Referee Report: Motivated Errors

From Exley and Kessler (2024)

Yiheng You

March 7, 2025

Contents

1 Questions First

2 The purpose of this paper

2.1 Definitions

2.1.1 Motivated Errors:

Individuals are keen to view themselves in a more favorable light. They might appeal to the possibility of being confused or having made a mistake to help justify their behavior. **People may make errors to justify their behavior** in a wide variety of contexts.

Appealing to the possibility that make a mistake ⇒ small possibility of taking a selfish action ⇒ They make a mistake due to confusion ⇒ choosing a selfish action cannot be viewed as a definitive desire to be selfish, since there is always some probability mass on the selfish action chosen due to confusion ⇒ an agent who wants to avoid **image costs** of being selfish can then **"mimic"** "confused type" (an excuse for selfish behavior)

acting confusion can facilitate selfish behavior ⇒ more evidence of supposed confusion. So what is confusion type? how to measure them? Says miscalculate when filing taxes.

2.1.2 Correlation Neglect

To rationalize favorable views of a preferred political candidate, they could display more correlation neglect when correlated reports about the candidate are positive but less correlation neglect when reports about the candidate are negative.

2.1.3 Anchoring Bias

When they want an excuse to drive rather than walk, they could display more of an anchoring bias when it results in them calculating a longer distance to their destination.

To justify an indulgent purchase to themselves, they could miscalculate the post-tax price even though they calculate tax correctly for goods they are less eager to buy.

After a purchase, they could conveniently think that they received the correct amount of change if the cashier provides too much but identify a shortfall if the cashier provides too little.

They could justify selfishness by making an error when multiplying a restaurant bill by a particular fraction, leading them to leave a smaller tip for the staff.

2.2 Three Studies of motivated errors

2.2.1 Adding Study

In the Adding Study, subjects make a series of decisions, choosing between receiving a fixed amount of money for themselves and a sum of amounts for charity. When a zero is added to the sum for charity, subjects become less likely to choose the sum for charity. However, when selfish motives are removed—when subjects instead choose between two payoffs that both benefit charity—decisions are no longer influenced by the addition of a zero. Subjects only act as if they cannot add a zero when doing so can help justify selfish decisions. Further treatments of our Adding Study show that decreasing the scope for confusion—by providing subjects with more information about the sum going to charity—reduces motivated errors. Motivated errors are cut in half when subjects can click a box to reveal the sum of payoffs or when they are shown the sum by default. Motivated errors are fully eliminated when the sum is shown by default and subjects must correctly report it back before making their choice.

Methodology
Study 1: Adding Study
Objective: Test whether subjects exploit arithmetic errors to justify selfish choices.
Design:
- Participants: 1,769 MTurk workers.
- Conditions:
- Self/Charity. Choose between money for self (X cents) and a sum for charity.
- Charity/Charity. Choose between two charity sums.
- Manipulation: Bundles contained 4 or 5 numbers (e.g., [51, 51, 51, 0] vs. [51, 51, 51]). Adding a zero lowered perceived charity value without changing the actual sum.
- Follow-Up Treatments:
- Sum Optional: Subjects could click to reveal the total.
- Sum Shown: Total displayed by default.
- Sum Unavoidable: Subjects must correctly report the total before choosing.

Key Measures:
- Rate of choosing charity bundles with vs. without added zeros.
- Accuracy in a supplemental Calculation Study (98% correct).

2.2.2 Correlation Neglect Study

In the Correlation Neglect Study, subjects are asked to make predictions about the average of correlated information. When they face selfish motives to provide certain predictions, they may justify selfishness by appealing to the possibility of being confused when making these predictions. We observe results consistent with such motivated errors. Indeed, we find that the opportunity to make motivated errors can both exacerbate and mitigate evidence for correlation neglect.

Study 2: Correlation Neglect Study
Objective: Examine how selfish motives influence neglect of signal correlations.

Design:
- Participants: 1,200 MTurk workers.
- Task: Predict the average of four estimates (Estimate 1-4) after observing correlated signals:
- Low Estimate 1: Initial estimate was the smallest (e.g., 3), leading to systematic underestimation.
- High Estimate 1: Initial estimate was the largest (e.g., 93), leading to overestimation.
- Conditions:
- Control: Correct answers rewarded with X (self) or 150 cents (charity).
- Underestimate: Underpredicting rewarded with X cents.
- Overestimate: Overpredicting rewarded with X cents.

