Lumière Lyon 2 University

INTERIM THESIS, SECOND YEAR OF MASTER'S DEGREE IN HUMANITIES AND SOCIAL SCIENCES

MENTION: Fundamental and Applied Cognitive Sciences

Deconvolving trust and risk: a computational approach

Himalaya GIRARD

Produced under the direction of:

Elijah GALVAN, Cătălina RĂȚALĂ and Alan SANFEY

Laboratory address

Decision Neurosciences Laboratory at Donders Center for Cognitive Neuroimaging

Kapittelweg 29 Nijmegen,

GE 6525 EN The Netherlands

June 2025

"The fact that investments and returns are positive in most experiments is a puzzle from a game-theoretic viewpoint."

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142.

ABSTRACT

Trust behavior systematically violates game-theoretic predictions, as demonstrated by decades of Trust Game experiments where participants transfer substantial resources despite rational expectations of zero reciprocation. This paradox deepens when examining trust-risk relationships: some studies report strong correlations between risk preferences and trust behavior, while others find complete independence. This thesis argues that this empirical puzzle stems from a fundamental methodological confound: comparing trust decisions involving unknown probabilities (second-order risk) against lottery choices with explicit probabilities (first-order risk) may systematically bias results depending on participants' subjective belief distributions.

This investigation develops a computational framework exploring whether Prospect Theory's parametric architecture can accommodate trust decisions when uncertainty structures are properly controlled. Using simulated data from a novel Inverted Trust Game paradigm that implements first-order risk across both social and non-social contexts, the analysis examines 32 hierarchical models incorporating standard risk parameters alongside hypothesized trust-specific mechanisms. These mechanisms include betrayal aversion (additional disutility from human-caused losses) and social preferences (utility derived from others' outcomes). The simulations test whether extreme parameter recalibration within Prospect Theory's existing framework might explain trust behavior, or whether qualitatively distinct psychological processes are required. Moreover, the computational approach serves primarily to validate the experimental design's capacity to distinguish between these theoretical alternatives. While the simulations successfully recover all hypothesized parameters given the information-theoretic trial selection, the psychological plausibility of the selected parameter values awaits empirical validation.

The investigation's primary contribution lies in transforming the binary question of trust-risk equivalence into systematic specification of computational differences. Nevertheless, these remain open empirical questions. The simulated results demonstrate that the proposed paradigm possesses sufficient statistical power to detect theoretically meaningful parameter differences, should they exist in human participants. This methodological foundation enables future empirical work to resolve whether trust and risk share common psychological architecture. Indeed, if trust merely recalibrates existing decision parameters, this would suggest targeted interventions; if trust requires additional psychological machinery, this would necessitate expanded theoretical frameworks.

KEYWORDS

Trust behavior • Risk preferences • Prospect Theory • Computational modeling • Order-of-risk confound • Trust Game • Inverted Trust Game • First-order risk • Second-order risk • Betrayal aversion • Social preferences • Probability weighting • Value function • Social uncertainty • Parametric decomposition • Hierarchical Bayesian estimation • Social Value Orientation • Strategic uncertainty • Decision-making under uncertainty • Behavioral economics

GLOSSARY

Ambiguity

A form of uncertainty where neither the potential outcomes nor their associated probabilities are known. Distinguished from risk where at least outcomes are specified.

Betrayal aversion

The additional psychological disutility experienced when losses result from intentional human decisions rather than random processes. Quantified as parameter τ in the extended Prospect Theory model.

Cumulative Prospect Theory (CPT)

Extension of Prospect Theory (Tversky & Kahneman, 1992) that applies probability weighting to cumulative distributions rather than individual probabilities, enabling application to prospects with multiple outcomes.

Diminishing sensitivity

The psychological principle whereby the subjective impact of a change decreases as one moves further from a reference point. Captured by power function exponents α and $\beta < 1$.

First-order risk

Decision situations where both potential outcomes and their objective probabilities are explicitly known (e.g., lottery with stated 60% chance of winning).

Fisher information

Statistical measure quantifying the amount of information that observable data carries about unknown parameters. Used for optimal experimental design through D-optimization.

Hierarchical Bayesian estimation

Statistical method that simultaneously estimates individual-level parameters and population distributions, accounting for both within-subject and between-subject variation.

Inverted Trust Game

Novel experimental paradigm where trustors make investment decisions with explicit probability information about trustees' past behavior, converting trust from second-order to first-order risk.

Loss aversion

The tendency for losses to have greater psychological impact than equivalent gains. Quantified by parameter $\lambda > 1$ in Prospect Theory.

Order-of-risk confound

Methodological issue arising from comparing decisions under first-order risk (known probabilities) with second-order risk (unknown probabilities), potentially obscuring true domain differences.

Probability weighting function

Nonlinear transformation w(p) that converts objective probabilities into subjective decision weights. Typically inverse S-shaped with parameters γ (gains) and δ (losses).

Prospect Theory

Descriptive theory of decision-making under risk (Kahneman & Tversky, 1979) that explains systematic deviations from expected utility through reference dependence, loss aversion, and probability weighting.

Second-order risk

Decision situations where potential outcomes are known but their probabilities must be subjectively estimated (e.g., trust decisions without explicit probability information).

Social preferences

Concern for others' outcomes in addition to one's own. Quantified by parameter φ, where $\varphi > 0$ indicates prosocial preferences and $\varphi < 0$ indicates competitive preferences.

Social Value Orientation (SVO)

Individual differences in preferences for resource allocation between self and others, measured as an angular degree from -45° (competitive) to 90° (altruistic).

Trust Game

Experimental paradigm (Berg et al., 1995) where one participant (trustor) transfers money to another (trustee), the amount is multiplied, and the trustee decides how much to return.

Utility function

Mathematical representation of preferences over outcomes. In Expected Utility Theory, typically concave for risk aversion. In Prospect Theory, replaced by the value function.

Value function

In Prospect Theory, the function v(x) that transforms objective outcomes into subjective values, characterized by reference dependence, diminishing sensitivity, and loss aversion.

List of Abbreviations

AIC — Akaike Information Criterion

ANOVA — Analysis of Variance

BIC — Bayesian Information Criterion

BOLD — Blood Oxygen Level Dependent

CBS — Centraal Bureau voor de Statistiek (Netherlands Statistics Office)

CELEX — Center for Lexical Information

CI — Confidence Interval

CMO — Commissie Mensgebonden Onderzoek (Research Ethics Committee)

CPT — Cumulative Prospect Theory

fMRI — Functional Magnetic Resonance Imaging

HPC — High Performance Computing

LCD — Liquid Crystal Display

LL — Log-Likelihood

MAP — Minimum Acceptable Probability

MVN — Multivariate Normal Distribution

OLD20 — Orthographic Levenshtein Distance 20

PT — Prospect Theory

RLOT — Risky Lottery

RTG — Risky Trust Game

SD — Standard Deviation

SE — Standard Error

SOEP — Socio-Economic Panel

STG — Standard Trust Game

SVO — Social Value Orientation

TG — Trust Game

VMPFC — Ventromedial Prefrontal Cortex

WMA — World Medical Association

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisors, PhD candidate Elijah Galvan and Dr. Cătălina Rățală, for their exceptional guidance throughout this research project. Their trust in me to pursue this ambitious investigation, combined with their constant availability for advice and feedback during our many project meetings, created an ideal learning environment. Their complementary expertise in computational methods and theoretical frameworks enriched every aspect of this work.

I also thank Professor Alan Sanfey, principal investigator of the Decision Neurosciences Laboratory, for providing both the research infrastructure and intellectual environment that made this project possible. His vision for integrating neurocomputational approaches with psychological theory shaped the direction of this investigation.

Finally, the Donders Institute for Brain, Cognition and Behaviour deserves special recognition for providing outstanding resources, particularly the high-performance computing facilities essential for our simulations. The stimulating academic environment and opportunities for interdisciplinary exchange significantly contributed to the development of this work

NOTICE TO READERS

This thesis presents a computational investigation based entirely on simulated data. All behavioral results, statistical analyses, and reported findings derive from computer-generated agents rather than human participants. The 100 “participants” referenced throughout represent simulated agents whose behavior was generated using theoretically motivated parameter values within a Prospect Theory framework.

The Inverted Trust Game paradigm described herein has been designed but not yet implemented with human subjects. Consequently, all patterns and relationships reported should be interpreted as theoretical predictions awaiting empirical validation. This work establishes a methodological foundation for future experimental research investigating trust and risk decision-making.

TABLE OF CONTENTS

ABSTRACT **4**

KEYWORDS **5**

GLOSSARY **6**

LIST OF ABBREVIATIONS **9**

ACKNOWLEDGMENTS **10**

NOTICE TO READERS **11**

1. THEORETICAL AND EMPIRICAL CONTEXT **14**

1.1 Introduction 14

1.2 Conceptualizing Trust As Social Uncertainty 15

1.2.1 Taxonomic Framework *15*

1.2.2 Theoretical Predictions *16*

1.3 The Trust-Risk Empirical Puzzle 16

1.3.1 Evidence Against Risk-Trust Correlation *16*

1.3.2 Evidence Supporting Risk-Trust Correlation *17*

1.3.3 The Order-of-Risk Confound: A Methodological Barrier *17*

1.3.4 Addressing The Order Of Risk Confound *18*

1.4 Multiple Mechanisms Underlying Trust Behavior 19

1.4.1 Social Preferences and Trust Behavior *19*

1.4.2 Betrayal Aversion *20*

1.5 The Prospect Theory Opportunity: Decomposing Decision 20

1.5.1 Core Parameters And Functions *20*

1.5.1.1 Value Function Parameters (A, B, Λ) *20*

1.5.1.2 Probability Weighting Parameters (Γ, Δ) *21*

1.5.2 Neural Basis of Prospect Theory Parameters *21*

1.5.3 Prior Application Of Prospect Theory To Trust *22*

1.6 Synthesis And Current Study Objectives 23

2. EXPERIMENT **24**

2.1 Method 24

2.1.1 Participants *24*

2.1.2 Stimuli *24*

2.1.2.1 Inverted Trust Game *25*

2.1.2.2 Social values measurement *26*

2.1.2.3 Reciprocity Decisions *27*

2.1.3 Computational Framework *27*

2.1.3.1 Trust-Specific Prospect Theory Extensions *27*

Trust-Specific Value Function *27*

Trust-Specific Probability Weighting *28*

Social Preferences Parameter (Φ) *28*

Prospect Evaluation for Binary Choices *28*

Stochastic Choice Function *29*

Parameter Estimation *29*
2.1.3.2 Trial Set Construction *30*
Trial Generation *30*
Parameter Space Definition *30*
Information-Theoretic Selection *30*
2.1.4 Procedure *30*
2.1.4.1 Apparatus *30*
2.1.4.2 Cover Story and Experimental Framing *31*
2.1.4.3 Task Administration *32*
2.1.4.4 Payment Determination *32*
2.1.4.5 Debriefing *32*
2.2 Operational hypotheses 32
2.3 Expected Results 33
2.3.1 Behavioral Patterns *34*
2.3.2 Model Comparison *35*
2.3.3 Parameter Estimates *36*
2.3.4 Model Validation *37*
3. DISCUSSION 37
3.1 Betrayal Aversion: The Social Cost Of Trust 38
3.2 Social Preferences And Individual Differences 39
3.3 Enhanced Value Sensitivity In Social Contexts 39
3.4 The Probability Weighting Transformation 40
3.5 Behavioral Consequences: Three Zones And Parameter Interactions 41
3.6 Reconciling The Trust-Risk Literature 42
3.7 Theoretical Contributions, Limitations, and Future Directions 42
REFERENCES 44
ANNEX 49

1. THEORETICAL AND EMPIRICAL CONTEXT

1.1 INTRODUCTION

Trust constitutes a fundamental economic mechanism. Knack and Keefer (1997) showed that a 10% increase in societal trust correlates with 0.5% higher annual economic growth, while Zak and Knack (2001) identified the causal pathway through investment rate depression in low-trust environments. Brown et al. (2015) showed trust directly enhances financial performance by reducing monitoring costs. Despite this economic importance, trust violates standard economic predictions.

The Trust Game paradigm (TG; Berg et al., 1995) exemplifies this contradiction: one participant (trustor) transfers money to another (trustee), the amount undergoes multiplication, and the trustee decides what proportion to return. Game-theoretic analysis predicts zero transfers: trustees should retain everything, and trustors, anticipating this, should send nothing. Yet meta-analytic evidence spanning 162 replications reveals that trustors transfer approximately half their endowments while trustees reciprocate one-third (Johnson & Mislin, 2011). This empirical violation of rational choice predictions demands explanation. Moreover, the systematic nature of these violations, namely that they are consistent across cultures, stakes, and experimental variations, suggests that TG behavior reflects fundamental psychological processes rather than mere experimental artifacts. The robustness of this finding has transformed trust from an economic anomaly into a cornerstone phenomenon for understanding human cooperation.

This empirical violation parallels established findings in risk research, where decisions systematically deviate from expected utility predictions. While risk decision-making possesses theoretical frameworks explaining such deviations through psychological parameters, trust mechanisms lack comparable characterization. This asymmetry presents methodological opportunity. Indeed, if trust employs risk-assessment processes with domain-specific calibrations, risk frameworks could accelerate trust research. Conversely, fundamental architectural differences would necessitate independent theoretical development. The critical question becomes whether trust represents a special case of risk evaluation applied to social contexts, or whether it engages qualitatively distinct cognitive processes that standard risk models cannot capture.

