. Betrayal, First-Person	54.92	<.001	.060	0.66, 1.58
Non-betrayal, Third-Person vs. Betrayal, Third-Person	67.44	<.001	.157	1.40, 2.28
Betrayal, First-Person vs. Betrayal, Third-Person	1.46	.228	.004	-0.17, 0.71
Non-betrayal, First-Person vs. Non-betrayal, Third-Person	3.70	.055	.010	-0.90, 0.10

Figure 3

Perceptions of Target (i.e., Partner) Trustworthiness Across Conditions in Experiments 4-6



Note. Error bars represent *SE*s.

Experiment 7

Experiments 4-6 contain a potential confound: Participants in first-person conditions read about an existing close relationship with the target, while those in the third-person conditions read about the target as a non-affiliate. Because people likely trust their close partners more than non-affiliates, this difference across first- and third-person conditions—rather than the benefit to self of a target's betrayal—may have led participants to deem the target as more trustworthy in the first- versus the third-person conditions (Consistent with this possibility, in Experiments 4-5, participants deemed even non-betraying targets as more trustworthy in the first- versus the third-person conditions). To the extent that close partners are more likely to behave in ways that benefit oneself rather than harm oneself, this pattern of results is not discordant with the idea that judgments of target trustworthiness integrate cues of the expected impact on the self of trusting that target. However, the inclusion of this confounding variable prevents clear interpretation of how the mind integrates different cues of how a target's behavior will likely impact the self (i.e., cues to fitness interdependence via close relationships vs cues from the target's present behavior).

Experiment 7 thus replicates Experiment 4[3], eliminating this confound by describing the target and other characters in the vignettes as the participant's friends in both first-and third-person conditions. For example, in the first-person vignettes, the participant was friends with Abby and Bernadette, with Abby betraying Bernadette's secret to the participant (or not), and in

[3] Experiment 4 was replicated because the existence of close relationships between the participant and each of the other characters in the scenario was only realistic in the friendship context (versus Experiment 5's romantic context).

the third-person vignettes, the participant was friends with Abby, Bernadette, and Claire, with Abby betraying Bernadette's secret to Claire (or not) (see Supplementary Materials for vignettes). Other than this modification, Experiment 7's methods and measures were identical to those used in Experiment 4.

See Table 7 for sample details.

Table 7

Sample Characteristics for Experiments 7

Recruited *N*	Final *N*	Sex	M_{age}	SD_{age}
516	383	51.7% male	41.60	12.82

Results

Do People Deem Non-Betrayers More Trustworthy Than Betrayers?

Yes. A 2(Target behavior: betrayal vs. non-betrayal) X 2(Participant role: affected first-person vs. unaffected third-person) ANOVA revealed a main effect of Target behavior; Participants deemed targets who betrayed a friend's secret as less trustworthy than targets who kept a friend's secret. See Tables 8-9, Figure 4.

Do People Deem Targets Whose Betrayal Benefits Them More Trustworthy?

Yes. A main effect of Participant role indicated that participants deemed partners who betrayed a friend's secret *to oneself* (benefitting oneself) as more trustworthy than partners who betrayed a friend's secret *to another friend* (not impacting oneself). Importantly, a significant interaction also emerged, indicating that participants' role in the scenario did not impact perceptions of non-betraying partners' trustworthiness. This suggests that the difference observed in perceptions of the target's trustworthiness between the first- and third-person non-

betrayal conditions in Experiments 4-5 was driven by the existence of a close relationship between the participant and the target in the first-person conditions (and the lack of such a relationship in the third-person conditions). When this additional variable was removed (i.e., when participants have a close relationship with the target across all conditions), the likely impact of betrayer-targets' behavior on participants still shaped participants' perceptions of that target's trustworthiness.

Table 8

Means (SEs) for Target Trustworthiness in Experiment 7

	Non-betrayal	**Betrayal**
	***M*(*SE*)**	***M*(*SE*)**
First-Person (Participant is Affected by Target's Behavior)	4.56(0.14)	4.06(0.14)
Third-Person (Participant is Unaffected by Target's Behavior)	4.56(0.14)	3.00(0.14)

