# Kantian Morality in Non-Human Primates:

## Structural Estimation from Ultimatum and Social Dilemma Games

Benson Njogu Mbau[1]

Lixing Sun[2]

Francesca De Petrillo[3]

Wei Lu[4]

Daniel L. Chen[5]

[1] oTree

[2] Department of Biological Sciences, Central Washington University

[3] School of Natural and Environmental Sciences, Newcastle University

[4] Zicklin School of Business, Baruch College, City University of New York

[5] Toulouse School of Economics; CNRS; Harvard Radcliffe Institute; Hoover Institution

June 2026

**Abstract**

Do non-human primates exhibit behaviour consistent with Kantian-consistent preference structure? We bring the structural preference framework of Van Leeuwen and Alger (2024) to bear on three primate datasets spanning chimpanzees, bonobos, and squirrel monkeys across ultimatum and social dilemma games (n = 3,957 trials; 24 subjects). For each individual, we estimate a Kantian-consistent preference weight $\kappa$, the weight placed on the universalized outcome of one's strategy, alongside standard inequity aversion parameters. Our primary identification comes from the squirrel monkey multi-game dataset (Dataset 3), where cross-game variation in the Kantian differential $\Delta K$ provides within-individual identification of $\kappa$ independent of inequity aversion. The pooled estimate of $\bar{\kappa} = 0.593$ in squirrel monkeys is statistically consistent with the Alger-Weibull (2013) evolutionarily stable strategy prediction of $\kappa = 0.5$, and two individuals reject $\kappa = 0$ at the 5% level. The two great ape datasets suffer from $\kappa$–$\alpha$ collinearity that renders $\kappa$ non-identified in those designs; their likelihood profiles are flat across the full $\kappa$ range, and we treat them as non-identification demonstrations rather than

independent evidence. The BIC-preferred structured specification for Dataset 3 is the Pure Kantian form ($\kappa$ free, $\alpha = \beta = 0$); the even simpler heuristic model achieves a marginally lower raw BIC, indicating the data are too noisy to strongly distinguish structured preference models. Our contribution is primarily methodological: we show that a multi-game design with cross-game $\Delta K$ variation can structurally identify Kantian-consistent preferences in non-human animals, and we provide suggestive evidence of positive $\kappa$ in squirrel monkeys.

**Keywords:** Kantian morality, *Homo moralis*, primate behaviour, structural estimation, ultimatum game, social dilemma, inequity aversion, evolutionary economics

## 1. Introduction

Whether moral behaviour has deep evolutionary roots, predating language, culture, and philosophical reflection, is a foundational question in evolutionary biology, behavioural economics, and moral philosophy (Trivers, 1971; de Waal, 2006; Henrich, 2016). One of the most debated forms of moral reasoning is Kantian universalizability: the principle that an agent evaluates an action not only by its immediate material consequences but by asking "what if everyone acted this way?" (Kant, 1785). If this preference architecture is uniquely human, we would expect it to emerge only with the cognitive and linguistic capacities of *Homo sapiens*. If, instead, universalizability reflects an evolutionarily stable strategy in social environments characterised by assortative matching, we would expect traces or prototypes of it to appear in species with rich social lives that predate the human lineage (Alger and Weibull, 2013).

This paper asks whether structural tools developed for human laboratory experiments can detect Kantian-consistent preference structure in non-human primates in decision making. We treat this as a test of out-of-sample predictive power rather than a claim about conscious moral reasoning: $\kappa$ is a revealed preference parameter that captures the degree to which behaviour is better predicted by a utility function that includes the Kantian payoff term $K(s)$ than by one that ignores it.

The preference framework we employ sits within a rapidly growing research programme. Alger and Weibull (2013) provided the theoretical foundation, showing that $\kappa > 0$ emerges as an evolutionarily stable strategy under assortative social matching, with $\kappa = 0.5$ as the equilibrium value under moderate assortativity. Van Leeuwen and Alger (2024) translated this theory into an estimable utility function and estimated it on human laboratory data. Most recently, Lu, Dhanda, Chen, and Hansen (2025) applied the same utility function to large language models.

These important findings lead to the central question of our paper: whether the preference structure itself has evolutionary roots that predate our species.

We estimate κ from three existing datasets with varying identification strength. The squirrel monkey multi-game dataset (Dataset 3) is the inferential centrepiece: the same individuals play three games with different Kantian payoff differentials ΔK, providing within-individual identification of κ independent of inequity aversion. The two great ape datasets (Datasets 1 and 2) serve as non-identification demonstrations: their designs suffer from κ–α collinearity that renders κ non-identified in those designs. Profile likelihood exercises (Appendix E) confirm that any value of $\kappa \in [0, 1]$ is observationally equivalent in these datasets; the high point estimates ($\bar{\kappa} \approx 0.94$ and 0.90) reflect optimizer corner solutions, not evidence of high κ. This identification hierarchy is the organising principle of our analysis.

Our central finding is that a multi-game design with cross-game ΔK variation can structurally identify κ in non-human animals, and that applying this design to squirrel monkeys yields an estimated $\bar{\kappa} = 0.593$, statistically consistent with the Alger-Weibull ESS prediction of $\kappa = 0.5$. The great ape datasets do not identify κ owing to collinearity; we report their optimizer outputs for completeness but draw no comparative inference from them. We emphasise that DS3 provides suggestive but not conclusive evidence of positive κ, and that the cross-species pattern is descriptive rather than an identified phylogenetic gradient.

The remainder of the paper is structured as follows. Section 2 presents the theoretical framework. Section 3 describes the three datasets and identification strategies. Section 4 outlines the estimation approach. Section 5 reports results. Section 6 discusses interpretation, limitations, and implications. Section 7 concludes.

## 2. Theoretical Framework

### 2.1 The Utility Function

We adopt the preference specification from Van Leeuwen and Alger (2024), which nests Kantian morality within a broader framework that also accommodates distributional preferences. For an agent choosing strategy s, utility is:

$$U(s; \kappa, \alpha, \beta) = (1 - \kappa) \cdot V(s) + \kappa \cdot K(s)$$

where the distributional component V(s) follows the Fehr-Schmidt (1999) inequity aversion specification:

$$V(s) = x(s) - \alpha \cdot \max\{x(s) - y(s), 0\} - \beta \cdot \max\{y(s) - x(s), 0\}$$

In these expressions, x(s) is the agent's own payoff, y(s) is the partner's payoff, K(s) is the Kantian payoff, the payoff the agent would receive if both players chose strategy s, and the four parameters are: $\kappa \in [0,1]$, the Kantian-consistent preference weight; $\alpha \geq 0$, aheadness aversion; $\beta \geq 0$, behindness aversion; and $\lambda > 0$, a logit precision parameter. The model nests several special cases: $\kappa = 0$ with free $\alpha$, $\beta$ is the standard Fehr-Schmidt model; $\kappa = 1$ with $\alpha = \beta = 0$ is pure Kantian utility; $\kappa = 0$, $\alpha = \beta = 0$ is pure self-interest.

### 2.2 The Kantian Term K(s) Across Game Formats

In the bilateral ultimatum game (Dataset 1), K(5,1) = 5, K(3,3) = 3, and K(1,5) = 1. A Kantian agent is drawn toward the equal split because K(3,3) = 3 is the most universalizable outcome. In the triadic competitive ultimatum game (Dataset 2), K(offer) = 8 − offer. In the social dilemma games (Dataset 3), the Kantian payoff is the diagonal of the payoff matrix. The Kantian differential $\Delta K$ = K(Coop) − K(Defect) is 1 in the Prisoner's Dilemma and 2 in both the Stag Hunt and Hawk-Dove game. This cross-game variation in $\Delta K$ is the key identification device for $\kappa$ in Dataset 3.