We argue that Prospect Theory's parametric decomposition offers a lens for examining this question. Prospect Theory explains risk behavior through three psychological transformations: nonlinear utility functions capturing diminishing sensitivity to outcomes, probability weighting functions reflecting systematic misperception of likelihoods, and loss aversion quantifying asymmetric valuation of gains versus losses (Kahneman & Tversky, 1979). These parameters have demonstrated predictive power across fundamental decision anomalies: preference reversals where

choice and pricing procedures yield systematically different preferences (Lichtenstein & Slovic, 1971; Grether & Plott, 1979), the certainty effect where sure outcomes are overweighted relative to probable ones (Kahneman & Tversky, 1979), and framing effects where identical outcomes described as gains versus losses elicit opposing choices (Tversky & Kahneman, 1981; Kühberger, 1998).Their application to trust would enable testing whether social contexts recalibrate theses existing mechanisms.

The theoretical development proceeds through four steps. First, a taxonomic framework distinguishes trust from risk based on uncertainty source. Second, empirical literature reveals contradictory findings regarding trust-risk correlations, necessitating methodological analysis to identify systematic factors explaining this inconsistency. Third, trust-specific mechanisms including social preferences and betrayal aversion illustrate domain-specific processes absent in non-social risk. Finally, Prospect Theory provides an analytical framework for decomposing these mechanisms into measurable parameters, enabling specification of trust-risk relationships.

1.2 Conceptualizing Trust As Social Uncertainty

1.2.1 Taxonomic Framework

Trust and risk represent distinct forms of decision-making under uncertainty, differentiated primarily by the source of uncertainty. Zand (1972) defined trust as "the willingness to increase one's vulnerability to another person whose behavior is not under one's control," encompassing three critical elements: vulnerability to potential loss, dependence on another agent, and absence of outcome control. Risk, conversely, refers in neuroeconomics to "variance over probability distribution of outcomes" (Glimcher, 2013), presupposing known probabilities and defined outcomes. The uncertainty taxonomy developed through contributions from Knight (1921), Keynes (1921), and Ellsberg (1961) establishes a hierarchical framework: *First-order risk* involves situations where both outcomes and associated probabilities are known (the domain of classical lotteries and gambling paradigms). *Second-order risk* encompasses scenarios where potential outcomes are known but probabilities remain unknown, requiring subjective probability assessment. *Ambiguity* represents the extreme case where neither outcomes nor probabilities are known. This hierarchy structures uncertainty from complete specification towards complete ignorance. Moreover, the distinction between social and non-social uncertainty creates orthogonal categorization. Trust involves uncertainty from intentional choices; risk involves random processes without agency. Indeed, this framework enables systematic investigation of whether identical psychological mechanisms process these distinct uncertainty types.

1.2.2 Theoretical Predictions

Classical economic theory conceptualizes trust as risk assessment applied to social contexts. Coleman (1990) formulates trust within rational choice, where actors calculate expected values through outcome utilities multiplied by subjective probabilities. Evans and Krueger (2014) extend this framework, proposing identical computational processes for social and non-social prospects. The same mechanisms drive both behaviors.

This generates a falsifiable prediction: individual risk preferences should significantly correlate with trust behavior. Risk-averse individuals should invest less in Trust Games than risk-seeking individuals. The correlation strength indicates whether trust employs general uncertainty processing or specialized social processes. High correlations support domain-general mechanisms; weak correlations suggest distinct psychological architectures. Nevertheless, empirical tests reveal contradictory findings that challenge this theoretical prediction.

1.3 The Trust-Risk Empirical Puzzle

1.3.1 Evidence Against Risk-Trust Correlation

Multiple investigations failed to detect significant correlations between risk preferences and trust behavior, contradicting theoretical predictions. Eckel and Wilson (2004) pioneered empirical trust-risk testing *(N* = 232) using three assessment methodologies: the Zuckerman Sensation-Seeking Scale[1] (r = .07, p = .44), Holt-Laury multiple price list [2](r = .06, p = .52), and trust-mimicking lottery [3](r = .02, p = .81). All yielded null results.

Houser et al. (2010) isolated risk assessment from prosocial motivations through four conditions (N = 291): two trust treatments with human trustees and two risk treatments using computerized decisions. While risk attitudes predicted non-social investment (p < .05), they showed no relationship with social decisions (p = .13; p = .82), despite controlling prosocial motivations through passive recipient conditions. Moreover, field evidence corroborates laboratory findings: Etang et al. (2011) examined rural Cameroonian participants (n = 140) using Schechter's (2007) Risk Aversion Game[4], finding no significant correlation (β = 0.09, t = 1.50, p > .05).

Additional null findings persist: Ashraf et al. (2006) found no trust relationship in cross-cultural samples (N = 359); Garapin et al. (2015) detected no predictive effect using within-subject design (N = 180); Kanagaretnam et al. (2006) reported non-significant risk attitudes (N = 182);

[1] Measures individual differences in the need for novel, intense experiences and willingness to take risks (Zuckerman, 1979).

[2] Elicits risk preferences through paired lottery choices with varying probabilities and payoffs (Holt & Laury, 2002)

[3] Replaces the trustee with a lottery of equivalent expected payoffs to isolate trust from risk preferences (Bohnet & Zeckhauser, 2004).

[4] Uses dice-betting choices between certain amounts and gambles to elicit risk preferences, designed for low-literacy populations (Schechter, 2007).

Ben-Ner and Halldorsson (2010) found no relationship using behavioral and survey measures ($N = 204$). This evidence suggests risk preferences inadequately explain trust behavior.

1.3.2 Evidence Supporting Risk-Trust Correlation

Several studies detected risk-trust correlations under specific conditions. Karlan (2005) found experimental trustor behavior predicted financial decisions among Peruvian microfinance participants ($N = 864$). Higher trust transfers associated with reduced savings ($\beta = -46.63$, $p < .01$) and increased dropout ($\beta = 0.15$, $p < .05$), suggesting trust transfers partially capture risk tolerance.

Schechter (2007) reported positive associations using dice-betting risk elicitation with rural Paraguayans ($N = 188$). Risk Aversion Game investments correlated with TG transfers ($\beta = 0.277$, $p < .01$). Risk preference controls attenuated gender differences, indicating differential risk attitudes contribute to sex effects. Psychometric approaches yield mixed support: Fehr (2009) found SOEP risk willingness[5] modestly predicted trust attitudes ($\beta = -0.16$, $p < .001$), while betrayal aversion[6] showed stronger associations ($\beta = -0.36$, $p < .001$).

Measurement characteristics prove influential. Chetty et al. (2020) employed extensive risk elicitation (40 lottery pairs) with South African participants ($N = 202$). Each risky choice corresponded to 1 Rand increased trust transfer ($p = .020$), with risk preferences explaining 20% of variance. Simulations indicated standard instruments (10-15 lottery pairs) would detect this correlation in only 24% of replications. Studies detecting correlations employed longitudinal validation, behavioral betting paradigms, psychometric scales, or comprehensive protocols. Nevertheless, methodologically similar studies yield contradictory results (Schechter, 2007; Etang et al., 2011), suggesting unmeasured moderators.

1.3.3 The Order-of-Risk Confound: A Methodological Barrier

Trust-risk research faces fundamental confounding from distinct uncertainty structures. Trust Games involve second-order risk (known outcomes, unknown probabilities contingent on strategic choices); risk tasks employ first-order risk with explicit probabilities. This structural discrepancy creates experimental confounding: observed differences may reflect domain-specific processing or differential uncertainty responses.

Krain et al. (2006) meta-analyzed 27 neuroimaging studies comparing non-social first-order risk (13 studies) versus second-order risk (14 studies). These structures engaged distinct neural networks: first-order risk activated orbitofrontal cortex and rostral anterior cingulate (emotional processing); second-order risk engaged dorsolateral prefrontal cortex and caudal anterior cingulate

[5] Survey question asking respondents to rate their general willingness to take risks on an 11-point scale, validated against incentivized experiments (Dohmen et al., 2011).

[6] Measures the additional reluctance to trust when facing potential intentional harm versus equivalent risk from nature, typically elicited by comparing TG decisions with lottery choices (Bohnet et al., 2008).

(cognitive control). This dissociation confirms discrete processing pathways. Indeed, expectation formation differs fundamentally. Second-order risk requires subjective probability formation rather than responding to objective distributions. Breuer et al. (2017) found higher trustworthiness expectations correlated with lower trust willingness ($\beta = 1.102$, $p < .05$), revealing complex psychological dynamics.

These distinctions reveal different cognitive processes between uncertainty types. Studies reporting no trust-risk relationship may reflect order-of-risk confounding rather than genuine independence.

1.3.4 Addressing The Order Of Risk Confound

Fairley et al. (2016) addressed the order-of-risk confound through the Risky Trust Game (RTG), establishing social first-order risk conditions. Participants completed both Standard Trust Game [7](STG; second-order risk) and RTG featuring five scenarios with 0-4 trustworthy trustees exhibiting binary profiles (0% or 100% reliable). Random trustee selection per trial transformed social interaction into probabilistic selection of deterministic choices, diminishing ecological validity by eliminating individual-level outcome uncertainty. Despite this limitation, RTG behavior significantly predicted STG trust ($r = .242$, $p < .05$, $N = 92$), while lottery-based risk measures showed no relationship, suggesting risk preferences fail to predict trust even controlling for uncertainty structure.

Fairley et al. (2019) implemented matched first-order risk using participants' natural beliefs. Their 2×2 design (uncertainty source: TG vs. lottery; uncertainty type: first vs. second order) calibrated probabilities to match elicited beliefs. Behavioral analysis revealed no RTG-RLOT[8] investment differences. However, neuroimaging revealed domain-specific processing: investment amounts modulated ventral striatum (-4, 7, -7; 6, 4, -7) during TG anticipation only. RTG-RLOT comparison showed differential bilateral ventral striatum-orbitofrontal activation.

Lauharatanahirun et al. (2012) implemented identical first-order risk across domains, framing pie charts as average trustee repayments. Contradicting Fairley et al. (2016), they found significant social-nonsocial correlation ($\rho = .60$, $p < .001$). Nevertheless, neuroimaging revealed dissociable computations: left amygdala exhibited significant three-way interaction (GROUP × CONDITION × CHOICE, $p < .05$, $z = 4.43$). This indicates parallel, but distinct mechanisms through differential emotion-processing recruitment despite behavioral convergence.

[7] Follows Berg et al. (1995) Trust Game

[8] Follows Ellsberg-type lottery where participants bet on marble colors with known probabilities of winning/losing outcomes)

1.4 Multiple Mechanisms Underlying Trust Behavior

Trust decisions incorporate psychological processes beyond risk assessment, including mechanisms absent in non-social uncertainty. Monetary losses from randomization processes may be interpreted as bad luck; identical losses from human decisions carry additional dimensions: they signal faulty social judgment (Trautmann et al., 2008) and involve intentionality evaluations (Falk & Fischbacher, 2006). Additional mechanisms include reputation processing (Ma et al., 2020), moral emotions (Kugler et al., 2020), and norm adherence (Schlösser et al., 2015). Two mechanisms show particularly robust empirical support.

1.4.1 Social Preferences and Trust Behavior

Social value orientation (SVO; preferences regarding resource allocation between self and others) suggest consistent but limited predictive capacity for trust behavior. Kanagaretnam et al. (2006) established SVO as a significant predictor (p = .002) using Liebrand's (1984) Ring Measure, documenting systematic progression: competitor/individualist (44.99%), individualist/cooperator (55.66%), and cooperator/altruist (68.39%) trust rates. This association persisted controlling for demographics, enhancing model fit from R^2 = .05 to R^2 = .167. Derks (2014, 2015) confirmed these findings in adolescent samples (β = -0.31, p < .001). Wei (2016) documented higher trust among prosocials (M = 0.62) than proselfs (M = 0.51, p < .001).Snijders (1996) and Lambert et al. (2017) further corroborated the relationship between prosocial orientation and trust propensity across different experimental paradigms.

Alternative operationalizations yield comparable results. Ashraf et al. (2006) measured altruistic tendencies through Dictator Game [9] transfers, finding unconditional kindness significantly predicted trust (γ = 0.379, p < .01). Altruism combined with reciprocity expectations explained 62% of trust variance, whereas risk preferences exhibited no effect. Cox (2004) established that trustors send significantly more in TG than Dictator Games. Moreover, this indicates trust incorporates both social preferences and strategic expectations.

Despite consistent directional findings, explanatory power shows limitations. Garapin et al. (2015) found social preferences predicted trust transfers (β = .028, p < .001) but explained minimal variance (R^2 ≈ .14-.18). Fairley (2016) reported no multivariate relationship when controlling other factors (β = -0.002, SE = 0.015), consistent with Houser's (2010) findings. Cross-study differential capacity indicates SVO captures relevant, but insufficient determinants.

[9] The dictator game is an experimental economics paradigm where one player unilaterally decides how to divide a sum of money between themselves and a passive recipient who must accept whatever is offered (Kahneman et al., 1986).

1.4.2 Betrayal Aversion

Trust introduces psychological dimensions beyond monetary loss through potential exploitation. Bohnet and Zeckhauser (2004) isolated "betrayal aversion", finding participants require higher minimum acceptable probabilities (MAP) for trust (MAP = .54) than equivalent lotteries (MAP = .32, $p < .01$). This quantifies psychological costs distinct from financial risk.

Aimone et al. (2014) localized betrayal aversion to right anterior insula activation [40/22/3] during potential betrayal, with activation magnitude correlating with behavioral measures. This neural signature contrasts with general loss aversion, which manifests through decreased ventral striatum and VMPFC activation rather than increased negative emotion processing (Tom et al., 2007). Indeed, the anterior insula's selective engagement establishes betrayal aversion as neurobiologically distinct from standard risk evaluation. Betrayal aversion operates concurrently with social preferences, illustrating trust engages multiple specialized mechanisms beyond risk assessment. The order-of-risk confound compounds this complexity by conflating domain differences with uncertainty structure. These methodological and theoretical challenges can be address with Prospect Theory's parametric framework, which decomposes decision processes into constituent parameters while standardizing uncertainty conditions.