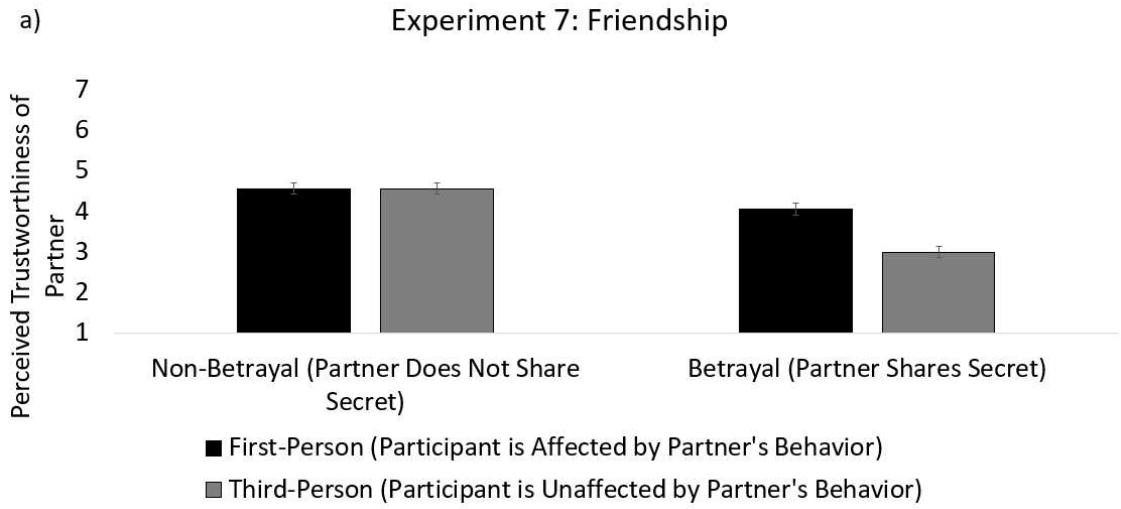
Table 9

Inferential Statistics for Tests of Target (i.e., Partner) Trustworthiness Across Conditions in Experiment 7

	F	***p***	η_p^2	**95% CI**
Main effect of Target Behavior	55.11	<.001	.127	--
Main effect of Participant Role	14.96	.000	.038	--
Interaction of Target Behavior X Participant Role	14.63	.000	.037	--
Non-betrayal, First-Person vs. Betrayal, First-Person	6.49	.011	.017	0.11, 0.89
Non-betrayal, Third-Person vs. Betrayal, Third-Person	63.15	<.001	0.14	1.18, 1.95
Betrayal, First-Person vs. Betrayal, Third-Person	28.91	<.001	.071	0.68, 1.46
Non-betrayal, First-Person vs. Non-betrayal, Third-Person	.001	.975	.000	-0.38, 0.39

Figure 4

Perceptions of Target (i.e., Partner) Trustworthiness in Experiment 7



Discussion

Judgments of target trustworthiness seem driven by whether that target is known to betray—as both intuition and some past work would predict—but these judgments are additionally driven by another, potentially less obvious factor: how target betrayal impacts the actor. Across experiments, people deem non-betrayers more trustworthy than betrayers. This suggests that trustworthiness judgments are regulated, in part, by targets' reluctance to betray, where reluctance to betray people in general may be a potent cue of reluctance to betray *oneself*—the evolutionarily relevant variable. But not all betrayers are deemed equally (un)trustworthy. People deem betrayers more trustworthy when those betrayers' behavior benefits oneself (e.g., by providing useful information) versus costs oneself (Experiments 1-3) or leaves oneself unaffected (Experiments 4-5).[4] This pattern was largely consistent across contexts—friendships, romantic relationships, and the professional domain. Effect sizes indicate that the TRUSTWORTHINESS concept depends primarily on whether a target betrays trust and, secondarily, when the target does betray, on the effect of the target's betrayal on the actor. This aligns with adaptationist approaches to personality, which view concepts relevant to person perception as having adaptive functions: predicting others' behavior based on past actions and communicating strategic information about others' behaviors (Buss 1996, 2011; Lukaszewski et al., 2020). From this perspective, the personality descriptor "(un)trustworthy" helps the actor

[4] The decrease in target trustworthiness when the target betrays the actor appears to be greater in magnitude than the increase in a target's judged trustworthiness when the target betrays in favor of the actor, an instance of bad being stronger than good (Baumeister et al., 2001).

determine whom to trust and to what extent, as well as to communicate this judgment to others, including close associates.

Limitations and Future Directions

Cultural and local norms influence decisions to extend trust (e.g., Yamagishi & Yamagishi, 1994) in ways that might affect how much people weigh cues of general trustworthiness (e.g., past betrayal) versus impact on oneself. For example, compared to Americans, Japanese seem reluctant to trust strangers, preferring to place trust in others with whom they have enduring relationships (Yamagishi, 2011). This suggests that Japanese might be especially attuned to individuals' target-specific (i.e., relational) trustworthiness and place less weight on an individual's general trustworthiness when making trust inferences. Whether the present findings generalize to other populations remains to be determined.

Judgments of hypothetical targets' trustworthiness might differ from trustworthiness judgments of real-world partners with whom one has had repeated interactions. Future work could address this via recall studies examining past behavior with real partners (Pedersen et al., 2020) or longitudinal studies tracking inferences over time (e.g., before and after friends share others' secrets). In-lab economic games could also increase the stakes of making (in)correct trustworthiness inferences and allow researchers to capture participants' trust *behavior* toward real-world partners (e.g., strangers, friends).

The claim that estimates of others' trustworthiness are ecologically and adaptively rational should be regarded as a data-informed conjecture. The veracity of the conjecture depends on whether, actuarially and ancestrally, the payoffs derived by trustors from their trustworthiness estimates would have qualitatively aligned with our findings. Thus, further

research could profitably study trustee behavior over time and payoffs to trustors in naturalistic interactions.