### 2.3 The Two Game Families: Players, Decisions, and the Role of κ

The three datasets we analyse fall into two game families: ultimatum games (Datasets 1 and 2) and simultaneous social dilemmas (Dataset 3). The two families differ in their strategic structure, but the Kantian morality parameter $\kappa$ plays the same conceptual role in both: it shifts the agent's preference between actions that maximise their own material payoff and actions that maximise the universalised payoff K(s).

Ultimatum games (Datasets 1 and 2). A proposer chooses how to divide a fixed pool of tokens between themselves and a responder. In Dataset 1 (bilateral UG, chimps, and bonobos), the proposer chooses one of three pre-specified allocations: (5,1), (3,3), or (1,5), where the first number is the proposer's share and the second is the responder's. In Dataset 2 (triadic UG, chimpanzees), two proposers simultaneously submit integer offers between 0 and 8 tokens to a single responder, who selects the higher offer; the winning proposer keeps the remainder of their endowment. In both designs, a purely self-interested proposer ($\kappa = 0$, $\alpha = \beta = 0$) chooses the maximum-own-payoff allocation: (5,1) in Dataset 1, offer = 0 in Dataset 2. A Kantian proposer reasons differently: they ask "what would I receive if my partner made the same proposal to me?" In Dataset 1, this universalisation gives K(5,1) = 5, K(3,3) = 3, and K(1,5) = 1, so the equal split (3,3) becomes increasingly attractive as $\kappa$ rises. In Dataset 2, K(offer) = 8 − offer because if both proposers made the same offer, each would win with probability 0.5 and keep 8

− offer tokens. Higher κ thus pulls behaviour toward the equal split in Dataset 1 and toward keeping more tokens (lower offers) in Dataset 2.

Social dilemmas (Dataset 3). Two squirrel monkeys, A and B, simultaneously choose between a cooperative action (C) and a non-cooperative action (D) without observing each other's choice. Each player's payoff depends on the joint outcome. In the Prisoner's Dilemma, mutual cooperation gives each player 2, mutual defection gives each player 1, and one-sided defection gives the defector 3 and the cooperator 0. A purely self-interested agent defects: defection yields a higher own payoff regardless of what the partner does. A purely Kantian agent cooperates: K(C) = 2 (the payoff if both cooperate) exceeds K(D) = 1 (the payoff if both defect). The Stag Hunt and Hawk-Dove games share this structure but with different payoffs and different Kantian differentials $\Delta K = K(C) - K(D)$, which is the key variation that identifies κ in Dataset 3.

The role of κ. In every game format, κ governs the weight placed on the universalised outcome relative to the agent's own material payoff. When κ = 0, only own payoff and inequity aversion (α, β) matter; when κ = 1, only the universalised outcome K(s) matters; intermediate values blend the two. Identifying κ requires payoff variation that moves K(s) independently of own payoff and the distributional gap (x − y), the multi-game structure of Dataset 3, with its cross-game variation in ΔK, provides exactly this within-individual variation. Datasets 1 and 2 do not, which is the source of the κ–α collinearity discussed in Section 3.5.

### 2.4 Evolutionary Foundations

Alger and Weibull (2013) demonstrate that κ = 0.5 is evolutionarily stable under moderate assortativity. Higher assortativity implies higher ESS κ; random matching implies κ = 0 as the ESS. We treat κ = 0.5 as a theoretically motivated null hypothesis and test whether each species' κ is consistent with it via a likelihood ratio test.

## 3. Data

### 3.1 Overview

We draw on three existing experimental datasets. Datasets are numbered in the order introduced: Dataset 1 is the chimp/bonobo bilateral ultimatum game, Dataset 2 is the chimpanzee triadic competitive ultimatum game, and Dataset 3 is the squirrel monkey multi-game social dilemmas. Dataset 3 is the identification centrepiece; Datasets 1 and 2 provide consistency evidence.

| Dataset | Species | Game | N trials | N subjects |
|---|---|---|---|---|
| DS1: rspb20211937_si_002.xlsx | Chimps & Bonobos | Bilateral ultimatum game | 1,535 (1,023 baseline) | 9 |
| DS2: competitive_altruism_dataset.csv | Chimpanzees | Triadic competitive UG | 879 (878 baseline) | 5 |
| DS3: NIHMS1028576-supplement-Raw_Data.csv | Squirrel Monkeys | PD + Stag Hunt + Hawk-Dove | 4,200 (2,056 simultaneous) | 10 (5 pairs) |

*Table 1. Summary of the three primate datasets.*

### 3.2 Dataset 1: Chimp/Bonobo Bilateral Ultimatum Game

The bilateral UG dataset comes from Sánchez-Amaro and Rossano (2021), covering nine great apes, five chimpanzees (*Pan troglodytes*: Lobo, Sandra, Frodo, Tai, Dorien), and four bonobos (*Pan paniscus*: Lome, Kuno, Luiza, Yasa), at the Wolfgang Köhler Primate Research Center, Leipzig Zoo, 2020–2021. Chimpanzees and bonobos are the two species most closely related to humans, sharing a common ancestor with our lineage approximately 6–8 million years ago. Both species live in large fission-fusion social groups and engage in cooperative hunting, food sharing, and coalition formation, but they differ behaviourally in ways relevant to fairness research: chimpanzees are more competitive and male-dominated, while bonobos are more egalitarian and tolerant of co-feeding. Prior work has documented that these great apes show sensitivity to inequity in food-distribution tasks (Brosnan et al., 2010), making them a natural starting point for examining whether structural fairness preferences extend beyond humans.

We chose this dataset for two reasons. First, for chimpanzees and bonobos, their phylogenetic proximity to humans makes them the most informative non-human comparison for any preference structure documented in human laboratory subjects: a positive finding here speaks directly to whether the preference architecture predates the divergence of our lineage from the other Hominidae. Second, the dataset includes a within-subjects social leverage manipulation, a design feature rare in the comparative literature, that shifts the strategic cost of equal-split proposals while holding the available payoff allocations constant. The dataset contains 1,535 trials, of which 1,023 are retained in the baseline sample (leverage $\in$ {yes, no}). The leverage=control condition, in which no partner is present, is excluded from the main estimation and used only as a descriptive baseline.

### 3.3 Dataset 2: Chimpanzee Triadic Competitive Ultimatum Game

The triadic UG dataset comes from Sánchez-Amaro, Maurits, and Haun (2024), covering five chimpanzees (Pan troglodytes: Tai, Sandra, Azibo, Riet, Frodo) at the Wolfgang Köhler Primate Research Center, Leipzig Zoo, 2022. Three of these subjects (Tai, Sandra, Frodo) also appear in Dataset 1, enabling within-individual qualitative comparison across the bilateral and triadic game formats. Chimpanzees engage in coalition formation, reciprocal exchange, and competition for status within their groups; the triadic game format speaks directly to these naturalistic strategic features by introducing a competitive element absent from the bilateral design. Prior work has documented chimpanzee sensitivity to outside options and to social context (Engelmann and Tomasello, 2017), suggesting they may modulate offer behaviour in response to the strategic environment.

We chose this dataset for two reasons. First, the triadic structure, in which two proposers simultaneously bid for a single responder, extends the strategic environment beyond the bilateral case and tests whether competitive pressure alters fairness behaviour. Second, the dataset's within-condition variation between triadic and dyadic sessions provides a behavioural counterfactual: dyadic sessions remove the competitive motive and isolate residual generosity, even if structural separation of preference parameters remains limited in this design. The dataset contains 879 trials, with 439 triadic and 440 dyadic sessions; the baseline sample retains simultaneous triadic trials (n = 219) and all dyadic trials (n = 440), excluding consecutive triadic trials in which the second proposer observed the first mover's offer. Consecutive trials are retained in robustness checks (Appendix C).