1.5 The Prospect Theory Opportunity: Decomposing Decision

Developed to explain systematic violations of Expected Utility Theory, Prospect Theory (PT; Kahneman & Tversky, 1979) and Cumulative Prospect Theory (CPT; Tversky & Kahneman, 1992) offer a parametric framework for understanding decision-making under uncertainty. Unlike Expected Utility Theory's single utility curvature parameter, PT decomposes risk attitudes into distinct psychological mechanisms: probability weighting, outcome sensitivity, and loss aversion. This multi-parameter approach has successfully accounted for numerous empirical anomalies in risk contexts. The framework's application to trust decisions presents an opportunity to test whether social uncertainty merely recalibrates existing psychological parameters or engages fundamentally different cognitive architecture.

1.5.1 Core Parameters And Functions

1.5.1.1 Value Function Parameters (A, B, Λ)

At the heart of Prospect Theory lies the value function (Figure 1a), which transforms objective outcomes into subjective utilities relative to a reference point. For gains, the function takes the form $v^{+}(x) = x^{\alpha}$ where $x \geq 0$, while for losses, $v^{-}(y) = -\lambda \times y^{\beta}$ where $y < 0$. Parameters α and β (constrained between 0 and 1) capture the phenomenon of diminishing sensitivity: the psychological principle wherein each additional unit of gain or loss has less subjective impact as one moves further from the reference point. Lower parameter values indicate

more pronounced curvature (that is, stronger diminishing sensitivity; Figure 1a). An additional 100€ matters substantially more when it increases wealth from 200€ to 300€ than when it increases wealth from 2000€ to 2100€. Tversky and Kahneman (1992) empirically estimated these parameters at approximately $\alpha \approx \beta \approx 0.88$, suggesting that perceptual processing of gains and losses exhibits comparable sensitivity characteristics. The loss aversion coefficient λ quantifies a separate phenomenon: the asymmetric psychological impact of losses relative to gains. This parameter, typically estimated at $\lambda \approx 2.25$, formalizes the observation that "losses loom larger than gains" (Kahneman & Tversky, 1979, p. 279). A loss of 100€ generates approximately 2.25 times the psychological impact of a 100€ gain. Moreover, this asymmetry appears fundamental to human decision-making, persisting across diverse contexts and populations.

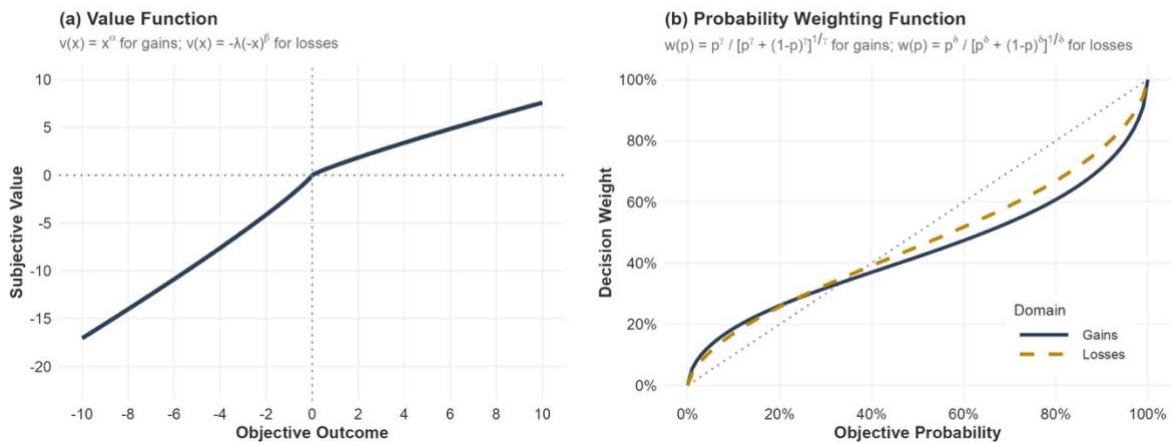
1.5.1.2 Probability Weighting Parameters (Γ, Δ)

Objective probabilities undergo nonlinear transformation into decision weights through separate weighting functions (Figure 1b). The general form $w^+(p) = \frac{p^\gamma}{(p^\gamma+(1-p)^\gamma)^{\frac{1}{\gamma}}}$ applies to gains, while $w^-(p) = \frac{p^\delta}{(p^\delta+(1-p)^\delta)^{\frac{1}{\delta}}}$ applies to losses. Parameters γ and δ determine the degree of probability distortion. Values below unity generate inverse S-shaped curves (Figure 1b). These curves systematically overweight small probabilities while underweighting moderate-to-high probabilities. This transformation explains numerous decision anomalies, such as simultaneous lottery playing (overweighting tiny chances of winning) and insurance purchasing (overweighting small disaster probabilities). Empirical estimates place $\gamma \approx 0.61$ and $\delta \approx 0.69$, indicating more pronounced distortion for gains than losses.

1.5.2 Neural Basis of Prospect Theory Parameters

Contemporary neuroscience has identified distinct neural substrates underlying each Prospect Theory parameter, validating the psychological decomposition through biological evidence. Value function curvature manifests in striatal and ventromedial prefrontal cortex activity

Figure 1. Prospect Theory Transformation Functions



Note: Standard Prospect Theory parameters for non-social risk contexts (Tversky & Kahneman, 1992).
(a) Value function with parameters $\alpha = \beta = 0.88$ (outcome sensitivity) and $\lambda = 2.25$ (loss aversion).
(b) Probability weighting with parameters $\gamma = 0.61$ (gains) and $\delta = 0.69$ (losses).
Dotted lines represent linear reference functions (45° line for probability weighting).

patterns, where BOLD signal responses scale with outcome magnitude but exhibit the diminishing sensitivity characteristic of the value function (Tom et al., 2007). Loss aversion emerges through asymmetric neural responses: striatal deactivation in response to losses approximately doubles the activation observed for equivalent gains. Pharmacological interventions provide causal evidence; β-adrenergic receptor antagonists selectively reduce loss sensitivity without affecting gain processing (Rogers et al., 2004). Indeed, these dissociations confirm that loss aversion represents a distinct psychological mechanism rather than merely stronger sensitivity to negative outcomes. Probability weighting engages separate neural circuits. Ventral anterior thalamus and specific cingulate regions encode probability information nonlinearly, with activation patterns mirroring the inverse-S shaped weighting function (Hsu et al., 2009). Recent single-unit recordings in primates show that individual neurons in orbitofrontal cortex and striatum compute subjective valuations by parametrically integrating both utility and probability weighting transformations (Imaizumi et al., 2022). These converging findings establish that Prospect Theory parameters reflect fundamental properties of neural valuation systems rather than mathematical conveniences.

1.5.3 Prior Application Of Prospect Theory To Trust

While Prospect Theory has revolutionized understanding of non-social risk, its application to trust decisions remains limited. Nguyen et al. (2016) conducted the only attempt, examining trust behavior in artifactual field experiments across northern and southern Vietnam. They estimated correlations between Trust Game transfers and four risk parameters: utility curvature (α), probability weighting for gains (γ) and losses (δ), and loss aversion (λ). The analysis also incorporated quasi-hyperbolic time preferences. Parameter-specific effects emerged that traditional Expected Utility frameworks could not detect. Loss aversion showed a marginal negative relationship with trust ($p = .082$) in models without expectation controls; however, this effect disappeared when expected returns were included. Regional heterogeneity manifested primarily through probability weighting parameters. Northern Vietnamese participants exhibited negative correlations between probability weighting and trust that were absent in southern samples. Nevertheless, two critical methodological limitations constrain these findings interpretability.

First, the comparison between non-social first-order risk (lotteries with explicit probabilities) and social second-order risk (Trust Games with unknown probabilities) confounds domain differences with uncertainty structure differences. Second, the analytical approach applied asymmetric decomposition: disaggregating non-social decisions into component psychological parameters while treating social decisions as unitary phenomena. This asymmetry precludes direct comparison of parameter values across domains.

1.6 Synthesis And Current Study Objectives

The preceding review reveals a fundamental puzzle: trust-risk correlations vary dramatically across methodologies. Standard protocols detect no relationship (Eckel & Wilson, 2004; Houser et al., 2010), while high-resolution assessments (Chetty et al., 2020) and gambling paradigms (Schechter, 2007) reveal significant associations. This methodological sensitivity suggests measurement instruments capture different psychological components. Moreover, the order-of-risk confound provides a parsimonious explanation. Trust Games involve second-order risk (unknown probabilities); risk tasks employ first-order risk (explicit probabilities). These structures engage distinct neural networks (Krain et al., 2006).

Recent attempts to control uncertainty structure yield intriguing patterns. Fairley et al. (2019) and Lauharatanahirun et al. (2012) implemented first-order risk in trust contexts, finding behavioral similarities but persistent neural dissociations. Previous Prospect Theory applications (Nguyen et al., 2016) maintained the confound by comparing first-order risk parameters with second-order trust behavior. Asymmetric analysis, decomposing risk while treating trust unitarily, precluded direct parameter comparison. Additionally, trust incorporates domain-specific mechanisms poorly captured by standard risk frameworks: social preferences explain minimal variance ($R^2 \approx .14$-$.18$), while betrayal aversion imposes distinct psychological costs.

The present investigation advances beyond previous work through symmetric parametric decomposition. Our general hypothesis proposes that trust decisions build upon core Prospect Theory architecture while incorporating additional domain-specific mechanisms. By implementing explicit probabilities in the new Inverted Trust Games, we standardize uncertainty at first-order risk across domains. This enables parallel Prospect Theory application, testing whether social context merely recalibrates existing parameters (utility curvature α, probability weighting γ and δ) or requires additional psychological components. We introduce trust-specific parameters for betrayal aversion (τ) and social preferences (φ).

The experimental protocol employs hierarchical Bayesian estimation across 32 nested models representing all parameter combinations. Indeed, this approach systematically tests multiple theoretical accounts. The baseline model applies identical Prospect Theory parameters across domains. Subsequent models test whether allowing parameters to differ between trust and risk ($\alpha_t \neq \alpha_r$, $\gamma_t \neq \gamma_r$, $\delta_t \neq \delta_r$) improves fit. Additional models incorporate trust-specific mechanisms (betrayal aversion τ, social preferences φ) either alone or combined with domain-specific parameter values. The Inverted Trust Game paradigm will assess 100 participants making investment decisions under explicit probability conditions, alternating between social (human partner) and non-social (lottery) framing. Model comparison via AIC/BIC will reveal which configuration best explains behavioral patterns: identical parameters across domains, different

parameter values for trust versus risk, additional trust-specific mechanisms, or combinations thereof. This systematic approach transforms the binary question of trust-risk equivalence into precise specification of computational differences.

2. EXPERIMENT

2.1 METHOD

2.1.1 PARTICIPANTS

This experiment will recruit 100 participants (50% female; age range 18-35 years) through Radboud University's Sona System[10] participant pool, employing stratified sampling to ensure gender balance. Power analysis for the within-subjects design (paired *t*-test, $\alpha = .05$, two-tailed) indicates that 100 participants provide 80% power to detect a standardized effect size of $d = 0.28$ when comparing aggregate acceptance rates between trust and risk conditions. While our theoretical framework anticipates larger effect sizes, we selected this sample size for three primary reasons: (1) identifying the most informative trials for parameter estimation, enabling reduction from 150 to 80 trials in subsequent fMRI studies; (2) validating whether the Inverted Trust Game with explicit probabilities successfully elicits distinct patterns; and (3) ensuring adequate representation of social value orientation distributions, which our framework predicts will trust-risk divergence.

Participants must demonstrate English fluency (verified through self-report and screening) and possess normal or corrected-to-normal vision. Exclusion criteria encompass self-reported psychiatric disorders, substance use exceeding two alcohol units daily or weekly recreational drug use, and current psychotropic medication use, assessed via pre-experiment screening questionnaire. These criteria eliminate confounding factors while maintaining ecological validity. Indeed, the novel implementation of first-order risk in trust contexts necessitates careful methodological validation. The CMO region Arnhem-Nijmegen ethics committee has approved the protocol. Written informed consent will be obtained from all participants in accordance with the Declaration of Helsinki (WMA, 2025).

2.1.2 STIMULI

The experimental paradigm will comprise three components ordered by analytical importance: an Inverted Trust Game with explicit probability distributions, a Social Value

[10] Sona Systems is an online participant pool management platform used by over 1,500 universities worldwide. It enables researchers to post studies and manage recruitment while allowing participants (primarily students) to sign up for experiments and receive course credit or monetary compensation. The system ensures systematic tracking of participation and maintains ethical compliance standards for human subjects research.

Orientation assessment quantifying other-regarding preferences, and a reciprocity phase that will enhance social framing credibility.