Key Measures:
- Deviation of predictions from the true average.
- Comparison of bias magnitude across incentive-aligned vs. misaligned conditions.

2.2.3 Anchoring Studies

In the Anchoring Studies, subjects are provided with an uninformative anchor and asked knowledge-based questions. Akin to what we find in the Correlation Neglect Study, we observe results consistent with motivated errors that can both exacerbate and mitigate evidence for an anchoring bias.

Study 3: Anchoring Studies
Objective: Test anchoring bias under selfish incentives.
Design:
- Participants: 1,195 (Study A) and 1,192 (Study B) MTurk workers.
- Tasks: Answer trivia questions (e.g., "How long does light take to reach Jupiter?") after exposure to random anchors (20 or 80).
- Conditions:
- Anchoring Study A: Slider defaulted to anchor; 15-second time limit.
- Anchoring Study B: No default anchor; self-paced.
- Reward Structures:
- Control: Correct answers rewarded with X/150 cents.
- Underestimate/Overestimate: Aligned payoffs with anchoring direction.

Key Measures:
- Magnitude of anchoring bias (difference between high/low anchor responses).

3 Results

3.1 Adding Study

- Self/Charity Condition: Adding a zero reduced charity choices by 6% ($p < 0.01$). - Charity/Charity Condition: No effect ($\Delta = 0\%, \mathrm{p} > 0.1$). - Mitigation: - Sum Optional: Effect halved to 3% ($\mathrm{p} < 0.05$). - Sum Unavoidable: Effect eliminated ($p > 0.1$).

3.2 Correlation Neglect Study

- Control Group: Underestimated Iow Estimate 1 questions by 19.5 points ($\mathrm{p} < 0.01$). - Motivated Modulation: - Underestimate Group: Underestimation amplified to 24.8 points ($p < 0.01$). - Overestimate Group: Underestimation reduced to 7.7 points ($p < 0.01$).

3.3 Anchoring Studies

- Control Group: Anchoring bias ranged from 14-20 points ($p < 0.01$). - Motivated Modulation: - Incentive-aligned conditions amplified bias ($\Delta = +6 - 8$ points, $\mathrm{p} < 0.01$). - Incentive-misaligned conditions reduced bias by 50-65% ($p < 0.01$).

4 Discussion

4.1 Theoretical Implications:

1. Motivated Errors as Strategic Tools: Errors are not passive mistakes but active justifications. Subjects exploited ambiguity to license selfishness, consistent with self-signaling models (Bénabou & Tirole, 2011). 2. Bias Plasticity: Cognitive biases (correlation neglect, anchoring) are modulated by incentives, challenging the view of biases as fixed irrationalities.

4.2 Practical Implications:

- Policy Design: Reducing ambiguity (e.g., auto-filled tax forms) and enforcing engagement (e.g., mandatory confirmations) can curb self-serving errors. - Charity Nudges: Transparent presentation of donation impacts reduces excuse-driven avoidance (Andreoni et al., 2017).

4.3 Limitations:

- External Validity: MTurk samples may not generalize to real-world high-stakes decisions. - Mechanisms: The paper focuses on whether errors occur, not why (e.g., self-deception vs. social image concerns).

5 Homework

Background and Significance of the Allais Paradox in Behavioral Economics The Allais Paradox reveals systematic deviations from Expected Utility Theory (EUT) in people's risk decision-making, serving as a key starting point for behavioral economics. Specifically:

- Problem 1: Most people choose a certain gain (Option A) over a high-risk, high-reward option (Option B).

- Problem 2: Most people choose a high-risk, high-reward option (Option D) over a low-risk, Iow-reward option (Option C). This contradicts the independence axiom of EUT but can be explained by Prospect Theory.

Core Mechanisms of Prospect Theory

1. Value Function

Formulated as $v(x) = x^{\alpha}$ (for gains) or $v(x) = -\lambda(-x)^{\alpha}$ (for losses), reflecting "loss aversion" ($\lambda > 1$) and "diminishing sensitivity to gains" ($\alpha < 1$).

2. Probability Weighting Function

Formulated as $w(p) = \frac{p^r}{(p^r+(1-p)^r)^{1/r}}$, where $r = 0.75$, used to adjust subjective probabilities and overweight small probabilities.

Numerical Analysis of the Allais Paradox Let $\gamma = 0.75, \alpha = 0.8, \lambda = 2$

Problem 1: Certain Gain vs. Risky Option

- Option A (Certain gain of 1 million):

Value $v(1) = 1^{0.8} = 1$, weight $\pi(1) = w(1) = 1$. Prospect value: $1 \times 1 = 1$.