### 3.4 Dataset 3: Squirrel Monkey Multi-Game Social Dilemmas

The squirrel monkey dataset comes from Vale, Williams, Schapiro, Lambeth, and Brosnan (2019), covering ten Bolivian squirrel monkeys (*Saimiri boliviensis*) forming five pairs at a North American primate facility. Squirrel monkeys are New World monkeys whose lineage diverged from the human lineage approximately 30–40 million years ago (Schneider et al., 2003; Aristide et al., 2015), substantially deeper in the primate family tree than great apes. They live in large multi-male, multi-female groups with frequent food sharing among related females, and have been shown to coordinate behaviour in cooperative tasks despite their phylogenetic distance from the Hominidae (Vale et al., 2019). Their inclusion in our analyses alongside great apes allows the analysis to span a much wider phylogenetic range than is typical in the comparative fairness literature.

We chose this dataset for two reasons. First, the phylogenetic distance from humans makes squirrel monkeys an unusually informative comparison: any preference structure detectable here would suggest a deep evolutionary origin or a parallel (convergent) evolution, rather than

a great-ape-specific feature. Second, the experimental design is the most analytically rich of the three; each pair plays three different games with different payoff structures, providing within-individual variation in the Kantian payoff differential $\Delta K = K(\text{Coop}) - K(\text{Defect})$ (1 in the Prisoners Dilemma; 2 in the Stag Hunt and Hawk-Dove). This cross-game variation, available for the same individuals, is rare in non-human game theory experiments and provides the analytical leverage exploited in our estimation strategy. Each pair played three games across multiple sessions: Prisoner's Dilemma (1,200 trials), Stag Hunt (1,500 trials), and Hawk-Dove (1,500 trials) with one pair (RamsesPeeWee) Prisoner's Dilemma data. We isolate simultaneous trials (firstplay = "Same", n = 1,028) for the baseline and include all trials in robustness analyses.

### 3.5 Identification

Identifying ($\kappa$, $\alpha$, $\beta$) requires three sources of variation: payoff variation that moves K(s) independently of the distributional gap ($x - y$); action variation that lets each parameter generate distinct predictions; and, where available, belief variation that shifts the strategic environment without altering K(s). The three datasets differ in what they offer.

**Payoffs.** Dataset 3 provides the variation needed: the same individuals play three games with different Kantian differentials ($\Delta K = 1, 2, 2$ for PD, Stag Hunt, and Hawk-Dove), and $\Delta K$ moves independently of $(x - y)$ across games. Datasets 1 and 2 do not. Dataset 1 offers only three fixed allocations, (5,1), (3,3), (1,5), in which K(s), own payoff, and the distributional gap all move together. In Dataset 2, $K(o) = 8 - o$ moves perfectly co-linearly with own payoff ($8 - o$), so K(s) cannot be shifted independently of x.

**Actions.** All three datasets show substantial within-individual choice variation, necessary for any likelihood-based estimation. This is sufficient in Dataset 3 because action variation interacts with the cross-game payoff variation, but insufficient in Datasets 1 and 2: in those designs, shifts in choice can be rationalised equally well by changes in $\kappa$ or in $\alpha$, because both parameters predict the same direction of behavioural change.

**Beliefs.** Dataset 1 includes a within-subject leverage manipulation that shifts beliefs about the responder's response without altering K(s). This provides a qualitative diagnostic; stability of equal-split behaviour across leverage conditions is more consistent with Kantian motivation than with strategic fear of rejection, but does not deliver structural separation of $\kappa$ from $\alpha$. Datasets 2 and 3 have no belief manipulation; Dataset 3 compensates through its payoff variation.

**Consequence.** Dataset 3 identifies $\kappa$ separately from $\alpha$ and $\beta$. Datasets 1 and 2 do not: $\kappa$ and $\alpha$ both predict shifts toward the equal split (Dataset 1) or higher offers (Dataset 2), so pooled LR

tests in those datasets fail to reject $\kappa = 0$ owing to collinearity, not absence of Kantian-consistent preferences. We make this non-identification explicit through $\kappa$ profiling exercises in Appendix E, which show flat likelihood profiles across $\kappa \in [0, 1]$ for both datasets.

## 4. Estimation

### 4.1 Individual-Level Maximum Likelihood

For each individual i, we estimate the parameter vector $\theta_i = (\kappa_i, \alpha_i, \beta_i, \lambda_i)$ by maximising the log-likelihood of observed choices. The choice probability follows a logistic (softmax) rule. We restrict the parameter space to $\kappa \in [0,1]$, $\alpha \geq 0$, $\beta \geq 0$, and $\lambda \in [0.01, 20]$. Optimisation uses L-BFGS-B with 20 random restarts. Standard errors are computed via the inverse numerical Hessian. For Dataset 2, we use a two-stage procedure: Stage 1 fixes $\lambda = 1.0$ and estimates ($\kappa$, $\alpha$, $\beta$); Stage 2 uses Stage 1 as a warm start for the full four-parameter model.

### 4.2 Pooled Estimation and Likelihood Ratio Tests

We pool all trials within each dataset and estimate a single representative-agent parameter vector on the pooled sample, providing more statistical power for three nested LR tests: $H_1$: $\kappa = 0$ (pure inequity aversion); $H_2$: $\alpha = \beta = 0$ (pure Kantian plus self-interest); $H_3$: $\kappa = 0.5$ (Alger-Weibull ESS). We use a 5% threshold for rejection.

### 4.3 Model Comparison

For Dataset 3, where identification is cleanest, we additionally compare four models by BIC: (a) Full Kantian ($\kappa$, $\alpha$, $\beta$ free); (b) Fehr-Schmidt ($\kappa = 0$, $\alpha$, $\beta$ free); (c) Pure Kantian ($\kappa$ free, $\alpha = \beta = 0$); (d) Heuristic bias ($\kappa = \alpha = \beta = 0$, $\lambda$ only). The BIC comparison directly addresses whether the Kantian term adds explanatory power beyond standard distributional preferences. We report both the Akaike Information Criterion ($AIC = -2 \cdot LL + 2k$) and the Bayesian Information Criterion ($BIC = -2 \cdot LL + k \cdot \ln n$). The two criteria differ in how heavily they penalise model complexity, BIC's penalty grows with sample size while AIC's is constant, so agreement between them provides a more robust basis for model selection than either alone. Results are reported in Section 5.3 and Table A4.

## 5. Results

### 5.1 Dataset 1: Chimps and Bonobos (Bilateral UG)

Table 2 reports individual-level estimates for the bilateral UG. All nine subjects converged, and all have $\kappa > 0.79$. The mean $\bar{\kappa} = 0.943$ is well above both the ESS value of 0.5 and the zero-Kantian baseline, but this figure should not be interpreted as evidence of high $\kappa$: as Appendix E demonstrates, the likelihood profile for DS1 is flat across the full range $\kappa \in [0, 1]$, meaning any value is equally consistent with the data. The point estimates reflect optimizer tendency toward corner solutions under collinearity, not identification. Three individuals hit the upper boundary at $\kappa = 1.0$ (Frodo, Lome, Tai); we report these for completeness.