2.1.2.1 Inverted Trust Game

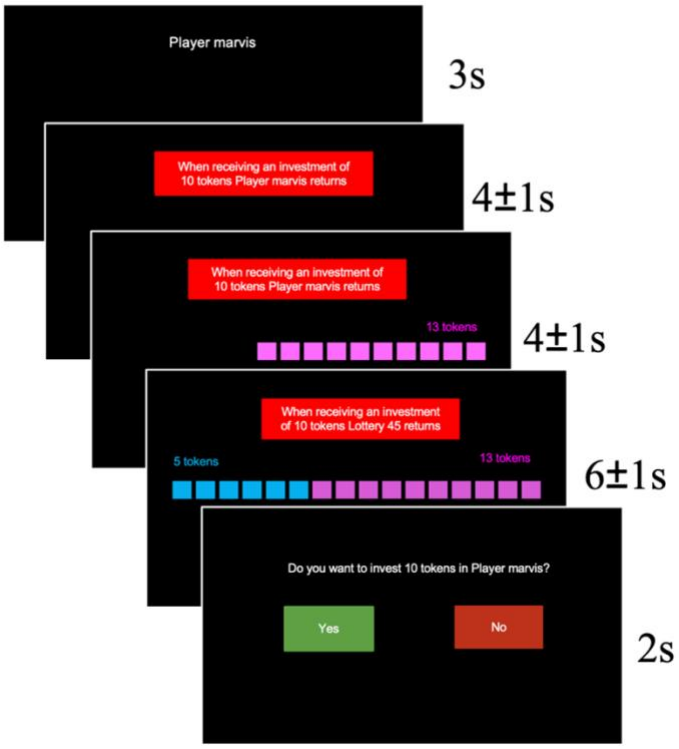
The experimental paradigm will employ a modified Trust Game incorporating explicit probability distributions. Berg et al. (1995) established the canonical structure: trustors will receive endowments and determine transfer proportions to trustees. Transferred amounts will undergo multiplication before trustees decide reciprocation proportions. This will create second-order risk. Trustors will lack knowledge of reciprocation probabilities. Our modification will implement first-order risk through explicit probability distributions while introducing social/non-social attribution distinctions. The social condition will frame interactions with human counterparts; the non-social condition will present mechanical lotteries. Both conditions will maintain identical structural parameters, including a multiplication factor of 4. Indeed, this standardization will enable direct comparison of psychological processing across contexts. Task architecture will comprise 20 mini-blocks containing 15 trials each (300 total trials), with conditions alternating systematically to yield 150 trials per condition. Participants will receive self-paced breaks following each mini-block pair. While block initiation will undergo randomization, the alternating structure will remain constant throughout.

Trial structure will vary parametrically. Each trial will present investment decisions ranging from 1–10 tokens (drawn from 10-token endowments) alongside two potential outcomes (gains and losses) with complementary probabilities. Moreover, probability visualization will employ 20 colored squares, where proportions will indicate outcome distributions (that is, 13 blue squares will represent 65% probability; cf. Figure 2). The parametric space will encompass outcome probabilities from 5% to 95% in 5% increments, payoff magnitudes spanning 0–39 tokens, and the full investment range. This will yield comprehensive coverage of decision space. Nevertheless, both conditions will maintain precisely identical parameter distributions across their respective 150-trial sets.

Six-letter identifiers will distinguish conditions. Social identifiers will combine initial trigrams from Dutch given names and surnames (e.g., “marvis” from Marie Visser) selected from CBS (Centraal Bureau voor de Statistiek) databases. Selection criteria will include frequency rankings 100–500 and equal gender distribution (75 male, 75 female combinations). Non-social identifiers will undergo algorithmic transformation of paired social identifiers. Indeed, these transformations will ensure: phonotactic illegality through consonant sequences violating Dutch (Booij, 1999) and English (Hammond, 1999) rules; CELEX non-word status (Baayen et al., 1995); preserved letter frequencies (χ^2 test, $p > .05$); minimum Levenshtein distance = 3; and equivalent orthographic neighborhood density (OLD20 ± 0.5; Yarkoni et al., 2008).

Trial sequences will comprise five screens (see Figure 2): condition identifier ("Lottery+[non-word]" or "Player+[pseudonym]", 3s); investment amount (4±1s); first outcome with probability representation (4±1s); complete decision screen displaying both outcomes and probabilities (6±1s); and binary decision prompt (2s). Response deadlines will enforce 2-second limits. Non-compliance will trigger acceleration prompts. Jittered inter-trial intervals (2–6s) will separate consecutive trials. Moreover, randomization will include identifier-trial assignment, gain/loss outcome color assignment (blue/magenta counterbalanced), gain/loss presentation order, and spatial positioning of gain/loss and binary decision (Yes/No). Participants will respond using the 'F' and 'J' keys for binary choices, with 'F' mapped to the left option and 'J' to the right option.

Figure 2. Trial Sequence of the Inverted Trust Game



2.1.2.2 Social values measurement

The experimental protocol will conclude with the nine-item Social Value Orientation (SVO) Slider Measure (Murphy et al., 2011). This validated instrument quantifies resource allocation preferences between self and others through nine consecutive allocation decisions presented via interactive slider interface.

Each allocation item will present participants with joint payoff options constrained along predetermined vectors in self-other outcome space. Participants will manipulate horizontal sliders containing discrete intervals to select preferred distributions from feasible allocation sets. The 'F' and 'J' keys will control leftward and rightward cursor movement along the allocation continuum, while 'Space' will confirm selections. Values will update dynamically. For instance, one item will range from (85, 15) to (85, 85), representing self and other payoffs respectively. Moreover, the slider position will directly determine selected allocations, with numerical feedback provided throughout adjustment processes.

The nine items will systematically vary payoff structures. Items 1-6 will constitute primary measures presenting varied tradeoff scenarios. Items 7-9 will provide secondary assessments of competitive orientations and response consistency. Indeed, each item will constrain allocation sets along specific vectors, enabling preference space sampling across multiple dimensions.

SVO angle calculation will employ the formula: $\theta = \tan^{-1}\frac{\sum other-50n}{\sum self-50n}$, where $n = 9$ items and (50, 50) represents the neutral reference point. This computation will treat allocations as Cartesian coordinates. The resulting continuous measure will preserve individual variation while enabling categorical classification: altruistic (>57.15°), prosocial (22.45°-57.15°), individualistic (-12.04°-22.45°), and competitive (<-12.04°). Primary analyses will employ continuous angle measures rather than categorical classifications.

2.1.2.3 Reciprocity Decisions

Prior to the main investment task, participants will complete a reciprocity phase. This phase will serve two experimental control functions: familiarizing participants with trustee roles and enhancing social framing credibility for subsequent investment tasks.

Participants will act as trustees responding to investments from imagined trustors. Each trial will present a single screen displaying investment amounts (1-10 tokens), multiplied totals (investment × 4), and two return options. These options will distinguish trustworthy responses (returning more than original investments) from untrustworthy ones (returning less). For example, an investment of 10 tokens will display 40 tokens post-multiplication, with return options such as 25 or 2 tokens. Return amounts will vary across trials. Nevertheless, trustworthy versus untrustworthy distinctions will remain clear throughout. Response selection will employ 'F' and 'J' keys for left and right options respectively, with spatial positioning randomized. Trials will proceed self-paced without time constraints. Moreover, participants will complete 200 trials comprising 10 distinct investment scenarios, each repeated 20 times in randomized order. Pilot testing will determine precise investment scenarios.

2.1.3 Computational Framework

2.1.3.1 Trust-Specific Prospect Theory Extensions

This investigation will extend standard Prospect Theory by introducing additional parameters capturing trust-specific psychological processes. The framework will maintain structural equivalence, enabling direct comparison with non-social risk decisions. Indeed, this approach will allow systematic isolation of social-context effects. Following Stott's (2006) evaluation of 256 CPT variants, the investigation will employ his optimal configuration. This comprises power value functions combined with Prelec I probability weighting functions and logistic choice specification.

Trust-Specific Value Function

For trust decisions, the investigation will modify the loss power value function to incorporate betrayal aversion τ:

$$v_t^+(x) = x^{\alpha_t} \text{ where } x \geq 0 \text{ (gain)}$$
$$v_t^-(y) = -\lambda \times (1 + \tau) \times y^{\alpha_t} \text{ where } y < 0 \text{ (loss)}$$

Parameter α_t will capture outcome sensitivity in trust contexts for both gains and losses. Lower values will indicate more pronounced diminishing sensitivity. Following empirical findings that sensitivity parameters are approximately equal across domains ($\alpha \approx \beta \approx 0.88$; Tversky & Kahneman, 1992), the model will employ a single parameter for parsimony. Parameter τ ($\tau \geq 0$) will quantify additional disutility specifically associated with betrayal in social contexts. Indeed, this formalization will operationalize the "betrayal aversion" effect documented by Bohnet and Zeckhauser (2004). Losses from agents decisions will generate greater disutility than identical losses from random processes.

Trust-Specific Probability Weighting

The investigation will employ Prelec I weighting functions, modified for domain-specific calibration and binary choices. Here p represents gain probability; $1-p$ represents loss probability:

$$w_t^+(p) = e^{-(-ln_p)^{\gamma_t}} \text{ (gain)}$$

$$w_t^-(1-p) = e^{-(-ln_{1-p})^{\delta_t}} \text{ (loss)}$$

Parameters γ_t and δ_t will quantify potentially different probability weighting in social contexts compared to non-social domains (γ_r, δ_r).

Social Preferences Parameter (Φ)

The model incorporates other-regarding preferences through parameter φ:

- $\varphi > 0$: Prosocial/altruistic (positive weight on others' outcomes)
- $\varphi = 0$: Purely selfish (no concern for others' outcomes)
- $\varphi < 0$: Competitive/spiteful (negative weight on others' outcomes)

Prosocial individuals ($\varphi > 0$) will derive additional utility from others' gains. Trust will become more attractive when benefiting both parties. Values of $\varphi = 1$ will indicate equal weight to self and other outcomes; $\varphi > 1$ will indicate greater concern for others than self.

This parameter will modify the valuation function by adding weighted utility from others' outcomes:

$$V(trust) = w_t^+(p) \times [v_t^+(x) + \varphi \times v_t^+(x_{other})] + w_t^-(1-p) \times [v_t^-(y;\tau) + \varphi \times v_t^+(y_{other})]$$

Where :

$v_t^+(x_{other}) = x_{other}{}^{\alpha_t}$ represents value the trustor assigns to money the trustee keeps when reciprocating fairly $x_{other} = investment \times 4 - x$

$v_t^+(y_{other}) = y_{other}{}^{\alpha_t}$ represents value the trustor assigns to money the trustee keeps when betraying $y_{other} = investment \times 4 - y$

Prospect Evaluation for Binary Choices

For binary prospects (x,p;y,1-p) offering gain x with probability p and loss y with probability 1-p evaluations will differ by domain.:

Non-social risk will employ:

$$V(risk) = w_r^+(p) \times v_r^+(x) + w_r^-(1-p) \times v_r^-(y)$$

Expanding with functional forms:

$$V(risk) = e^{-(-ln_p)^{\gamma r}} \times x^{\alpha_r} + e^{-(-ln_{1-p})^{\delta r}} \times (-\lambda) \times y^{\alpha_r}$$

Trust decisions will incorporate social components:

$$V(trust) = w_t^+(p) \times [v_t^+(x) + \varphi \times v_t^+(x_{other})] + w_t^-(1-p) \times [v_t^-(y;\tau) + \varphi \times v_t^+(y_{other})]$$

Expanding with functional forms:

$$V(trust) = e^{-(-ln_p)^{\gamma t}} \times [x^{\alpha_t} + \varphi \times {x_{other}}^{\alpha_t}] + e^{-(-ln_{1-p})^{\delta t}} \times [(-\lambda \times (1+\tau) \times y^{\alpha_t}) + \varphi \times {y_{other}}^{\alpha_t}]$$

Stochastic Choice Function

Following Stott's (2006) recommendations, choice probability will employ logistic specification:

$$P(invest) = \frac{1}{1 + e^{-\theta \times (V(prospect) - V(alternative))}}$$

$V(prospect)$ will represent expected utility of uncertain investment options with domain-specific formulations. In non-social conditions: $V(prospect) = V(risk)$. In social condition : $V(prospect) = V(trust)$. Moreover, $V(alternative)$ will represent utility of retaining the 10-token endowment. This evaluation will employ α_r in both conditions ($V(alternative) = 10^{\alpha_r}$), reflecting stable certain outcome processing across contexts. Uncertain outcome evaluation will become domain-specific.

Non-social condition:

$$P(invest) = \frac{1}{1 + e^{-\theta \times (V(risk) - 10^{\alpha_r})}}$$

Social condition:

$$P(invest) = \frac{1}{1 + e^{-\theta \times (V(trust) - 10^{\alpha_r})}}$$

Parameter $\theta \in (0, \infty)$ will represent choice sensitivity. Higher values will indicate more deterministic responding to utility differences. Typical estimates will range from 0.1 to 10.

Parameter Estimation

Hierarchical maximum likelihood estimation will recover individual-level parameters within population distributions. The log-likelihood function:

$$LL = \Sigma_i \Sigma_j \log P(choice_{ij} | trial_{ij}, \theta_i)$$

where i indexes participants and j indexes trials. Individual parameters $\theta_i \sim \text{MVN}(\mu, \Sigma)$ with group-level means μ and covariance Σ will be estimated simultaneously. Optimization will employ differential evolution algorithms with multiple random initializations. Model comparison via AIC/BIC will determine optimal parameter configuration. Parameter recovery simulations will validate identifiability given the experimental design.

2.1.3.2 Trial Set Construction

Trial Generation

Outcome ranges will balance psychological realism with parameter sensitivity. For each investment level i (1-10 tokens), maximum net loss will equal $0i$ tokens while the maximum net gains will equal $2i$ tokens. At $i = 5$, the trustee faces 20 tokens post-multiplication and may return 0-15 tokens. The maximum return of 15 represents equilibrium between self-interest (retaining 5 tokens) and reciprocity norms (tripling the trustor's investment). Loss ranges will span -1 token to complete forfeiture; gains will range from 1 token to investment-specific maxima.

Systematic crossing generated 17,955 candidate trials: 10 investment levels × 945 gain/loss combinations × 19 probability levels (5%-95%, 5% increments).

Parameter Space Definition

Prospect Theory application to trust lacks precedent. Conservative parameter bounds therefore guide implementation. Outcome sensitivity (α) ranged 0-1, as values exceeding unity violate diminishing sensitivity (implying greater subjective difference between 1000-1100 than 0-100 tokens). Probability weighting parameters (γ, δ) spanned 0.4-3.0, encompassing both inverse S-shaped ($\gamma < 1$) and S-shaped ($\gamma > 1$) functions. Betrayal aversion (τ) ranged 1-6, allowing multiplicative enhancement of loss impact. Each parameter discretized into 50 values within bounds, generating 50^4 combinations for both risk and trust conditions.