- Option B ($\mathbf{10}\%$ chance of 5 million, 89% chance of 1 million, 1% chance of 0) :

Value: $v(5) = 5^{0.8} \approx 3.623, v(1) = 1, v(0) = 0$.

Weight calculation:

- $\pi(5M) = w(0.10) \approx 0.1563$,
- $\pi(1M) = w(0.99) - w(0.10) \approx 0.8056$,
- $\pi(0) = 1 - w(0.99) \approx 0.0381$.

Prospect value: $3.623 \times 0.1563 + 1 \times 0.8056 \approx 1.372$.

Contradiction: Theory predicts choosing B ($1.372 > 1$), but in reality, people choose A.

Explanation: Prospect Theory underestimates the "certainty effect" (overweighting of certain gains), requiring parameter adjustments (e.g., reducing γ to enhance small probability overweighting).

Problem 2: Comparison of Risky Options

- Option C ($\mathbf{11}\%$ chance of 1 million, 89% chance of 0) : Weight $\pi(0.11) = w(0.11) \approx 0.1667$, prospect value: $1 \times 0.1667 \approx 0.167$.

- Option D ($\mathbf{10}\%$ chance of $\mathbf{5}$ million, $\mathbf{90}\%$ chance of $\mathbf{0}$): Weight $\pi(0.10) \approx 0.1563$, prospect value: $3.623 \times 0.1563 \approx 0.566$.

Consistency: Theory predicts choosing $\mathrm{D}(0.566 > 0.167)$, aligning with reality. Reason: Both options are risky, and the certainty effect is not involved; probability weighting dominates the decision.

Limitations of Parameters and Behavioral Extensions

1. Parameter Adjustment: Reducing γ (e.g., to 0.6) can enhance small probability overweighting, making the theoretical prediction for Problem 1 closer to reality.

2. Editing Phase: Additional modeling of the "framing effect" (e.g., categorizing certain gains separately) is needed.

3. Loss Aversion: The Allais Paradox does not involve losses, but the Ellsberg Paradox (ambiguity aversion) requires analysis incorporating λ.

Why Prospect Theory Predictions Seem Strange

1. Nonlinear Probability Weighting

- Overweights tiny probabilities (e.g., buying lottery tickets) and underweights near-certain outcomes (e.g., preferring a sure loss over a small risk of a larger loss).

- Example: People may prefer a 1% chance of 1,000 overa 2500 , violating expected utility.

2. Loss Aversion

- Losses loom larger than gains $(\lambda > 1)$, leading to risk-seeking behavior in losses (e.g., holding losing stocks) and risk aversion in gains (e.g., selling winning stocks early).

3. Reference Dependence

- Decisions depend on framing (e.g., "saving 200 lives" vs. " 400 deaths"). Identical outcomes can be perceived differently based on context.

4. Certainty Effect

- Overvaluation of guaranteed outcomes (e.g., Allais paradox), even when mathematically dominated by riskier options.

Ways to Improve Prospect Theory

1. Dynamic Reference Points

- Allow reference points to shift over time or with experience (e.g., adapting to wealth changes).

- Example: Incorporate habit formation or expectation-based reference points.

2. Context-Dependent Parameters

- Make parameters $(\alpha, \lambda, \gamma)$ context-specific rather than fixed.

- Example: Loss aversion λ might increase under financial stress.

3. Hybrid Models

- Combine prospect theory with:

- Mental Accounting: Model how people categorize gains/losses (e.g., "house money effect").

- Ambiguity Aversion: Address Ellsberg-like uncertainty (unknown probabilities).

4. Empirical Calibration

- Use large-scale experiments to calibrate probability weighting functions $(w(p))$ for different domains (e.g., health vs. finance).

- Example: Power-law weighting $w(p) = p^{\gamma} / \left(p^{\gamma} + (1-p)^{\gamma}\right)^{1/\gamma}$ could be replaced with more flexible forms.

5. Neural/Behavioral Foundations

- Integrate findings from neuroscience (e.g., how dopamine encodes prediction errors) to ground parameters in biological mechanisms.

6. Bounded Rationality

- Incorporate cognitive constraints (e.g., limited attention, heuristic processing) to explain why some anomalies persist.

References

Exley, Christine L and Judd B Kessler (2024) "Motivated errors," *American Economic Review*, 114 (4), 961–987.