| | **Lobo** | **Sandra** | **Frodo** | **Lome** | **Tai** | **Dorien** | **Kuno** | **Luiza** | **Yasa** |
|---|---|---|---|---|---|---|---|---|---|
| | *Chimp* | *Chimp* | *Chimp* | *Bonobo* | *Chimp* | *Chimp* | *Bonobo* | *Bonobo* | *Bonobo* |
| **κ** | 0.985 | 0.792 | 1.000† | 1.000† | 1.000† | 0.985 | 0.843 | 0.944 | 0.939 |
| | (0.245) | — | — | — | — | — | (4.215) | — | (2.439) |
| | [0.000] | — | — | — | — | — | [0.842] | — | [0.700] |
| **α** | 7.256 | 0.470 | 0.000 | 0.000 | 0.866 | 4.629 | 1.697 | 3.221 | 0.907 |
| | (115.317) | — | — | — | — | — | (45.648) | — | (59.123) |
| | [0.950] | — | — | — | — | — | [0.970] | — | [0.988] |
| **β** | 0.899 | 0.881 | 0.451 | 0.000 | 1.755 | 0.006 | 0.547 | 2.036 | 1.767 |
| | (49.834) | — | — | — | — | — | (47.379) | — | (89.771) |
| | [0.986] | — | — | — | — | — | [0.991] | — | [0.984] |
| **λ** | 0.881 | 0.830 | 0.469 | 0.973 | 0.508 | 0.636 | 1.155 | 1.227 | 1.017 |
| | — | — | — | — | — | — | — | — | — |
| ***Log-likelihood*** | *−58.29* | *−79.00* | *−76.05* | *−24.11* | *−37.05* | *−34.81* | *−50.66* | *−41.48* | *−43.54* |
| ***N trials*** | *128* | *192* | *128* | *64* | *64* | *64* | *127* | *128* | *128* |

*Table 2. Individual-level MLE estimates, bilateral ultimatum game (Dataset 1, chimps and bonobos). Each column reports estimates for one individual (5 chimpanzees, 4 bonobos). Point estimates appear in the top row of each parameter block; standard errors via the inverse numerical Hessian in parentheses; two-sided p-values for the null parameter = 0 in square brackets. † indicates a boundary estimate ($\kappa \geq 0.995$); standard errors are not well defined and these point estimates should be interpreted as $\kappa \geq 0.85$. $\lambda$ has no p-value row because the null $\lambda = 0$ is on the boundary of the parameter space. Optimisation uses L-BFGS-B with 20 random restarts. The mean $\bar{\kappa}$ across well-identified interior estimates (Lobo, Sandra, Dorien, Kuno, Luiza, Yasa) is 0.916. The wide standard errors and high p-values for many $\alpha$ and $\beta$ estimates reflect the $\kappa$–$\alpha$ collinearity diagnosed in Section 3.5: the flat likelihood surface produces near-flat Hessians and inflated standard errors.*

*Figure 1, Equal-split choice rates by leverage condition and species (Chimps vs Bonobos; DS1, Bilateral Ultimatum Game). Control condition is shown for descriptive reference only and is excluded from the main estimation.*

## 5.2 Dataset 2: Chimpanzees (Triadic UG)

Table 3 reports estimates for the triadic competitive UG. All five individuals converged using the two-stage procedure, and all have $\kappa > 0.619$. Mean $\bar{\kappa} = 0.897$, but as with DS1 this figure should not be read as evidence of high $\kappa$: the profile likelihood for DS2 is flat across the full $\kappa$ range (Appendix E), confirming non-identification. As in DS1, three individuals (Azibo, Riet, Frodo) have $\alpha$ at or near the upper boundary of 10, a further symptom of $\kappa$–$\alpha$ collinearity.

| | **Tai** | **Sandra** | **Azibo** | **Riet** | **Frodo** |
|---|---|---|---|---|---|
| | *Chimp* | *Chimp* | *Chimp* | *Chimp* | *Chimp* |
| **κ** | 1.000† | 0.619 | 0.930 | 0.986 | 0.949 |
| | — | (3.023) | — | (0.043) | (0.210) |
| | — | [0.838] | — | [0.000] | [0.000] |
| **α** | 0.000 | 1.855 | 9.992 | 9.994 | 10.000 |
| | — | (11.712) | — | (21.495) | (41.505) |
| | — | [0.874] | — | [0.642] | [0.810] |
| **β** | 0.000 | 1.389 | 0.020 | 0.001 | 0.000 |
| | — | (13.344) | — | (15.009) | (0.000) |
| | — | [0.917] | — | [1.000] | [0.000] |
| **λ** | 0.322 | 0.510 | 0.066 | 0.217 | 0.319 |
| | — | — | — | — | — |
| ***Log-likelihood*** | *−367.18* | *−359.68* | *−261.10* | *−383.33* | *−407.41* |
| ***N trials*** | *192* | *192* | *119* | *183* | *192* |

*Table 3. Individual-level MLE estimates, triadic competitive ultimatum game (Dataset 2, chimpanzees). Each column reports estimates for one individual. Point estimates in the top row, standard errors in parentheses, two-sided p-values in square brackets. † indicates a boundary estimate ($\kappa \geq 0.995$). Three of the five individuals (Tai, Sandra, Frodo) also appear in Dataset 1, enabling within-individual qualitative comparison across game formats. Estimation uses a two-stage procedure: Stage 1 fixes $\lambda = 1.0$ and estimates ($\kappa$, $\alpha$, $\beta$); Stage 2 uses Stage 1 as a warm start for the full four-parameter model. Mean $\bar{\kappa} = 0.897$. As in Table 2, wide standard errors and high p-values reflect $\kappa$–$\alpha$ collinearity in this design (see Section 3.5 and the $\kappa$ profile likelihood plot in Appendix E).*

The pooled LR test for $H_2$ ($\alpha = \beta = 0$) strongly rejects (LR = 1,975.1, $p < 0.001$). The tests for $H_1$ ($\kappa = 0$) and $H_3$ ($\kappa = 0.5$) both fail to reject, reflecting $\kappa$–$\alpha$ collinearity. The $\kappa$ profiling exercise for DS2 (Appendix E) confirms that the likelihood surface is flat across the full range of $\kappa$, making $\kappa$ non-identified in this design independently of the pooled LR test result.

### 5.3 Dataset 3: Squirrel Monkeys (Multi-Game), Identification Centrepiece

Table 4 reports estimates for the squirrel monkey multi-game dataset. Individual estimates are more heterogeneous than in the chimpanzee datasets, with κ ranging from 0.000 (SirzeeMidge_sub1) to 1.000 (FiftyDrew_sub1). Mean $\bar{\kappa}$ = 0.593 across all ten individuals. We flag five individuals with λ ≤ 0.016 as potentially unreliable.

| | **MulanBug_1** | **MulanBug_2** | **ArielOlivia_1** | **ArielOlivia_2** | **FiftyDrew_1** | **FiftyDrew_2** | **SirzeeMidge_1** | **SirzeeMidge_2** | **RamsesPW_1** | **RamsesPW_2** |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Sq.Mky* | *Sq.Mky* | *Sq.Mky* | *Sq.Mky* | *Sq.Mky* | *Sq.Mky* | *Sq.Mky* | *Sq.Mky* | *Sq.Mky* | *Sq.Mky* |
| **κ** | 0.112 | 0.923 | 0.411 | 0.724 | 1.000† | 0.432 | 0.000 | 0.684 | 0.822 | 0.822 |
| | (15.818) | (0.164) | (0.404) | (0.223) | — | (3.975) | — | (0.164) | (1.503) | (1.519) |
| | [0.994] | [0.000] | [0.308] | [0.001] | — | [0.913] | — | [0.000] | [0.584] | [0.588] |
| **α** | 10.000 | 0.000 | 0.002 | 0.274 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | (30.957) | (0.000) | (0.970) | (1.237) | (0.000) | (0.000) | — | (0.001) | (0.000) | (0.000) |
| | [0.747] | [1.000] | [0.999] | [0.824] | [1.000] | [1.000] | — | [1.000] | [1.000] | [1.000] |
| **β** | 2.567 | 10.000 | 0.000 | 0.000 | 0.000 | 10.000 | 0.000 | 0.000 | 10.000 | 10.000 |
| | (19.368) | (23.762) | (0.000) | (0.000) | (0.000) | (56.199) | — | (0.001) | (41.395) | (39.678) |
| | [0.895] | [0.674] | [1.000] | [1.000] | [1.000] | [0.859] | — | [1.000] | [0.809] | [0.801] |
| **λ** | 0.010* | 0.294 | 0.160 | 0.288 | 0.010* | 0.016* | 0.161 | 0.460 | 0.010* | 0.010* |
| | — | — | — | — | — | — | — | — | — | — |
| ***Log-likelihood*** | *−147.10* | *−144.13* | *−192.99* | *−188.70* | *−134.53* | *−133.96* | *−121.44* | *−116.05* | *−113.77* | *−113.77* |
| ***N trials*** | *213* | *213* | *280* | *280* | *194* | *194* | *177* | *177* | *164* | *164* |