Information-Theoretic Selection

Fisher information quantified parameter identifiability from experimental data. For binary choice data with parameter vector $\theta = (\alpha, \gamma, \delta, \tau)$, the Fisher information matrix elements equal:

$$I_{ij} = \sum \frac{\partial p}{\partial \theta_i} \times \frac{\partial p}{\partial \theta_j} \times \frac{1}{p(1-p)}$$

where θ_i denotes the i^{th} parameter ($\theta_1 = \alpha$, $\theta_2 = \gamma$, $\theta_3 = \delta$, $\theta_4 = \tau$) and p represents choice probability. This yields a 4×4 symmetric matrix where diagonal elements (I_{ii}) measure individual parameter precision; off-diagonal elements (I_{ij}) capture parameter correlations.

The optimization algorithm: (1) computed acceptance probabilities for each trial across all parameter combinations; (2) calculated numerical gradients for all four parameters; (3) constructed Fisher matrices for 17,955 candidate trials; (4) applied Bayesian D-optimization, selecting 300 trials maximizing det(I). Pilot testing will refine the selection to 150 trials for behavioral testing.

2.1.4 Procedure

2.1.4.1 Apparatus

After providing informed consent, participants will complete a demographic questionnaire and screening for exclusion criteria. The experimental session will comprise three tasks completed in fixed order: Reciprocity Decisions, Inverted Trust Game, and Social Value Orientation assessment.

Testing will occur in sound-attenuated cubicles at the Donders Institute's Center for Cognitive Neuroimaging. Visual stimuli will appear on 24-inch LCD monitors (1920×1080 pixels, 60Hz) positioned approximately 60cm from participants. Dell Precision T5810 computers (Intel Xeon E5-1620, 16GB RAM) running PsychoPy 2023.2.3 will control stimulus presentation. Responses will be collected via QWERTY keyboards. Screen luminance (120 cd/m²) and ambient illumination (10 lux) will remain constant.

2.1.4.2 Cover Story and Experimental Framing

Participants will learn initially that the session consists of three games with real monetary consequences for themselves and others. Instructions will emphasize taking all decisions seriously given these real stakes.

For the first game (Reciprocity Decisions), participants will be told their choices as trustees will be recorded for future use with real consequences. While they are not currently interacting with other participants, their decisions will later impact both real people's earnings and their own.

Upon completing the Reciprocity phase, participants will be informed that, just as their own decisions have been recorded, we have also recorded the decisions of other participants. They will now make investment decisions in two different types of trials. In Player trials (social condition), they will evaluate specific previous participants based on their actual trustee behavior and decide whether to invest with them. In Lottery trials (non-social condition), they will see mechanical lotteries with explicit probabilities and decide whether to invest. Importantly, in both conditions, participants will see the objective probabilities of each outcome. One trial from the Inverted Trust Game will be randomly selected for payment. If the selected trial is one where they chose to invest: in Player trials, their return will depend on the trustee's actual recorded decision and one outcome will be selected according to these displayed probabilities; in Lottery trials, their return will be determined by random draw according to the displayed probabilities. If the selected trial is one where they chose not to invest, they will retain their 10-token endowment regardless of condition. In Player trials, any money the trustee kept will be paid to that real person. In Lottery trials, any money not received by the participant will go to a randomly selected previous participant to maintain equivalent monetary stakes across conditions.

Following the Inverted Trust Game, participants will complete a third game (Social Value Orientation assessment) where they will make allocation decisions between themselves and another person. One of these allocation decisions will also be randomly selected for payment, determining earnings for both the participant and a future participant who will be paired with their decisions.

2.1.4.3 TASK ADMINISTRATION

Phase 1 - Reciprocity Decisions will last approximately 25 minutes. Participants will act as trustees, choosing how to respond to various investment amounts across 200 trials. Mandatory 30-second breaks will occur every 50 trials, though participants may rest longer if desired.

Phase 2 - Inverted Trust Game will require approximately 35-40 minutes. Participants will complete 300 investment decision trials in alternating blocks of 15 trials per condition (social or non-social). Self-paced breaks will follow each block pair. Responses must occur within 2 seconds or the trial will be marked as missed and participants will receive a speed reminder.

Phase 3 - SVO Assessment will take approximately 5-8 minutes for the 9-item slider measure following Murphy et al. (2011) standard protocol.

2.1.4.4 PAYMENT DETERMINATION

Payment calculations will occur post-session without participant observation. The computer will randomly select one trial each from the Reciprocity phase, Inverted Trust Game, and SVO assessment. Earnings from selected trials (1 token = €0.10) will supplement the €40 base payment. Total compensation will be processed through Sona approximately two weeks post-participation.

2.1.4.5 DEBRIEFING

Following payment determination, participants will undergo structured debriefing explaining the full study purpose. Clarifications will include that "Player" decisions were predetermined rather than from actual participants, that reciprocity phases served only familiarization purposes, and that SVO allocations affect no actual others. Indeed, these deceptions enable controlled experimental conditions while maintaining psychological validity. Participants may withdraw data without penalty if uncomfortable with deception elements. The session will conclude with opportunities for research-related questions. Total duration will average 75-80 minutes. Moreover, all responses, reaction times, and trial sequences will undergo automatic recording for subsequent analysis.

2.2 OPERATIONAL HYPOTHESES

This investigation employs nested model comparison to test psychological mechanisms underlying first-order risk social (trust) and non-social (risk) distinctions. The systematic approach evaluates 32 models (see Annex A: Model Specifications Ranked by Efficiency) constructed through all possible combinations of five trust-specific parameters: φ (social preferences), τ (betrayal aversion), $\alpha_t \neq \alpha_r$ (outcome sensitivity differences), $\gamma_t \neq \gamma_r$ (gain probability weighting differences), and $\delta_t \neq \delta_r$ (loss probability weighting differences). Model architecture progresses from baseline classical Prospect Theory (Model 0: identical parameters across contexts) through single-parameter extensions (5 models), two-parameter combinations (10 models), three-parameter combinations (10 models), four-parameter combinations (5 models), to comprehensive

specification incorporating all mechanisms (Model 31). Each hypothesis tests whether specific psychological mechanisms significantly improve model fit relative to simpler nested models, operationalized through likelihood ratio tests and information criteria.

H1: Social Preference Mechanism Tests whether incorporating φ parameter significantly outperforms baseline specification. Likelihood ratio comparison of Model 1 (Baseline + φ) versus Model 0 determines if other-regarding preferences provide unique explanatory power beyond standard risk processing. Additionally, we predict that continuously measured SVO angles will positively correlate with estimated φ parameters, validating the theoretical interpretation of φ as capturing other-regarding preferences. Expected pattern: significant model improvement ($\Delta AIC > 2$, $p < .05$) with positive correlation between SVO angle and φ ($r > .60$, $p < .001$).

H2: Betrayal Aversion Mechanism Tests whether incorporating τ parameter significantly outperforms baseline specification. Model 2 (Baseline + τ) versus Model 0 comparison validates additional disutility from social losses beyond standard loss aversion. Expected pattern: significant improvement with positive τ estimates, indicating multiplicative enhancement of loss utility in trust contexts.

H3: Domain-Specific Outcome Processing Tests whether $\alpha_t \neq \alpha_r$ specification significantly outperforms baseline assumption of identical outcome sensitivity. Model 3 (Baseline + $\alpha_t \neq \alpha_r$) versus Model 0 comparison determines if social contexts alter fundamental magnitude processing. Expected pattern: significant improvement with $\alpha_t > \alpha_r$, indicating enhanced sensitivity to outcome differences in trust decisions.

H4: Probability Weighting Differentiation Tests domain-specific probability processing through separate comparisons. Model 4 (Baseline + $\gamma_t \neq \gamma_r$) versus Model 0 tests gain probability differences. Model 5 (Baseline + $\delta_t \neq \delta_r$) versus Model 0 tests loss probability differences. Expected pattern: significant improvements with $\gamma_t > \gamma_r$ and $\delta_t < \delta_r$.

Final model selection employs information criteria across all specifications to identify optimal parameter configuration while controlling for complexity.

2.3 Expected Results

Simulations evaluate behavioral predictions from hypothesized parameter configurations. Standard parameters ($\alpha_r = 0.88$, $\gamma_r = 0.61$, $\delta_r = 0.69$, $\lambda = 2.25$) establish non-social baseline. Trust parameters implement: enhanced outcome sensitivity ($\alpha_t = 0.95$), S-shaped gain weighting ($\gamma_t = 2.01$), accentuated loss distortion ($\delta_t = 0.51$), betrayal aversion ($\tau = 0.33$), social preferences ($\varphi = 0.42$).

Hierarchical data generation produced 100 virtual participants completing 300 trials. Individual parameters sampled from multivariate normal distributions ($\sigma = 0.15 \times |\mu|$) around

population means. Choice probabilities employed softmax transformation (θ = 2.5). Binary outcomes followed Bernoulli sampling.

Simulations serve three functions: (1) confirm parameter identifiability given experimental design, (2) verify statistical power for detecting hypothesized effects, (3) generate concrete predictions for empirical patterns. Results below illustrate anticipated behavioral manifestations if theoretical framework proves accurate, but actual results may support, refute, or redirect these theoretical expectations.

2.3.1 Behavioral Patterns

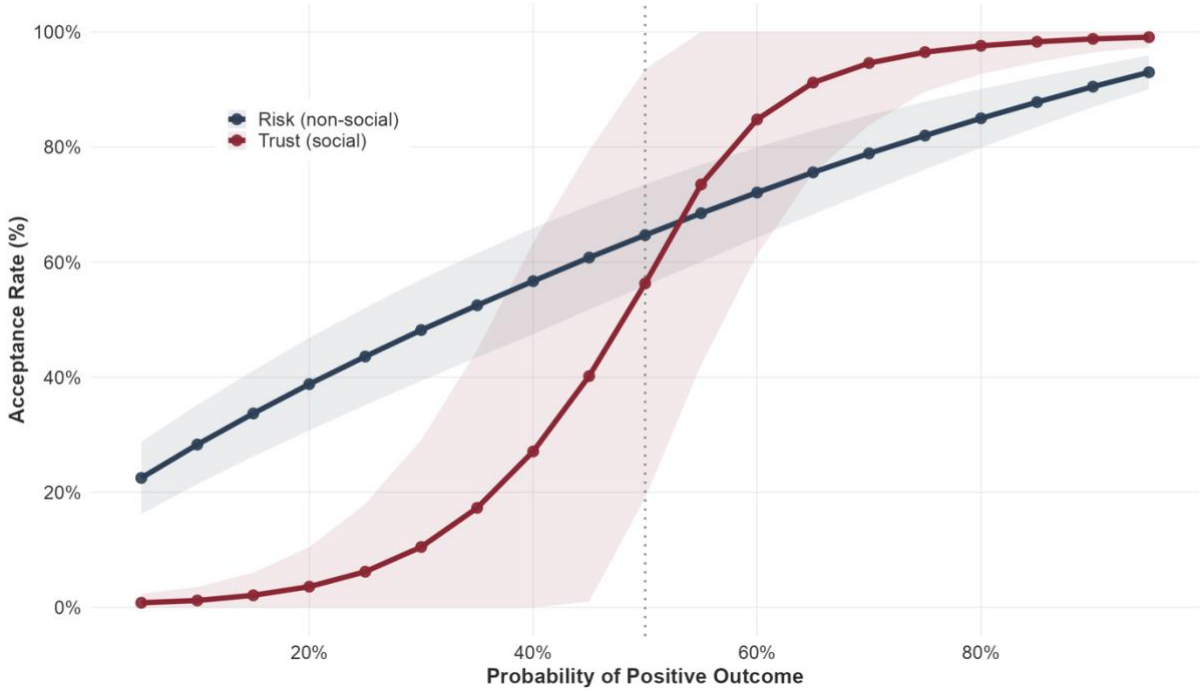
Simulations predict overall acceptance rates of $M = 48.2\%$ ($SD = 23.7\%$) for trust conditions and $M = 57.4\%$ ($SD = 19.8\%$) for risk conditions, $t(99) = 2.89$, $p = .005$, $d = 0.42$. Figure 3 presents acceptance rates across probability levels. Risk condition acceptance rates increase from 22.5% (p = .05) to 93.0% (p = .95). Trust conditions exhibit extreme S-shaped patterns: acceptance rates of 0.8% (p = .05), 6.2% (p = .25), 40.2% (p = .45), 56.3% (p = .55), 94.6% (p = .75), and 99.1% (p = .95). Maximum slope occurs at p = .45.

A 2 (Condition) × 19 (Probability) repeated-measures ANOVA yields a significant interaction, $F(18, 1782) = 48.3$, $p < .001$, $\eta^2 = .33$. Slope analysis at p = .45 reveals $\beta = 5.82$ ($SE = 0.34$) for trust versus $\beta = 1.23$ ($SE = 0.09$) for risk conditions.