Conclusion

Trust is crucial for human social interaction, from close relationships to large-scale cooperation. The mind infers a target's trustworthiness by accounting for more than the target's willingness to betray. The mind also accounts for whether targets betray in ways that benefit or cost the actor. Findings may help explain the incongruence between, for example, wide endorsement of popular notions such as "once a cheater, always a cheater" and one's feelings of trust and liking in friends who have just shared another friend's secret.

References

Basyouni, R., & Parkinson, C. (2022). Mapping the social landscape: Tracking patterns of interpersonal relationships. *Trends in Cognitive Sciences*, *26*(3), 204-221.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5(4)*, 323–370.

Bedrov, A., Gable, S., & Liberman, Z. (2021). It takes two (or more): The social nature of secrets. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(6), e1576.

Brewer, M. B. (1997). On the social origins of human nature. In C. McGarty & S. A. Haslam (Eds.), *The message of social psychology: Perspectives on mind in society* (pp. 54 – 62). Cambridge, MA: Blackwell.

Buss, D. M. (1996). Social adaptation and five major factors of personality. In J. S. Wiggins (Ed.), The five factor model of personality: Theoretical perspectives (pp. 180–207). New York: Guilford Press.

Buss, D. M. (2011). Personality and the adaptive landscape: The role of individual differences in creating and solving social adaptive problems. In D. M. Buss & P. Hawley (Eds.), The evolution of personality and individual differences (pp. 29 –57). New York, NY: Oxford University Press.

Cosmides, L., & Tooby, J. (2015). Neurocognitive adaptations designed for social exchange. In David M. Buss (Ed.) *The handbook of evolutionary psychology*, 584-627.

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology*, *92*(2), 208.

Dasgupta, P. (1988). Trust as a commodity. In D. Gambetta (Ed.)*, Trust: Making and breaking cooperative relations* (pp. 49-72). Oxford: Basil Blackwell.

Delton, A. W., & Sell, A. (2014). The co-evolution of concepts and motivation. *Current Directions in Psychological Science, 23(2)*, 115–120.

DeScioli, P., & Kurzban, R. (2009). The alliance hypothesis for human friendship. *PloS ONE*, *4*(6), e5802.

Fonseca, M. A., & Peters, K. (2018). Will any gossip do? Gossip does not need to be perfectly accurate to promote trust. *Games and Economic Behavior*, *107*, 253-281.

Hess, N. C., & Hagen, E. H. (2002). Informational warfare.

Hess, N. H., & Hagen, E. H. (2023). The impact of gossip, reputation, and context on resource transfers among Aka hunter-gatherers, Ngandu horticulturalists, and MTurkers. *Evolution and Human Behavior*, *44*(5), 442-453.

Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are punishment and reputation for?. *PLoS ONE 7*(9): e45662.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science, 27(3)*, 405–418.

Krems, J. A., Hahnel-Peeters, R. K., Merrie, L. A., Williams, K. E., & Sznycer, D. (2023). Sometimes we want vicious friends: People have nuanced preferences for how they want their friends to behave toward them versus others. *Evolution and Human Behavior*, *44*(2), 88-98.

Krems, J. A., Merrie, L. A., Rodriguez, N. N., & Williams, K. E. (2024). Venting makes people prefer—and preferentially support—us over those we vent about. *Evolution and Human Behavior*, *45*(5), 106608.

Lukaszewski, A. W., Lewis, D. M. G., Durkee, P. K., Sell, A. N., Sznycer, D., & Buss, D. M. (2020). An adaptationist framework for personality science. *European Journal of Personality, 34*, 1151–1174.

Lukaszewski, A. W., & Roney, J. R. (2010). Kind toward whom? Mate preferences for personality traits are target specific. *Evolution and Human Behavior*, *31*(1), 29-38.

Merrie, L. A., Krems, J. A., & Sznycer, D. (2024). Dyads in networks: We (dis) like our partners' partners based on their anticipated indirect effects on us. *Evolution and Human Behavior*, *45*(2), 203-213.

Pedersen, E. J., McAuliffe, W. H., Shah, Y., Tanaka, H., Ohtsubo, Y., & McCullough, M. E. (2020). When and why do third parties punish outside of the lab? A cross-cultural recall study. *Social Psychological and Personality Science*, *11*(6), 846-853.

Peters, K., Jetten, J., Radova, D., & Austin, K. (2017). Gossiping about deviance: Evidence that deviance spurs the gossip that builds bonds. *Psychological Science*, *28*(11), 1610-1619.

Roberts, G. (2020). Honest signaling of cooperative intentions. *Behavioral Ecology*, *31*(4), 922-932.

Shaw, A., DeScioli, P., Barakzai, A., & Kurzban, R. (2017). Whoever is not with me is against me: The costs of neutrality among friends. *Journal of Experimental Social Psychology*, *71*, 96-104.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In *The adapted mind: Evolutionary psychology and the generation of culture* (Eds.: J. Barkow, L. Cosmides, & J. Tooby) (pp. 19–136). Oxford University Press.

Yamagishi, T. (2011). Trust as social intelligence. In *Trust* (pp. 107-131). Springer, Tokyo.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129-166.