*Table 4. Individual-level MLE estimates, multi-game social dilemmas (Dataset 3, squirrel monkeys). Each column reports estimates for one individual across the Prisoner's Dilemma, Stag Hunt, and Hawk-Dove games. Point estimates in the top row, standard errors in parentheses, two-sided p-values in square brackets. † indicates a boundary estimate (κ ≥ 0.995). * indicates a low-precision flag (λ ≤ 0.016): MulanBug_1, FiftyDrew_1, FiftyDrew_2, RamsesPW_1, and RamsesPW_2. These individuals are reported but excluded from primary inference because near-zero λ implies a near-flat likelihood surface and unreliable point estimates. RamsesPeeWee_1 and RamsesPeeWee_2 produce near-identical estimates because this pair has no Prisoner's Dilemma data, leaving very limited cross-game variation in ΔK. Mean $\bar{\kappa}$ across all 10 individuals = 0.593; mean across the 5 credible-precision individuals (MulanBug_2, ArielOlivia_1, ArielOlivia_2, SirzeeMidge_1, SirzeeMidge_2) = 0.548. Pooled likelihood ratio test for $H_1$ (κ = 0): LR = 3.754, p = 0.053. Pooled LR test for $H_3$ (κ = 0.5): LR = 0.965, p = 0.326. Two credible-precision individuals individually reject κ = 0 at the 5% level: SirzeeMidge_2 (p = 0.001) and MulanBug_2 (p = 0.015).*

Among the five credible-precision individuals (MulanBug_sub2, ArielOlivia_sub1, ArielOlivia_sub2, SirzeeMidge_sub1, SirzeeMidge_sub2), κ ranges from 0.000 to 0.923, and the mean is 0.549, remarkably close to the Alger-Weibull ESS prediction of 0.5. The pooled LR test for $H_3$ (κ = 0.5) fails to reject (LR = 0.965, p = 0.326). The test for $H_1$ (κ = 0) is marginally

insignificant, failing to reach the conventional threshold (LR = 3.754, p = 0.053). Notably, two credible-precision individuals individually reject κ = 0 at the 5% level: SirzeeMidge_sub2 (LR = 11.27, p = 0.001) and MulanBug_sub2 (LR = 5.92, p = 0.015), with ArielOlivia_sub2 approaching significance (LR = 5.22, p = 0.022). We interpret the result as suggestive of positive κ in squirrel monkeys rather than conclusive.

Reduced-form evidence. Within-pair cooperation rates increase monotonically from ΔK = 1 (Prisoner's Dilemma) to ΔK = 2 (Stag Hunt and Hawk-Dove), consistent with the Kantian payoff structure predicting behaviour independently of the utility model. The within-pair regression of cooperation rate on ΔK yields a positive slope (β = 0.027, SE = 0.021, t = 1.28), directionally consistent with the Kantian payoff structure but statistically underpowered (limited by sample size) given the small number of pair-game observations (n = 14). We treat this as corroborating rather than independently conclusive evidence.

Model comparison. Table A4 reports BIC scores for four models fit to DS3. Among the four specifications, the heuristic model (λ only, BIC = 2,857.7) achieves the lowest raw BIC, followed by the Pure Kantian model (κ free, α = β = 0, BIC = 2,861.0), then the Fehr-Schmidt model (BIC = 2,872.4), and finally the Full Kantian model (BIC = 2,876.3). The heuristic model's advantage reflects BIC's parsimony penalty: it uses only one parameter. However, the heuristic model cannot account for cross-game variation in cooperation rates; it predicts the same behaviour in all three games, while the Pure Kantian model does so with only one additional parameter (κ). Among models that can explain why cooperation differs across games, the Pure Kantian model fits best. The data do not require distributional parameters (α, β) once the Kantian payoff structure is included. This conclusion is reinforced by the Akaike Information Criterion: the AIC also selects the Pure Kantian model (AIC = 2,849.8) as best-fitting, and under the AIC the heuristic model falls to third place (AIC = 2,852.1), behind both the Pure Kantian and the Full Kantian specifications. The heuristic's competitive raw BIC therefore reflects parsimony rather than explanatory power; the convergence of both information criteria on the Pure Kantian model is the strongest model-selection evidence the data can provide.

*Figure 2, Squirrel monkey cooperation rates by game, plotted against the Kantian differential ΔK. Each coloured line is one pair; the thick dashed line is the population mean. Stag Hunt and Hawk-Dove are both plotted at ΔK = 2 and may overlap. RamsesPeeWee, which has no Prisoner's Dilemma data, is omitted from the ΔK = 1 endpoint.*

## 5.4 Cross-Species Comparison

Table 5 summarises the cross-species comparison of κ for descriptive purposes. We caution strongly against interpreting these figures comparatively: the great ape $\bar{\kappa}$ values (0.943 and 0.897) are optimizer outputs from flat likelihood surfaces and carry no evidential weight about

the true κ in those species. The squirrel monkey $\bar{\kappa} = 0.593$ is identified, but even this estimate is based on 5 credible-precision individuals with a pooled test that falls marginally short of conventional significance (p = 0.053). The apparent monotone pattern across taxa should not be read as an identified gradient; it is a descriptive juxtaposition of one identified estimate and two non-identified optimizer artifacts.

| Species / Dataset | $\bar{\kappa}$ | SD | n | Boundary | $H_3$: κ=0.5 | $H_1$: κ=0 |
|---|---|---|---|---|---|---|
| Chimps & Bonobos (bilateral UG) | 0.943 | 0.063 | 9 | 3 | n.r. | n.r. |
| Chimpanzees (triadic UG) | 0.897 | 0.148 | 5 | 1 | n.r. | n.r. |
| Squirrel Monkeys (multi-game) | 0.593 | 0.328 | 10 | 2 | n.r. | n.s.* |

*Table 5. Cross-species κ comparison. n.r. = not rejected. $H_1$ and $H_3$ non-rejections for great ape datasets reflect κ–α collinearity in those designs, not absence of Kantian-consistent preferences. * p = 0.053; fails to reach the 5% threshold but approaches conventional significance. See Section 5.3.*

## 5.5 Finite Mixture Model

Table 6 reports the BIC comparison across J = 1, 2, 3 latent types. The BIC selects J = 1 for DS1 and DS3, and J = 2 for DS2. Table 7 reports parameter estimates for the BIC-preferred model in each case.

| Dataset | J | Log-likelihood | BIC | N obs | N types |
|---|---|---|---|---|---|
| DS1_bilateral | 1 | −466.55 | 960.82 | 1,023 | 1 |
| DS1_bilateral | 2 | −457.64 | 977.65 | 1,023 | 2 |
| DS1_bilateral | 3 | −457.64 | 1,012.30 | 1,023 | 3 |
| DS2_triadic | 1 | −1,834.69 | 3,696.48 | 878 | 1 |
| DS2_triadic | 2 | −1,806.18 | 3,673.36* | 878 | 2 |
| DS2_triadic | 3 | −1,791.56 | 3,678.01 | 878 | 3 |
| DS3_multgame | 1 | −1,422.89 | 2,876.29* | 2,056 | 1 |
| DS3_multgame | 2 | −1,418.14 | 2,904.93 | 2,056 | 2 |
| DS3_multgame | 3 | −1,418.14 | 2,943.07 | 2,056 | 3 |