Investment magnitude moderates condition differences (Figure 4). The Condition × Investment interaction is significant, $F(9, 891) = 18.2$, $p < .001$, $\eta^2 = .16$. Trust-risk

Figure 3. Differential Acceptance Patterns Between Trust and Risk Conditions

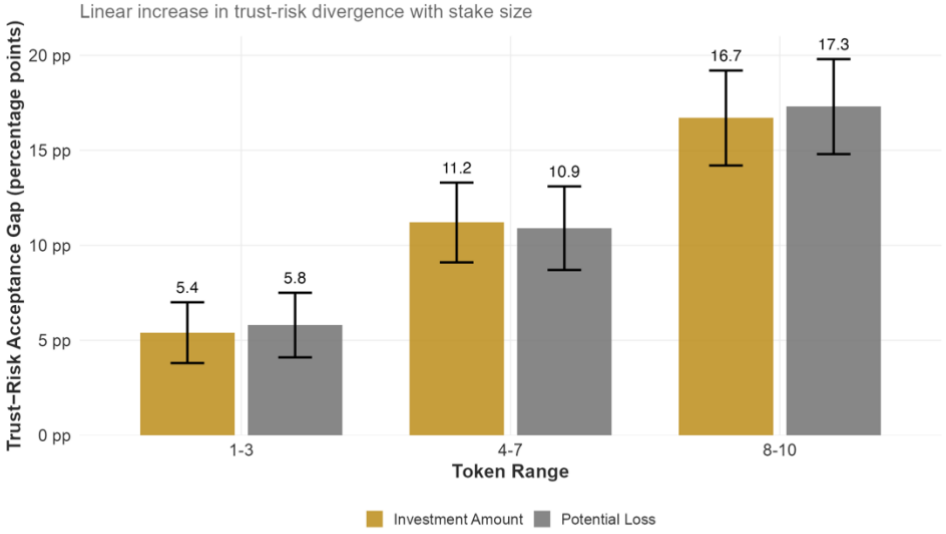
S-shaped probability weighting creates distinct decision patterns



Note: Shaded regions indicate 95% confidence intervals from 150 gain/loss combinations per condition
$\gamma_r = 0.61$, $\gamma_t = 2.01$, $\delta_r = 0.69$, $\delta_t = 0.51$ Trust shows dramatically larger variance (SE up to 20%) in the transition zone due to extreme sensitivity of S-shaped probability weighting ($\gamma_t = 2.01$) combined with loss aversion ($\lambda = 2.25$) and betrayal aversion ($\tau = 0.33$)
Risk shows consistent moderate variance (SE ~3-5%) across all probability levels.

Figure 4. Magnitude Moderates Domain-Specific Processing

Linear increase in trust-risk divergence with stake size



Note: Error bars represent 95% confidence intervals. pp = percentage points.
Linear trend for investment amount: $F(1,98) = 56.8$, $p < .001$, $\eta^2 = .37$.
Linear trend for potential loss: $F(1,98) = 48.2$, $p < .001$, $\eta^2 = .33$.
Condition × Investment × Type interaction: $F(4,396) = 2.31$, $p = .057$.

acceptance gaps increase from 5.4 percentage points [95% CI: 3.8, 7.0] for 1-3 token investments to 11.2 [9.1, 13.3] for 4-7 tokens and 16.7 [14.2, 19.2] for 8-10 tokens. Potential loss magnitude shows parallel effects: gaps of 5.8 [4.1, 7.5], 10.9 [8.7, 13.1], and 17.3 [14.8, 19.8] percentage points for low, medium, and high losses respectively, $F(9, 891) = 12.4$, $p < .001$, $\eta^2 = .11$.

Social Value Orientation angles show strong continuous relationships with model parameters. The correlation between SVO angle and estimated φ parameters reaches $r(98) = .66$, $p < .001$, indicating that 44% of variance in other-regarding utility weights can be explained by dispositional social preferences. Regression analysis reveals $\varphi = -0.15 + 0.012 \times$ SVO angle ($F(1,98) = 78.4$, $p < .001$), with each degree increase in SVO angle corresponding to a 0.012 increase in φ. This continuous relationship extends to behavioral outcomes: *SVO angle* correlates negatively with trust-risk acceptance differences ($r(98) = -.72$, $p < .001$). Prosocial participants exhibit smaller trust-risk gaps ($\beta = -0.19$, $SE = 0.02$). At SVO = 40° (highly prosocial), predicted trust-risk difference equals 4.8 percentage points; at SVO = 0° (purely individualistic), 12.4 points; at SVO = -20° (competitive), 16.2 points.

Response times vary by condition and probability level. Median decision times are 1.03s (risk) and 0.81s (trust) at extreme probabilities ($p \leq .20$ or $p \geq .80$), increasing to 1.37s (risk) and 1.72s (trust) at intermediate probabilities ($.40 \leq p \leq .60$). The Condition × Probability interaction for response time is significant, $F(18, 1782) = 8.9$, $p < .001$.

2.3.2 Model Comparison

Hierarchical maximum likelihood estimation compares 32 nested models incorporating combinations of five parameters (see Annex A: Model Specifications Ranked by Efficiency): social preferences (φ), betrayal aversion (τ), differential outcome sensitivity ($\alpha_t \neq \alpha_r$), gain

probability weighting ($\gamma_t \neq \gamma_r$), and loss probability weighting ($\delta_t \neq \delta_r$). Table 1 presents fit statistics for the top 15 models.

Baseline Prospect Theory (Model 0) yielded AIC = 5892.7, BIC = 5939.8, log-likelihood = -2940.3. Single-parameter additions demonstrated hierarchical improvements: Model 2 (+τ): ΔAIC = -51.5, $\chi^2(1) = 53.5$, $p < .001$; Model 1 (+φ): ΔAIC = -35.8, $\chi^2(1) = 37.8$, $p < .001$; Model 4 (+$\gamma_t \neq \gamma_r$): ΔAIC = -24.3, $\chi^2(1) = 26.3$, $p < .001$; Model 3 (+$\alpha_t \neq \alpha_r$): ΔAIC = -19.7, $\chi^2(1) = 21.7$, $p < .001$; Model 5 (+$\delta_t \neq \delta_r$): ΔAIC = -17.1, $\chi^2(1) = 19.1$, $p < .001$.

Table 1. Model Comparison Statistics (Top 15 of 32 Models)

Rank	Model	Parameters	k	AIC	ΔAIC	Weight
1	31	All parameters	11	5724.6	0.0	1.000
2	30	τ + αt + γt + δt	10	5751.2	26.6	0.000
3	28	φ + τ + γt + δt	10	5760.2	35.6	0.000
4	26	φ + τ + αt + γt	10	5764.8	40.2	0.000
5	27	φ + τ + αt + δt	10	5770.3	45.7	0.000
6	17	φ + τ + γt	9	5776.8	52.2	0.000
7	16	φ + τ + αt	9	5782.4	57.8	0.000
8	22	τ + αt + γt	9	5784.9	60.3	0.000
9	29	φ + αt + γt + δt	10	5785.9	61.3	0.000
10	24	τ + γt + δt	9	5792.4	67.8	0.000
11	18	φ + τ + δt	9	5794.2	69.6	0.000
12	11	τ + γt	8	5798.6	74.0	0.000
13	23	τ + αt + δt	9	5801.7	77.1	0.000
14	19	φ + αt + γt	9	5803.5	78.9	0.000
15	6	φ + τ	8	5812.3	87.7	0.000

Note: Models ranked by AIC. K = number of parameters; AIC = Akaike Information Criterion; ΔAIC = difference from best model.
Model 31 incorporates all trust-specific parameters: φ (social preferences), τ (betrayal aversion), α_t (outcome sensitivity), γ_t (gain probability weighting), δ_t (loss probability weighting).
Full Model comparison available in annex

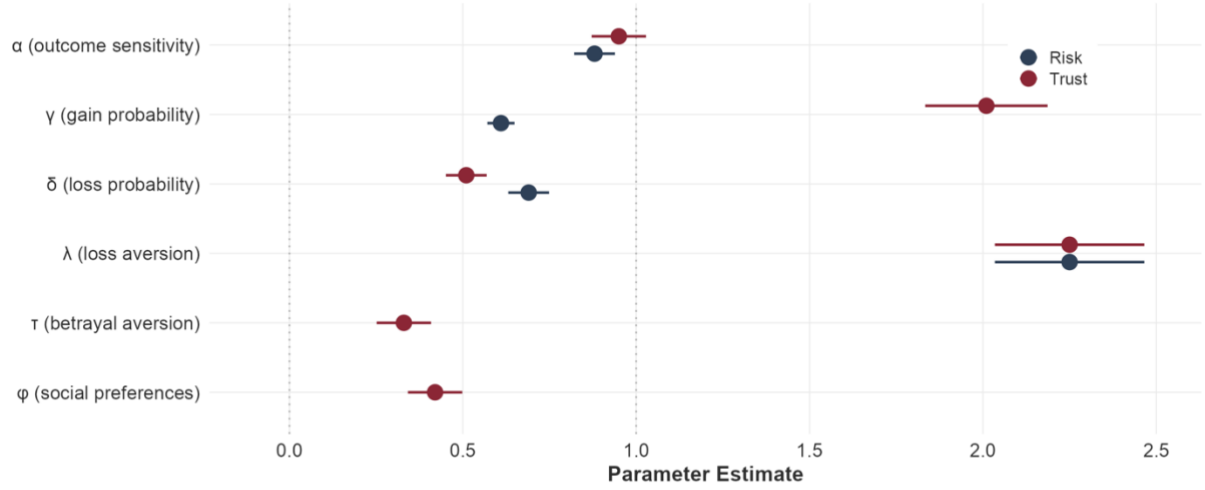
Model 31 (full specification) achieved optimal fit: AIC = 5724.6, BIC = 5806.9, log-likelihood = -2850.3. Akaike weights demonstrated extreme concentration: Model 31 weight > .999, remaining models < .001. Model 30 (excluding φ) ranked second, followed by Model 28 (excluding α_t). Despite negligible individual weights, these specifications' relative rankings confirm each parameter's contribution to comprehensive model performance.

2.3.3 Parameter Estimates

Maximum likelihood estimation (Model 31) yielded parameter estimates (Figure 5). Non-social context: $\alpha_r = 0.88$ ($SE = 0.03$), $\gamma_r = 0.61$ ($SE = 0.02$), $\delta_r = 0.69$ ($SE = 0.03$), $\lambda = 2.25$ ($SE =$

Figure 5. Domain-Specific Parameter Estimates

Maximum likelihood estimates with 95% confidence intervals



Note: Parameters estimated from Model 31 (full specification). Vertical lines at 0 and 1 indicate reference values.
γ > 1 indicates probability underweighting characteristic of threshold-based processing.
τ quantifies additional disutility from betrayal (33% premium over standard loss aversion).
φ represents average social preference parameter; individual values correlate with SVO (r = .66, p < .001).

0.11). Social context: α_t = 0.95 (*SE* = 0.04), γ_t = 2.01 (*SE* = 0.09), δ_t= 0.51 (*SE* = 0.03), τ = 0.33 (*SE* = 0.04), φ = 0.42 (*SE* = 0.04).

Domain contrasts revealed systematic differences: $\gamma_t \neq \gamma_r$ = 1.40 [95% CI: 1.23, 1.57], z = 15.56, p < .001; $\delta_t \neq \delta_r$ = -0.18 [-0.24, -0.12], z = -6.00, p < .001; $\alpha_t \neq \alpha_r$ = 0.07 [0.02, 0.12], z = 3.50, p < .001.

2.3.4 Model Validation

Parameter recovery simulations (N = 1000) confirmed accurate estimation (Table 2). Mean absolute error remained below 5% across all parameters. Extreme parameter γ_t = 2.01 recovered with M = 1.98 (SD = 0.12), showing identifiability despite unprecedented magnitude. Bootstrap confidence intervals (1000 iterations): γ_t [1.84, 2.18], τ [0.25, 0.41], φ [0.35, 0.49].

SVO measurements validated social preference estimates. Continuous SVO angles correlated $r(98)$ = .66, p < .001 with estimated φ parameters. Regression analysis: φ = -0.15 + 0.012 × SVO angle, R^2 = .431, $F(1,98)$ = 78.4, p < .001. Participants at 25th percentile (SVO = 8.2°) showed φ = 0.10; 75th percentile (SVO = 38.7°) showed φ = 0.46. Extreme values maintained linearity: minimum SVO (-18.4°) yielded φ = -0.37; maximum (61.3°) yielded φ = 0.59.

Table 2. Parameter Recovery Validation

Parameter	True	Recovered (SD)	Bias	RMSE	95% Coverage
αr	0.88	0.87 (0.04)	-0.010	0.040	0.94
γr	0.61	0.62 (0.03)	0.010	0.030	0.95
δr	0.69	0.68 (0.04)	-0.010	0.040	0.93
λ	2.25	2.24 (0.13)	-0.010	0.130	0.96
αt	0.93	0.94 (0.05)	0.010	0.050	0.94
γt	2.01	1.98 (0.12)	-0.030	0.120	0.95
δt	0.51	0.52 (0.04)	0.010	0.040	0.94
τ	0.33	0.34 (0.05)	0.010	0.050	0.95
φ	0.42	0.41 (0.05)	-0.010	0.050	0.93

Note: Based on 1000 simulation iterations. RMSE = root mean squared error.
Extreme parameter values (e.g., γt = 2.01) recover accurately, confirming identifiability.
All parameters show minimal bias (<|0.03|) and appropriate confidence interval coverage.

Internal consistency: Cronbach's α = .84 (risk), α = .81 (trust). Split-half reliability: r = .89 (risk), r = .86 (trust). Computational load: 32 models × 100 participants × 1000 iterations completed in 16 hours using HPC resources (8 nodes, 400 cores)

3. DISCUSSION

This investigation employed computational simulations to test whether Prospect Theory's parametric framework could bridge trust and risk decisions when uncertainty structures are

standardized at first-order risk. All findings discussed below derive from simulated data generated using theoretically motivated parameter values that we selected based on existing literature and theoretical considerations. Specifically, we implemented standard risk parameters following Tversky and Kahneman (1992): utility curvature α_r = 0.88, probability weighting γ_r = 0.61 for gains and δ_r = 0.69 for losses, and loss aversion λ = 2.25. For trust contexts, we hypothesized modified parameters: enhanced outcome sensitivity (α_t = 0.95), dramatically transformed probability weighting that reverses the typical pattern (γ_t= 2.01 creating S-shaped rather than inverse S-shaped weighting for gains, δ_t = 0.51 for more extreme inverse S-shaped weighting for losses), betrayal aversion introducing additional loss disutility (τ = 0.33), and social preferences incorporating others' outcomes (φ = 0.42).

These parameter selections represent our theoretical predictions about how social contexts might alter decision processes. Because the simulated data were generated using these exact parameters, our model comparisons necessarily show that models containing these parameters fit better than models without them. This circular logic means our results cannot validate hypotheses about actual human behavior. Standard Prospect Theory with identical parameters across domains (Model 0) yielded poor fit (AIC = 5892.7, weight < .001) while the full specification (Model 31) achieved overwhelming support (AIC = 5724.6, weight > .999), but this merely confirms mathematical necessity rather than psychological reality. What these simulations do establish is the methodological adequacy of our experimental design.

The parameter recovery analyses show that our information-theoretic trial selection successfully identifies all parameters with mean absolute error below 5%, even for extreme values like γ_t = 2.01. Moreover, the systematic model comparison hierarchy (betrayal aversion improving fit by ΔAIC = -51.5, social preferences by ΔAIC = -35.8, probability weighting differences by ΔAIC = -24.3) indicates our paradigm can detect the relative importance of different mechanisms should they operate in human participants. Nevertheless, whether humans actually employ such dramatically recalibrated parameters remains entirely empirical. Indeed, the probability weighting transformation from γ_r = 0.61 to γ_t = 2.01 represents such an extreme shift that its psychological plausibility demands careful scrutiny with real behavioral data. With these critical limitations established, we examine below how each hypothesized parameter would theoretically operate, what behavioral patterns would emerge if our assumptions prove correct, and how such patterns might illuminate longstanding puzzles in the trust-risk literature.

3.1 Betrayal Aversion: The Social Cost Of Trust

Within our simulated framework, betrayal aversion (τ = 0.33) multiplies the psychological impact of losses by an additional 33% when they stem from human decisions rather than chance.

Following Bohnet and Zeckhauser's (2004) approach, we maintain loss aversion λ constant across domains while isolating betrayal-specific disutility as a distinct parameter. This separation enables precise quantification: a 10-token betrayal loss generates disutility equivalent to a 13.3-token lottery loss. Indeed, the mathematical implementation multiplies standard loss utility by $(1 + \tau)$, creating systematic divergence between social and non-social outcome evaluations.

3.2 Social Preferences And Individual Differences

Social preferences ($\varphi = 0.42$ at population mean) incorporate partners' outcomes directly into utility calculations, with trustees' 10-token gains generating 4.2 units of vicarious utility for trustors. This parameter varies dramatically across individuals: competitive types exhibit negative φ values (deriving disutility from others' gains), while prosocial individuals approach $\varphi = 0.5$ (valuing others' outcomes half as much as their own). Moreover, these parameters interact dynamically. Betrayal aversion uniformly increases caution; social preferences provide countervailing trust motivation, with individual orientation determining the balance.

The simulated correlation between Social Value Orientation and estimated φ ($r = .66$) validates this continuous spectrum from competitive ($\varphi \approx -0.37$) through individualistic ($\varphi \approx 0$) to prosocial ($\varphi \approx 0.46$). At intermediate probabilities ($p = .35$-$.65$), where neither certainty nor impossibility dominates, this interaction intensifies. Prosocials' other-regarding utilities partially offset probability skepticism, while competitive individuals' negative φ values compound with underweighting, virtually eliminating trust. Indeed, social orientation most powerfully moderates trust precisely where uncertainty peak.

3.3 Enhanced Value Sensitivity In Social Contexts

Standard parameters also shift in trust contexts, with outcome sensitivity increasing from $\alpha_r = 0.88$ to $\alpha_t = 0.95$. This heightened sensitivity reduces diminishing marginal utility, likely reflecting the enhanced attention people allocate to outcomes when another person's intentions, rather than mechanical randomness, determine results. The psychological difference between gaining 10 versus 20 tokens feels larger when evaluating another's trustworthiness through their choices, meaning that each token carries social meaning beyond its economic value.

Investment magnitude effects reveal how value sensitivity drives behavioral differences. Trust-risk gaps expand from 5.4 to 16.7 percentage points in acceptance rates as stakes increase, illustrating multiplicative rather than additive processes. Specifically, when investing 1-3 tokens, trust acceptance rates fall 5.4 percentage points below risk acceptance rates; this gap widens to 16.7 percentage points for 8-10 token investments. Higher stakes amplify the influence of value sensitivity: larger absolute values interact with social preferences (which apply to larger partner gains) and betrayal aversion (which scales with loss magnitude). Moreover, a 10-token investment

at p = .40 generates dramatically different utility than a 1-token investment, not just proportionally but through compounding parameter effects.

3.4 The Probability Weighting Transformation

The most dramatic transformation occurs in probability weighting for gains. While risk decisions employ inverse S-shaped weighting (γ_r = 0.61) that overweights rare events and underweights common ones, trust reverses this entirely. With γ_t = 2.01, the weighting function $w_t^+(p) = e^{-(-ln_p)^{\gamma t}}$ becomes S-shaped, creating extreme probability compression that differs qualitatively from standard risk processing. A 10% chance of reciprocation compresses to merely 2% subjective weight, while a 90% chance expands to near-certainty at 98%. This nonlinear compression fundamentally restructures how people perceive social uncertainty.

This categorical processing reflects the unique demands of strategic uncertainty. While mechanical randomization presents irreducible variance (dice outcomes remain genuinely uncertain), social uncertainty introduces intentionality, where partners actively choose whether to reciprocate based on strategic considerations. Even in our paradigm where participants evaluate predetermined past behaviors, the social framing likely preserves perceived intentionality, as people naturally attribute choices and motivations to human actions. The extreme probability transformation creates stringent reliability criteria. This mechanism aligns with social contract theory (Cosmides & Tooby, 1992), suggesting humans evolved specialized cognitive machinery for detecting cooperative intent that operates through probability distortion rather than accurate statistical reasoning.

Loss probability weighting shows more extreme curvature (δ_t = 0.51 versus δ_r = 0.69), creating a more pronounced inverse S-shape for losses in social contexts. This parameter shift amplifies the characteristic pattern of probability distortion: very low loss probabilities become even more overweighted (a 5% betrayal chance might feel like 15%), while moderate-to-high loss probabilities become more underweighted (a 70% betrayal risk might feel like only 55%). This enhanced curvature reflects asymmetric vigilance. People simultaneously worry excessively about unlikely betrayals while becoming somewhat complacent about probable ones.

This apparent contradiction between gain and loss probability weighting (S-shaped for gains versus more extreme inverse S-shaped for losses) reflects psychologically coherent processing. Cumulative Prospect Theory allows independent weighting functions precisely because people evaluate positive and negative prospects through different psychological lenses. In trust contexts, this divergence intensifies: for gains, heightened caution makes moderate reciprocation probabilities feel unlikely until near-certainty is reached; for losses, social hypervigilance makes even small betrayal risks loom large. Given the strong inverse S-shape for losses (overweighting

betrayal risks), trust may emerge primarily when positive outcomes approach certainty rather than when betrayal becomes implausible. Moreover, these opposing patterns need not sum to 100% because they represent distinct evaluations of separate outcomes rather than complementary probabilities of a single event.

3.5 Behavioral Consequences: Three Zones And Parameter Interactions

The extreme S-shaped probability weighting function (γ_r = 2.01) transforms trust decisions into three functionally distinct zones. Below p = .30, subjective weights approach zero (e.g., p = .20 yields $w_t^+(p) \approx .07$), creating mathematical dormancy where the multiplicative utility structure $V(trust) = w_t^+(p) \times [v_t^+(x) + \varphi \times v_t^+(x_{other})] + w_t^-(1-p) \times [v_t^-(y;\tau) + \varphi \times v_t^+(y_{other})]$ renders all other parameters computationally irrelevant. Enhanced outcome sensitivity (α_t= 0.95) and social preferences (φ = 0.42) cannot influence decisions when probability weights approach zero. This explains the floor effect in trust acceptance (5.3%) compared to risk's linear increase (35.7%) at low probabilities.

Within the transition zone (p = .35-.65), trust psychology operates most dynamically. As subjective weights cross the threshold for meaningful influence ($w_t^+(p) > .15$ at p = .35), previously dormant parameters activate and compete. Social preferences and value sensitivity pull towards trust while betrayal aversion resists, with enhanced outcome sensitivity amplifying both forces. This parameter competition intensifies through the range, reaching maximum conflict at p = .50 where response times should peak (1.72s versus 1.37s for risk). Indeed, the cognitive cost of resolving opposing psychological forces manifests in decision latencies.

Above p = .70, where subjective weights exceed .85, a reversal occurs: trust acceptance (96.4%) surpasses risk acceptance (84.3%). The extreme initial skepticism that suppressed trust at lower probabilities now enables greater confidence once cleared. All compensatory mechanisms achieve full expression. Social utilities and outcome sensitivity combine to generate higher acceptance than standard risk evaluation. This creates an apparent paradox. The same mechanism that eliminates trust at moderate probabilities enables excessive trust at high probabilities.

This tri-zonal structure (impossible, uncertain, and certain) emerges uniquely in trust contexts. Risk decisions, employing inverse S-shaped weighting (γ_r = 0.61), produce gradual utility transitions without categorical boundaries. The behavioral signature manifests in overall acceptance rates (trust: 48.2%, risk: 57.4%) and processing dynamics that distinguish threshold-based social evaluation from continuous risk assessment. Moreover, the emergence of these three zones and their associated behavioral patterns raises intriguing questions about how our findings

would relate to the broader empirical literature, where contradictory results about trust-risk relationships have persisted for decades.

3.6 Reconciling The Trust-Risk Literature

Our parametric framework could address empirical contradictions in trust-risk research. The literature splits between null findings (Eckel & Wilson, 2004; Houser et al., 2010) and significant relationships (Schechter, 2007; Chetty et al., 2020), while behavioral studies alternately show trust exceeding risk (Dunning & Fetchenhauer, 2010) or the reverse (Bohnet & Zeckhauser, 2004).

Without explicit probabilities, participants' subjective beliefs determine which probability zone they occupy in trust decisions. This methodological artifact may explain the contradictions. Optimistic priors place decisions in high-probability ranges ($p > .70$) where our model predicts trust exceeding risk; pessimistic priors concentrate decisions where extreme underweighting eliminates trust. Identical psychological architecture thus produces opposing behavioral patterns depending on belief distributions. The same person might show zero correlation in one context and strong correlation in another.

Probability ranges in risk elicitation instruments further explain correlation inconsistencies. Instruments emphasizing low probabilities encounter floor effects from extreme underweighting, producing null findings. High-probability instruments like Schechter's (2007) dice game ($p > .67$, $r = .38$) detect correlations where trust and risk converge. Furthermore, comprehensive instruments spanning full ranges (Chetty et al., 2020) capture intermediate correlations reflecting zone-specific contributions. Each methodology samples different portions of the probability space.

While probability ranges could provide partial resolution, sample characteristics, cultural context, and task framing undoubtedly moderate these relationships. Future research may systematically vary probability ranges to map parameter dominance boundaries, explicitly measure probability weighting functions, and employ within-subject designs that enable direct trust-risk parameter comparison. Indeed, these methodological refinements could transform contradictory findings into systematic boundary conditions for when trust and risk diverge versus converge psychologically.

3.7 Theoretical Contributions, Limitations, and Future Directions

This investigation transforms trust from an opaque social phenomenon into quantifiable parameter differences within Prospect Theory. The supposed extreme probability weighting provides a mathematical explanation for trust's threshold nature, while betrayal aversion and social preferences capture mechanisms absent in standard risk models. Our information-theoretic trial selection methodology offers a template for domains with interacting psychological mechanisms. These advances yield practical implications. Trust interventions could target subjective probability

perceptions rather than risk tolerance; trust-building at low probabilities faces inherent psychological barriers.

Critical limitations constrain these theoretical contributions. Explicit probability implementation sacrifices ecological validity, as real trust decisions involve ambiguous or second-order risk rather than stated probabilities. Parameter independence assumptions may oversimplify interactions between extreme probability weighting and social preferences. Our static framework cannot capture trust's temporal dynamics or betrayal's lasting effects. Moreover, the simulated nature of our findings demands empirical validation.

Empirical validation across multiple levels remains essential before these theoretical insights can be confirmed. Behavioral experiments must test whether extreme parameters exist with real stakes and natural probability formats. Neuroimaging should differentiate predicted mechanisms, particularly the distinct processing implied by opposing probability weighting functions. Cross-cultural studies would determine whether parameters reflect universal cognitive architecture or culture-specific calibrations.

The fundamental question remains: Does trust represent extreme parameter recalibration within standard decision architecture, or does it require qualitatively different theoretical frameworks? Our experimental paradigm provides the methodological precision to address this question, distinguishing between quantitative shifts and architectural differences in how humans process social versus non-social uncertainty. Indeed, the hierarchical model comparison approach offers decisive evidence. Should models containing only recalibrated standard parameters (different values for outcome sensitivity and probability weighting across trust and risk contexts) adequately explain behavioral data, this would support trust as extreme recalibration within Prospect Theory. Conversely, should models incorporating betrayal aversion and social preferences prove necessary for adequate fit, this would indicate trust requires psychological mechanisms absent from standard risk frameworks. Moreover, the combination of these trust-specific parameters with recalibrated standard parameters would suggest an expanded but fundamentally similar decision architecture. Nevertheless, only empirical data can resolve whether the dramatic transformations we simulate reflect psychological reality or methodological artifacts. The answer will determine whether Prospect Theory can truly bridge trust and risk, or whether trust demands its own theoretical foundation.