*Table 6. BIC comparison across J = 1, 2, 3 types by dataset. * marks the BIC-preferred J.*

| Dataset | J | Type | π | κ | α | β | λ | Preferred J |
|---|---|---|---|---|---|---|---|---|
| DS2_triadic | 2 | 1 | 0.80 | 0.984 | 0 | 9.702 | 0.152 | 2 |
| DS2_triadic | 2 | 2 | 0.20 | 0.439 | 1.253 | 0.915 | 0.518 | 2 |
| DS3_multgame | 1 | 1 | 1.00 | 0.780 | 0 | 0 | 0.066 | 1 |
| DS1_bilateral | 1 | 1 | 1.00 | 0.992 | 9.985 | 0.003 | 0.866 | 1 |

*Table 7. Parameter estimates for the BIC-preferred mixture model.*

For DS3 (squirrel monkeys), a single type with κ = 0.780, α = 0.000, and β = 0.000 describes the population, a single Kantian type with no inequity aversion, consistent with both the individual-level finding that most squirrel monkey estimates have α ≈ 0 and with the model comparison result that the Pure Kantian specification fits best. For DS2 (triadic chimpanzees), the BIC prefers two types: Type 1 (80%, κ = 0.984, β = 9.702) and Type 2 (20%, κ = 0.439, α = 1.253, β = 0.915), the latter resembling the mixed type in Van Leeuwen and Alger (2024). For DS1 (bilateral UG), a single high-κ, high-α type describes the population, consistent with the κ–α collinearity.

*Figure 3, Individual κ̂ estimates with 95% confidence intervals, grouped by species. Four panels: Bonobos (bilateral UG), Chimpanzees (bilateral UG), Chimpanzees (triadic UG), Squirrel Monkeys (social dilemmas). Dashed vertical line at κ = 0.5 marks the Alger-Weibull ESS. Red borders indicate boundary estimates.*

*Figure 4, κ̂ vs α̂ scatter plot by individual and species. Solid markers = credible-precision individuals (λ > 0.016); open/hollow markers = low-precision individuals (λ ≤ 0.016) excluded from primary inference. Dashed vertical line at κ = 0.5.*

*Figure 5, Mixture type positions in (κ, α) space. Three panels: DS1 bilateral, DS2 triadic, DS3 multgame. Bubble size proportional to population share π. Dashed vertical line at κ = 0.5.*

## 6. Discussion

### 6.1 What the Estimates Mean

The central result, that squirrel monkeys show positive estimated κ, consistent with the Alger-Weibull ESS prediction, requires careful interpretation. We do not claim that squirrel monkeys consciously reason in the Kantian sense. Rather, we use κ as a revealed preference parameter: it is identified from choice patterns across varying payoff environments, not from any introspective report. An animal with positive κ is one whose decisions are better predicted by a utility function that includes the Kantian payoff term K(s) than by one that ignores it. The

Kantian term K(s) provides additional predictive power over and above inequity aversion in DS3, where the multi-game design provides within-individual identification. In DS1 and DS2, K(s) and inequity aversion are collinear and κ is not identified; we do not treat the optimizer outputs from those datasets as comparative evidence.

The model comparison result for DS3 warrants careful discussion. The heuristic model (λ only) achieves the lowest raw BIC, reflecting its parsimony, but it cannot explain cross-game variation in cooperation rates. Instead, it predicts the same behaviour in all three games. Among specifications capable of accounting for game-level variation, the Pure Kantian model (κ free, α = β = 0) fits best, outperforming the Full Kantian and Fehr-Schmidt models. This indicates that squirrel monkey cooperation patterns are explained by the Kantian payoff structure without requiring distributional parameters, but we acknowledge that the data are too noisy to strongly discriminate between structured preference models.

Individual heterogeneity in κ. The squirrel monkey estimates show substantial individual variation in κ, from $\hat{\kappa} = 0.000$ (SirzeeMidge_sub1) to $\hat{\kappa} = 1.000$ (FiftyDrew_sub1), even within the credible-precision subsample. This heterogeneity mirrors the type-level variation Van Leeuwen and Alger (2024) document in human subjects and suggests that κ may be best understood as an individual-level preference parameter rather than a species-level constant. The drivers of this variation — dominance rank, kinship, prior experimental experience, age, or developmental trajectory — lie beyond the present data, but they are a natural target for future studies designed to identify the proximate sources of Kantian-consistent preference structure.

## 6.2 Relationship to the Primate Fairness Literature

A large literature documents inequity aversion in non-human primates, beginning with the cucumber-versus-grape experiment of Brosnan and de Waal (2003). Our results offer a structural decomposition that the prior literature could not attempt: separating distributional concern (α, β) from Kantian-consistent preference weighting (κ). For great apes, κ and α co-occur and cannot be separated in the available designs. For squirrel monkeys, the multi-game design achieves this separation, and the data are best fit by a model with positive κ and near-zero α, a qualitatively different preference profile from the high-α, high-κ pattern in great apes.

The Fehr-Schmidt model (κ = 0) with inequity aversion alone is not rejected by our pooled LR tests for DS1 and DS2, reflecting identification limits, not evidence against κ. In DS3, the data are more informative: the pooled result approaches significance against κ = 0 ($p = 0.053$), and

two individuals individually reject $\kappa = 0$. The squirrel monkey evidence provides the strongest inferential foundation for the paper's core claim.

### 6.3 Limitations

Several limitations qualify our results. First, sample sizes are small (9, 5, and 10 individuals). Second, the $\kappa$–$\alpha$ collinearity in DS1 and DS2 is fundamental to those designs, not a fixable estimation problem, new data with payoff variation that independently moves K(s) and $(x - y)$ would be required. Third, low $\lambda$ values in several DS3 individuals suggest considerable choice noise, which may reflect either random behaviour or model misspecification that the utility function does not capture. Fourth, the RamsesPeeWee pair's near-identical estimates ($\kappa = 0.822$ for both) reflect their missing Prisoner's Dilemma data; both are correctly excluded under the low-precision flag. Fifth, the cross-species pattern is purely descriptive. Sixth, and importantly, the $\Delta$K-cooperation relationship identified in DS3 is not uniquely explained by Kantian-consistent preferences. Alternative mechanisms, including risk dominance heuristics, payoff-symmetry preferences, or reinforcement learning effects that differ across game types, could generate a similar pattern. The within-individual design absorbs animal-level heterogeneity, and the three-game structure limits some alternatives, but a decisive identification argument would require a design in which K(s) and the leading alternative mechanisms predict opposite choices, a Kantian mirror paradigm in which K(Coop) is high while the own payoff from cooperation is low.

### 6.4 Implications and Future Directions

If Kantian-consistent preference structure is present in squirrel monkeys, a New World monkey lineage that diverged from the human lineage approximately 30–40 million years ago, then the preference architecture underlying universalizability reasoning may have ancient origins predating the specific cognitive and linguistic capacities of modern humans. Alternatively, a convergence of a similar preference architecture may have evolved multiple times under the selective pressure of social living. We treat this as a motivating hypothesis, not a conclusion, in initiating a research direction in the evolution of rules in preference and decision making. The monotone pattern in $\bar{\kappa}$ across taxa is consistent with Kantian-consistent preferences co-varying with social complexity, as the Alger-Weibull model predicts, but with three species, descriptive cross-species comparisons cannot establish this.

Future work should address the collinearity limitation via a Kantian mirror paradigm: a game in which K(s) increases while the proposer's own payoff decreases, so that high $\kappa$ and high $\alpha$ predict opposite choices. A second direction is to extend the estimation to both New and Old

World monkeys, such as macaques and capuchins, which are also highly social yet represent different clades in the primate phylogeny, and apply the multi-game design that provides clean identification in DS3.

This will bring up one of the most important issues in the emergence of preference structure: convergent versus homologous evolution. The monotone pattern of estimated $\bar{\kappa}$ across taxa is consistent with two distinct evolutionary hypotheses. Under a homology account, Kantian-consistent preferences descend from a common primate ancestor, and the cross-species variation reflects subsequent divergence in social ecology. Under a convergence account, the preference structure has evolved independently in multiple lineages under common selection pressures associated with social living, group cooperation, repeated interaction, kin recognition, and assortative matching. These hypotheses are not mutually exclusive: a preference architecture may be partially inherited and partially elaborated under lineage-specific selection.

Distinguishing the two would require extending the structural framework beyond the primate order. A convergence account predicts that Kantian-consistent preferences should also be detectable in other socially complex mammals in which assortative matching and repeated cooperation are present, for example cetaceans (Krützen et al., 2005), elephants (Plotnik and de Waal, 2014), or social canids (Range et al., 2019). A strict homology account predicts no such finding outside the primate lineage. The multi-game structural design we apply here is, in principle, portable to these species, though the experimental and welfare considerations differ substantially. If future work documents structural $\kappa > 0$ in socially complex non-primate mammals, this would constitute strong evidence that the Kantian preference architecture evolved convergently in response to the universal selection pressures of social living, rather than being a primate-specific inheritance. In our view this is the most important and promising open question raised by the present analysis: whether the Kantian decision rule is a feature of the primate lineage specifically, or a recurrent evolutionary solution to the problem of sustaining cooperation under assortative matching wherever social living imposes that problem.

## 7. Conclusion

This paper provides the first structural estimates of Kantian-consistent preferences in non-human primates, using three existing datasets with varying identification strength. The squirrel monkey multi-game dataset is the inferential centrepiece: within-individual variation in the Kantian payoff differential $\Delta K$ identifies $\kappa$ independently of inequity aversion. Among structured specifications that can account for cross-game variation in cooperation, the Pure

Kantian model ($\kappa = 0.780$, $\alpha = \beta = 0$) fits best by BIC; the simpler heuristic model achieves a lower overall BIC but predicts no game-level variation. The pooled test fails to reject $\kappa = 0$ ($p = 0.053$), approaching but not reaching conventional significance, and the estimate of $\bar{\kappa} = 0.593$ is statistically consistent with the Alger-Weibull ESS prediction of $\kappa = 0.5$. Two individual squirrel monkeys independently reject $\kappa = 0$ at the 5% level.

The great ape datasets show behaviour consistent with high $\kappa$, $\bar{\kappa} = 0.943$ in bonobos and chimpanzees (bilateral UG) and $\bar{\kappa} = 0.897$ in chimpanzees (triadic UG), but $\kappa$ is not separately identified from inequity aversion in those designs. The $\kappa$ profiling exercises (Appendix E) make this non-identification explicit: likelihood profiles are flat across the full $\kappa$ range. We treat the great ape results as consistency evidence for the Kantian-consistent preference structure identified in squirrel monkeys.

The same utility function, with the same Kantian payoff term K(s), that describes behaviour in human laboratory subjects and can be instilled in AI agents (Lu et al., 2025) also provides the best-fitting model in squirrel monkey data. This positions $\kappa$ as a portable behavioural parameter that may serve as a common currency for comparing moral behaviour across species, cultures, and artificial systems.

## References


Alger, I., & Weibull, J. W. (2013). Homo moralis — preference evolution under incomplete information and assortative matching. Econometrica, 81(6), 2269–2302. https://doi.org/10.3982/ECTA10637

Aristide, L., Dos Reis, S. F., Machado, A. C., Lima, I., Lopes, R. T., Perez, S. I., & Flores, D. A. (2015). Brain shape convergence in the convergent evolution of new world monkeys. Proceedings of the National Academy of Sciences, 112(9), 2782–2787. https://doi.org/10.1073/pnas.1422823112

Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. Nature, 425, 297–299. https://doi.org/10.1038/nature01963

de Waal, F. B. M. (2006). Primates and philosophers: How morality evolved. Princeton University Press.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. Quarterly Journal of Economics, 114(3), 817–868. https://doi.org/10.1162/003355399556151

Henrich, J. (2016). The secret of our success. Princeton University Press.

Kant, I. (1998). Groundwork of the metaphysics of morals (M. Gregor, Trans.). Cambridge University Press. (Original work published 1785)

Krützen, M., Mann, J., Heithaus, M. R., Connor, R. C., Bejder, L., & Sherwin, W. B. (2005). Cultural transmission of tool use in bottlenose dolphins. Proceedings of the National Academy of Sciences, 102(25), 8939–8943. https://doi.org/10.1073/pnas.0500232102

Lu, W., Dhanda, A., Chen, D. L., & Hansen, C. B. (2025). Aligning AI agents with economic preferences in strategic environments. arXiv preprint arXiv:2507.20796. https://arxiv.org/abs/2507.20796

Plotnik, J. M., & de Waal, F. B. M. (2014). Asian elephants (Elephas maximus) reassure others in distress. PeerJ, 2, e278. https://doi.org/10.7717/peerj.278

Range, F., Marshall-Pescini, S., Kratz, C., & Virányi, Z. (2019). Wolves lead and dogs follow, but they both cooperate with humans. Scientific Reports, 9, 3796. https://doi.org/10.1038/s41598-019-40468-y

Sánchez-Amaro, A., & Rossano, F. (2021). Chimpanzees and bonobos use social leverage in an ultimatum game. Proceedings of the Royal Society B, 288(1962), Article 20211937. https://doi.org/10.1098/rspb.2021.1937

Sánchez-Amaro, A., Maurits, L., & Haun, D. B. M. (2024). Chimpanzees engage in competitive altruism in a triadic ultimatum game. Scientific Reports, 14, Article 3393. https://doi.org/10.1038/s41598-024-53973-6

Schneider, H., Sampaio, I., Harada, M. L., Barroso, C. M. L., Schneider, M. P. C., Czepielewski, E., & Goodman, M. (2003). Molecular phylogeny of the New World monkeys. American Journal of Physical Anthropology, 121(4), 305–319. https://doi.org/10.1002/ajpa.10199

Trivers, R. L. (1971). The evolution of reciprocal altruism. Quarterly Review of Biology, 46(1), 35–57. https://doi.org/10.1086/406755

Vale, G. L., Williams, L. E., Schapiro, S. J., Lambeth, S. P., & Brosnan, S. F. (2019). Responses to economic games of cooperation and conflict in squirrel monkeys (Saimiri boliviensis). Animal Behavior and Cognition, 6(1), 32–47. https://doi.org/10.26451/abc.06.01.03.2019

Van Leeuwen, B., & Alger, I. (2024). Estimating social preferences and Kantian morality in strategic interactions. Journal of Political Economy Microeconomics, 2(4), 665–706. https://doi.org/10.1086/732125

## Appendix A: Derivation of K(s) for the Triadic UG

The full expected Kantian own payoff is $E[x \mid o_i = o_j = o] = 0.5 \times (8 - o) + 0.5 \times 8 = 8 - 0.5 \times o$. In the main text and estimation, we use the simplified expression $K(o) = 8 - o$, which corresponds to the payoff upon winning. This simplification rescales the slope of the Kantian incentive by a factor of two but preserves the strictly decreasing monotone relationship between offer and Kantian payoff. We confirmed that substituting the full expression $8 - 0.5 \times o$

in estimation leaves all individual κ estimates within ±0.03 of the reported values and does not change any qualitative result. Results available on request.