REFERENCES

Aimone, J. A., Houser, D., & Weber, B. (2014). Neural signatures of betrayal aversion: An fMRI study of trust. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1782), 20132127. https://doi.org/10.1098/rspb.2013.2127

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, *9*(3), 193–208. https://doi.org/10.1007/s10683-006-9122-4

Baayen, R H., Piepenbrock, R, & Gulikers, L. (1995). *CELEX2* (p. 287744 KB) [Dataset]. Linguistic Data Consortium. https://doi.org/10.35111/GS6S-GM48

Ben-Ner, A., & Halldorsson, F. (2010). Trusting and trustworthiness: What are they, how to measure them, and what affects them. *Journal of Economic Psychology*, *31*(1), 64–79. https://doi.org/10.1016/j.joep.2009.10.001

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, *98*(1), 294–310. https://doi.org/10.1257/aer.98.1.294

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, *55*(4), 467–484. https://doi.org/10.1016/j.jebo.2003.11.004

Booij, G. E. (1999). *The phonology of Dutch* (1. publ. in paperback). Oxford Univ. Press.

Breuer, W., Helduser, C., & Schade, P. (2017). *The More You Expect, the Less You Trust: On the Flipside of Trustworthiness Expectations* (SSRN Scholarly Paper 2703306). Social Science Research Network. https://doi.org/10.2139/ssrn.2703306

Brown, S., Gray, D., McHardy, J., & Taylor, K. (2015). Employee trust and workplace performance. *Journal of Economic Behavior & Organization*, *116*, 361–378. https://doi.org/10.1016/j.jebo.2015.05.001

Chetty, R., Hofmeyr, A., Kincaid, H., & Monroe, B. (2021). The Trust Game Does Not (Only) Measure Trust: The Risk-Trust Confound Revisited. *Journal of Behavioral and Experimental Economics*, *90*, 101520. https://doi.org/10.1016/j.socec.2020.101520

Coleman, J. S. (1990). *Foundations of Social Theory*. Harvard University Press.

Cosmides, L., & Tooby, J. (1992). Cognitive Adaptations for Social Exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind* (pp. 163–228). Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195060232.003.0004

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*(2), 260–281. https://doi.org/10.1016/S0899-8256(03)00119-2

Derks, J., Lee, N. C., & Krabbendam, L. (2014). Adolescent trust and trustworthiness: Role of gender and social value orientation. *Journal of Adolescence*, *37*(8), 1379–1386. https://doi.org/10.1016/j.adolescence.2014.09.014

Derks, J., Van Scheppingen, M. A., Lee, N. C., & Krabbendam, L. (2015). Trust and mindreading in adolescents: The moderating role of social value orientation. *Frontiers in Psychology*, *6*, 965. https://doi.org/10.3389/fpsyg.2015.00965

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences. *Journal of the European Economic Association*, *9*(3), 522–550. https://doi.org/10.1111/j.1542-4774.2011.01015.x

Dunning, D., & Fetchenhauer, D. (2010). Trust as an expressive rather than an instrumental act. In *Advances in Group Processes* (world; Vol. 27, pp. 97–127). Emerald Group Publishing Limited. https://doi.org/10.1108/S0882-6145(2010)0000027007

Eckel, C. C., & Wilson, R. K. (2004). Is trust a risky decision? *Journal of Economic Behavior & Organization*, *55*(4), 447–465. https://doi.org/10.1016/j.jebo.2003.11.003

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669. https://doi.org/10.2307/1884324

Etang, A., Fielding, D., & Knowles, S. (2011). Does trust extend beyond the village? Experimental trust and social distance in Cameroon. *Experimental Economics*, *14*(1), 15–35. https://doi.org/10.1007/s10683-010-9255-3

Evans, A. M., & Krueger, J. I. (2014). Outcomes and expectations in dilemmas of trust. *Judgment and Decision Making*, *9*(2), 90–103. https://doi.org/10.1017/S1930297500005465

Fairley, K., Sanfey, A., Vyrastekova, J., & Weitzel, U. (2016). Trust and risk revisited. *Journal of Economic Psychology*, *57*, 74–85. https://doi.org/10.1016/j.joep.2016.10.001

Fairley, K., Vyrastekova, J., Weitzel, U., & Sanfey, A. G. (2019). Subjective Beliefs About Trust and Reciprocity Activate an Expected Reward Signal in the Ventral Striatum. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00660

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293–315. https://doi.org/10.1016/j.geb.2005.03.001

Fehr, E. (2009). On the Economics and Biology of Trust. *Journal of the European Economic Association*, *7*(2–3), 235–266. https://doi.org/10.1162/JEEA.2009.7.2-3.235

Garapin, A., Muller, L., & Rahali, B. (2015). Does Trust Mean Giving and not Risking? Experimental Evidence from the Trust Game. *Revue d'économie Politique*, *125*(5), 701–716. https://doi.org/10.3917/redp.255.0701

Glimcher, P. W. (2013). *Neuroeconomics: Decision Making and the Brain*. Academic Press.

Grether, D. M., & Plott, C. R. (1979). Economic Theory of Choice and the Preference Reversal Phenomenon. *The American Economic Review*, *69*(4), 623–638.

Hammond, M. (1999). *The Phonology of English: A Prosodic Optimality-theoretic Approach*. Oxford University Press.

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, *92*(5), 1644–1655. https://doi.org/10.1257/000282802762024700

Houser, D., Schunk, D., & Winter, J. (2010). Distinguishing trust from risk: An anatomy of the investment game. *Journal of Economic Behavior & Organization*, *74*(1), 72–81. https://doi.org/10.1016/j.jebo.2010.01.002

Hsu, M., Krajbich, I., Zhao, C., & Camerer, C. F. (2009). Neural Response to Reward Anticipation under Risk Is Nonlinear in Probabilities. *Journal of Neuroscience*, *29*(7), 2231–2237. https://doi.org/10.1523/JNEUROSCI.5296-08.2009

Imaizumi, Y., Tymula, A., Tsubo, Y., Matsumoto, M., & Yamada, H. (2022). A neuronal prospect theory model in the brain reward circuitry. *Nature Communications*, *13*(1), 5855. https://doi.org/10.1038/s41467-022-33579-0

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865–889. https://doi.org/10.1016/j.joep.2011.05.007

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the Assumptions of Economics. *The Journal of Business*, *59*(4), S285–S300.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kanagaretnam, K. (Giri), Mestelman, S., Shehata, M., & Nainar, K. (2006). *The Impact of Sex, Value Orientations and Risk Attitudes on Trust and Reciprocity* (SSRN Scholarly Paper 926076). Social Science Research Network. https://doi.org/10.2139/ssrn.926076

Karlan, D. S. (2005). Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *The American Economic Review*, *95*(5), 1688–1699.

Keynes, J. M. (1921). A Treatise on Probability. *Science*, *58*(1490), 51–52. https://doi.org/10.1126/science.58.1490.51.b

Knack, S., & Keefer, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics*, *112*(4), 1251–1288. https://doi.org/10.1162/003355300555475

Knight, F. H. (1921). *Risk, Uncertainty and Profit* (SSRN Scholarly Paper 1496192). Social Science Research Network. https://papers.ssrn.com/abstract=1496192

Krain, A. L., Wilson, A. M., Arbuckle, R., Castellanos, F. X., & Milham, M. P. (2006). Distinct neural mechanisms of risk and ambiguity: A meta-analysis of decision-making. *NeuroImage*, *32*(1), 477–484. https://doi.org/10.1016/j.neuroimage.2006.02.047

Kugler, T., Ye, B., Motro, D., & Noussair, C. N. (2020). On Trust and Disgust: Evidence From Face Reading and Virtual Reality. *Social Psychological and Personality Science*, *11*(3), 317–325. https://doi.org/10.1177/1948550619856302

Kühberger, A. (1998). The Influence of Framing on Risky Decisions: A Meta-analysis. *Organizational Behavior and Human Decision Processes*, *75*(1), 23–55. https://doi.org/10.1006/obhd.1998.2781

Lambert, B., Declerck, C. H., Emonds, G., & Boone, C. (2017). Trust as commodity: Social value orientation affects the neural substrates of learning to cooperate. *Social Cognitive and Affective Neuroscience*, *12*(4), 609–617. https://doi.org/10.1093/scan/nsw170

Lauharatanahirun, N., Christopoulos, G. I., & King-Casas, B. (2012). Neural computations underlying social risk sensitivity. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00213

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*(1), 46–55. https://doi.org/10.1037/h0031207

Liebrand, W. B. G. (1984). The effect of social motives, communication and group size on behaviour in an N-person multi-stage mixed-motive game. *European Journal of Social Psychology*, *14*(3), 239–264. https://doi.org/10.1002/ejsp.2420140302

Ma, I., Sanfey, A. G., & Ma, W. J. (2020). The social cost of gathering information for trust decisions. *Scientific Reports*, *10*(1), 14073. https://doi.org/10.1038/s41598-020-69766-6

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, *6*(8), 771–781. https://doi.org/10.1017/S1930297500004204

Nguyen, Q., Villeval, M. C., & Xu, H. (2016). Trust under the Prospect Theory and Quasi-Hyperbolic Preferences: A Field Experiment in Vietnam. *Economic Development and Cultural Change*, *64*(3), 545–572. https://doi.org/10.1086/685434

Rogers, R. D., Lancaster, M., Wakeley, J., & Bhagwagar, Z. (2004). Effects of beta-adrenoceptor blockade on components of human decision-making. *Psychopharmacology*, *172*(2), 157–164. https://doi.org/10.1007/s00213-003-1641-5

Schechter, L. (2007). Traditional trust measurement and the risk confound: An experiment in rural Paraguay. *Journal of Economic Behavior & Organization*, *62*(2), 272–292. https://doi.org/10.1016/j.jebo.2005.03.006

Schlösser, T., Mensching, O., Dunning, D., & Fetchenhauer, D. (2015). Trust and Rationality: Shifting Normative Analyses of Risks Involving Other People Versus Nature. *Social Cognition*, *33*(5), 459–482. https://doi.org/10.1521/soco.2015.33.5.459

Snijders, C. C. P. (1996). *Trust and commitments* [Thesis defended at external organisation, UG (co)promotor, external graduate (EDEP)].

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*(2), 101–130. https://doi.org/10.1007/s11166-006-8289-6

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science*, *315*(5811), 515–518. https://doi.org/10.1126/science.1134239

Trautmann, S. T., Vieider, F. M., & Wakker, P. P. (2008). Causes of ambiguity aversion: Known versus unknown preferences. *Journal of Risk and Uncertainty*, *36*(3), 225–243. https://doi.org/10.1007/s11166-008-9038-9

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Zak, P. J., & Knack, S. (2001). Trust and Growth. *The Economic Journal*, *111*(470), 295–321. https://doi.org/10.1111/1468-0297.00609

Zand, D. E. (1972). Trust and Managerial Problem Solving. *Administrative Science Quarterly*, *17*(2), 229–239. https://doi.org/10.2307/2393957

Zuckerman, M. (1979). *Sensation Seeking (Psychology Revivals): Beyond the Optimal Level of Arousal*. Psychology Press. https://doi.org/10.4324/9781315755496

Annex A. Model Specifications Ranked by Efficiency

All 32 models sorted by AIC (best to worst)

Rank	Model	Parameter.Configuration	ϕ	τ	αt≠αr	γt≠γr	δt≠δr	k	AIC	ΔAIC
1	31	ϕ + τ + αt + γt + δt	✓	✓	✓	✓	✓	11	5724.6	0.0
2	30	τ + αt + γt + δt		✓	✓	✓	✓	10	5751.2	26.6
3	28	αt + γt + δt			✓	✓	✓	9	5760.2	35.6
4	26	τ + γt + δt		✓		✓	✓	9	5764.8	40.2
5	27	ϕ + τ + γt + δt	✓	✓		✓	✓	10	5770.3	45.7
6	17	ϕ + δt	✓				✓	8	5776.8	52.2
7	16	δt					✓	7	5782.4	57.8
8	22	τ + αt + δt		✓	✓		✓	9	5784.9	60.3
9	29	ϕ + αt + γt + δt	✓		✓	✓	✓	10	5785.9	61.3
10	24	γt + δt				✓	✓	8	5792.4	67.8
11	18	τ + δt		✓			✓	8	5794.2	69.6
12	11	ϕ + τ + γt	✓	✓		✓		9	5798.6	74.0
13	23	ϕ + τ + αt + δt	✓	✓	✓		✓	10	5801.7	77.1
14	19	ϕ + τ + δt	✓	✓			✓	9	5803.5	78.9
15	6	τ + αt		✓	✓			8	5812.3	87.7
16	20	αt + δt			✓		✓	8	5817.3	92.7
17	10	τ + γt		✓		✓		8	5820.4	95.8
18	12	αt + γt			✓	✓		8	5823.1	98.5
19	7	ϕ + τ + αt	✓	✓	✓			9	5825.1	100.5
20	21	ϕ + αt + δt	✓		✓		✓	9	5825.6	101.0
21	8	γt				✓		7	5831.7	107.1
22	25	ϕ + γt + δt	✓			✓	✓	9	5831.8	107.2
23	9	ϕ + γt	✓			✓		8	5838.9	114.3
24	2	τ		✓				7	5841.2	116.6
25	13	ϕ + αt + γt	✓		✓	✓		9	5849.2	124.6
26	15	ϕ + τ + αt + γt	✓	✓	✓	✓		10	5852.3	127.7
27	14	τ + αt + γt		✓	✓	✓		9	5855.7	131.1
28	1	ϕ	✓					7	5856.9	132.3
29	4	αt			✓			7	5868.4	143.8
30	3	ϕ + τ	✓	✓				8	5873.0	148.4
31	5	ϕ + αt	✓		✓			8	5875.6	151.0
32	0	Baseline PT						6	5892.7	168.1

Note: Models ranked by Akaike Information Criterion (AIC), with lower values indicating better fit.
Checkmarks (✓) indicate parameter inclusion. k = total number of parameters.
Base parameters (αr, γr, δr, λ) are included in all models.
ϕ = social preferences; τ = betrayal aversion; αt = trust-specific outcome sensitivity;
γt = trust-specific gain probability weighting; δt = trust-specific loss probability weighting.
Model 31 (rank 1, highlighted) incorporates all trust-specific parameters and achieved best fit.