## Appendix B: Payoff Matrices (Dataset 3)

| Game / Outcome | (Coop, Coop) | (Coop, Defect) | (Defect, Defect) |
|---|---|---|---|
| Prisoner's Dilemma | (2, 2) | (0, 3) | (1, 1) |
| Stag Hunt | (4, 4) | (0, 2) | (2, 2) |
| Hawk-Dove | (2, 2) | (1, 3) | (0, 0) |

*Table A1. Payoff matrices for Dataset 3. K(Coop) is the (Coop, Coop) own payoff; K(Defect) is the (Defect, Defect) own payoff.*

## Appendix C: Robustness, DS2 Consecutive Trials

| Individual | κ_base | α_base | β_base | λ_base | bdry_base | κ_rob | α_rob | β_rob | λ_rob | bdry_rob |
|---|---|---|---|---|---|---|---|---|---|---|
| Tai | 1.000 | 0.000 | 0.000 | 0.322 | Yes | 1.000 | 0.958 | 5.683 | 0.314 | Yes |
| Sandra | 0.619 | 1.855 | 1.389 | 0.510 | No | 0.537 | 1.322 | 0.043 | 1.000 | No |
| Azibo | 0.930 | 9.992 | 0.020 | 0.066 | No | 0.962 | 10.000 | 0.000 | 0.010 | No |
| Riet | 0.986 | 9.994 | 0.001 | 0.217 | No | 1.000 | 10.000 | 0.000 | 0.177 | Yes |
| Frodo | 0.949 | 10.000 | 1.238 | 0.319 | No | 0.719 | 1.851 | 3.164 | 0.316 | No |

*Table A2. DS2 parameter estimates, baseline vs robustness sample (including consecutive trials, n = 1,318).*

Mean $\bar{\kappa}$ is 0.897 under the baseline and 0.844 in the robustness sample. Frodo's κ drops from 0.949 to 0.719 when consecutive trials are added, with α dropping from the boundary (10.000) to 1.851, suggesting that consecutive trials partially separate κ from α for this individual. The overall interpretation is unchanged.

## Appendix D: Reduced-Form Evidence, ΔK Predicts Cooperation (DS3)

As a model-light test of Kantian-consistent preference structure, we regress pair-level mean cooperation rates on the Kantian payoff differential ΔK using a within-pair (pair fixed effects) estimator. This approach requires no utility function: if ΔK predicts cooperation after absorbing pair-level heterogeneity, that is direct reduced-form evidence that the Kantian payoff structure shapes behaviour.

Formally, we demean cooperation rates and ΔK within each pair and estimate the OLS slope on the demeaned data. Standard errors use HC1 heteroskedasticity correction. Results are reported in Table A3 and Figure 6. The within-pair slope β(ΔK) is positive (β = 0.027, SE = 0.021, t = 1.28, p ≈ 0.23), consistent with cooperation increasing as the Kantian payoff advantage of the cooperative action rises, though underpowered to detect significance at conventional levels given the small number of pair-game observations (n = 14). This reduced-form result is directionally consistent with the structural κ estimates without requiring the utility model, but should be treated as corroborating rather than independently conclusive evidence.

| Estimator | β(ΔK) | SE (HC1) | t-stat | $R^2$ (within) | N pairs × games |
|---|---|---|---|---|---|
| Within-pair (pair FE) | 0.027 | 0.021 | 1.283 | 0.069 | 15 (5 pairs × 3 games) |

*Table A3. Reduced-form regression of cooperation rate on ΔK with pair fixed effects (DS3).*

*Figure 6, Reduced-form evidence: pair-level cooperation rates against ΔK with within-pair regression line. Each point is one pair × game observation; slope shows the within-pair effect of increasing the Kantian payoff advantage.*

## Appendix E: κ Profile Likelihood, DS1 and DS2

To make the non-identification of κ in the ape datasets transparent, we fix κ at eleven grid points from 0 to 1 and re-estimate α, β, and λ at each. The resulting profile log-likelihoods are plotted in Figure 8. If the profile is flat, as collinearity predicts, the figure directly shows that all κ values are observationally equivalent given the available payoff variation. This turns the identification weakness into an explicit, honest result.

Tables A5a and A5b report the full profile tables for DS1 and DS2, respectively. In both cases, the ΔLL profile is nearly flat across the range $\kappa \in [0, 1]$, with no value outside the 95% confidence interval threshold (ΔLL = −1.92). This confirms that any κ between 0 and 1 is consistent with the ape data, and that the high point estimates ($\bar{\kappa} \approx 0.94$ and 0.90) reflect the tendency of the optimiser to fit corner solutions rather than precise identification.

*Figure 7, κ profile likelihood for DS1 (bilateral UG, left) and DS2 (triadic UG, right). Each panel shows ΔLL relative to the maximum across the κ grid. The dashed horizontal line marks the 95% CI threshold (ΔLL = −1.92). Flat profiles confirm that κ is not identified in these designs.*

## Appendix F: Model Comparison, DS3 Squirrel Monkeys

Table A4 reports BIC scores for four models fit to the DS3 pooled baseline sample (n = 2,056 simultaneous trials). The heuristic model (λ only, 1 parameter) achieves the lowest raw BIC (2,857.7), followed by the Pure Kantian model (2 parameters, BIC = 2,861.0), the Fehr-Schmidt model (3 parameters, BIC = 2,872.4), and the Full Kantian model (4 parameters, BIC = 2,876.3). The heuristic model's raw BIC advantage reflects parsimony: it uses one fewer parameter than the Pure Kantian model. Critically, the heuristic model predicts the same cooperation rate in all games and therefore cannot explain the cross-game ΔK variation that motivates the structural approach. Among models that can explain game-level variation in cooperation, the Pure Kantian specification fits best.

| Model | k | Log-likelihood | AIC | ΔAIC | BIC | ΔBIC |
|---|---|---|---|---|---|---|
| Pure Kantian (κ free, α=0, β=0) | 2 | −1,422.89 | 2,849.78 | 0.00 | 2,861.03 | 0.00 |
| Heuristic / bias (κ=0, α=0, β=0) | 1 | −1,425.06 | 2,852.11 | +2.33 | 2,857.74 | −3.29 |
| Full Kantian (κ, α, β free) | 4 | −1,422.89 | 2,853.78 | +4.00 | 2,876.29 | +15.26 |
| Fehr-Schmidt (κ=0, α, β free) | 3 | −1,424.77 | 2,855.53 | +5.75 | 2,872.42 | +11.39 |

*Table A4. Model comparison for DS3 (squirrel monkeys, n = 2,056 trials). k = number of free parameters. AIC = −2·LL + 2k; BIC = −2·LL + k·ln(n). ΔAIC and ΔBIC are relative to the Pure Kantian model; lower values indicate better fit. Both criteria select the Pure Kantian model (κ free, α = β = 0) as best-fitting. The Heuristic model attains the lowest raw BIC because of its extreme parsimony (one parameter), but it ranks third under the AIC and cannot explain the cross-game variation in cooperation rates that the Kantian models capture; the convergence of both information criteria on the Pure Kantian model is the strongest model-selection evidence the data can provide.*

*Figure 8, BIC comparison across four models fit to DS3. Lower BIC indicates better fit. The Pure Kantian model (κ free, α=β=0) outperforms both the Full Kantian and Fehr-Schmidt specifications, indicating that the Kantian payoff structure explains squirrel monkey cooperation without requiring distributional parameters.*

## Appendix G: Robustness, Credible-λ Subsample (DS3)

Table A6 compares κ estimates for the five credible-precision squirrel monkeys (λ > 0.016) between the full sample (all 10 individuals) and the restricted credible-λ sample. Mean κ in the credible-λ subsample is 0.549, almost identical to the full-sample mean of 0.593, confirming that the exclusion of low-precision individuals does not materially change the point estimates. Figure 9 shows the comparison graphically.

*Figure 9, κ estimates for credible-precision squirrel monkeys (λ > 0.016). Grey bars show full-sample κ; coloured bars show credible-λ subsample κ. The dashed line marks the credible-λ mean (0.549); the dotted line marks the ESS prediction (0.5).*