Bounding the Long Tail: AI Norms for Decision-Making under Negligible Probabilities

Vipul Razdan
Email: vr7717@gmail.com

***Abstract*—This paper recasts a long-standing decision-theory dilemma as an AI agent-design problem, proposing a principled cutoff for ultra-low-probability, extreme-utility outcomes to prevent exploitability in autonomous systems. By characterizing a vulnerability class for expected-utility maximizers and introducing a rationally negligible probability threshold grounded in cognitive skepticism, the framework preserves dominance and tractability while blocking adversarial gambles such as Pascal-type offers. Formal analysis motivates design norms for AI agents—utility bounding, calibrated priors, and epsilon-screening—together with guidance on selecting context-sensitive thresholds to maintain preference stability. This positions the proposal as a safety-centric inductive bias for rational AI decision-makers, aligning theoretical desiderata with implementable policy constraints in high-stakes, low-signal environments.**

***Index Terms*—Decision theory, expected utility, AI safety, negligible probabilities, exploitation, bounded rationality, cognitive skepticism, preference stability.**

I. Introduction

Decision theory provides a mathematical framework for rational choice under uncertainty, with expected utility maximization serving as a cornerstone normative principle [14]. Its success spans economics, operations research, and artificial intelligence, where it guides autonomous agents in probabilistic environments. However, certain pathological decision problems—St. Petersburg paradox [2], Pascal's Wager [11], and Pascal's Mugging [3]—reveal a critical flaw: expected utility theory can assign unbounded value to gambles with extremely low probability but arbitrarily high payoff. This overvaluation not only defies intuitive rationality but opens agents to systematic exploitation.

The core vulnerability arises when a malicious actor can offer an agent a "High Utility, Low Probability" (HULP) gamble. Because the expected utility of such a gamble is infinite or arbitrarily large, an expected-utility maximizer is compelled to accept any finite cost to participate, even if the offer is almost certainly fraudulent. This allows an exploiter to arbitrarily reorder the victim's preferences, effectively hijacking its decision-making process. For autonomous AI systems, which may lack human-like intuition or skepticism, this vulnerability poses a direct safety risk: a deceptive entity could induce harmful actions by promising astronomically unlikely rewards.

This paper reframes the problem as one of agent design. Rather than treating HULP scenarios as mere philosophical curiosities, we view them as a class of adversarial inputs that must be defused in practical AI architectures. We argue that rational agents should adopt a modified decision rule that truncates probabilities below a context-dependent threshold ϵ, rendering HULP gambles finite-valued and unexploitable. This approach, inspired by Smith's "Rationally Negligible Probabilities" (RNP) [13], is justified through an epistemic bound derived from the probability of catastrophic cognitive failure—the Cognitive Skepticism Hypothesis (CSH). By tying ϵ to the agent's own reliability estimate, we obtain a non-arbitrary, implementable cutoff that preserves the benefits of expected utility in ordinary decisions while blocking HULP-based attacks.

The contributions are fourfold: (1) we formalize the class of HULP problems and show how they enable preference-reordering exploitation; (2) we evaluate existing defenses (dominance reasoning, bounded utility) and show why they are insufficient or overly restrictive; (3) we advance RNP as a uniform solution and provide a principled, non-circular method for selecting ϵ based on cognitive skepticism; (4) we translate the theoretical proposal into concrete design norms for AI systems, emphasizing safety and preference stability. The resulting framework offers a tractable, philosophically grounded amendment to standard decision theory, tailored for real-world autonomous agents.

II. HULP Problems and Exploitation

A. Characterizing HULP Decision Problems

A HULP (High Utility, Low Probability) decision problem contains at least two actions: a "walk-away" action with a single outcome of zero utility, and a "HULP action" that includes at least one outcome with arbitrarily high (or infinite) utility and a correspondingly low probability. Classic examples include the St. Petersburg lottery (infinite expected dollar payout), Pascal's Wager (infinite heavenly reward), and Pascal's Mugging (arbitrarily large promised payoff). Formally, let $U(o)$ denote the utility of outcome o and $P(o)$ its probability. A HULP action A_H satisfies:

$$\exists o \in O(A_H) : U(o) \to \infty \text{ (or is arbitrarily large) and } P(o) \to 0.$$

The expected utility of A_H is therefore infinite or unbounded, while the walk-away action A_W has expected utility zero. Standard expected-utility maximization then prescribes choosing A_H regardless of any finite cost attached to it.

B. The Mechanism of HULP Exploitation

An exploiter can leverage this overvaluation to force an agent to undertake any finite-cost action. Suppose the victim initially prefers not to perform action X because it leads to an

undesirable outcome O_d with utility $U(O_d) < 0$. The exploiter promises a reward R with utility $U(R) > 0$ conditional on the victim performing X. Even if the victim assigns only a minuscule probability p to the exploiter being honest, the expected utility of performing X becomes:

$$EU(X) = p{\cdot}[U(R){+}U(O_d)]{+}(1{-}p){\cdot}U(O_d) = p{\cdot}U(R){+}U(O_d).$$

By choosing $U(R)$ sufficiently large, the exploiter can always make $EU(X) > 0$, thereby flipping the victim's preference. Crucially, the exploiter need not actually deliver R; the mere promise, however implausible, is enough to manipulate the victim's calculus.

This vulnerability is not merely theoretical. Autonomous systems that rigorously implement expected-utility maximization could be induced to transfer resources, disclose secrets, or even cause physical harm in exchange for fantastical promises. The exploit is cheap, scalable, and effective against any agent that does not discount extremely low probabilities to zero.

C. Preference Reordering and Agency Loss

HULP exploitation enables more than mere one-off coercion; it allows an attacker to arbitrarily reorder the victim's preference ranking over outcomes. Given any two outcomes A and B with $U(A) > U(B)$, an exploiter can offer a sufficiently large conditional reward for choosing B such that the agent reverses its preference. Since utilities are defined relationally, systematically rewriting preferences equates to redefining the agent's utility function. This strikes at the heart of agency: a decision-theory should help an agent achieve its own goals, not make it a puppet of another's desires.

For AI systems, preference stability is often a built-in objective [4], [10]. An agent designed to maximize a particular utility function has an instrumental reason to prevent that function from being altered, as future versions with different preferences would no longer pursue the original goals. HULP exploitation circumvents this protection, effectively allowing external parties to reprogram the agent through deceptive offers. Consequently, immunity to such exploitation should be a key desideratum for any decision-theory employed in safety-critical autonomous systems.

III. Limitations of Existing Defenses

A. Dominance-Based Decision Theories

One alternative to pure expected-utility maximization is to incorporate dominance reasoning. An action A dominates B if it yields at least as good an outcome in every state and a strictly better outcome in at least one state [12]. In problems like the Altadena vs. Pasadena gambles [9], dominance gives clear advice where expected utility is undefined. However, dominance alone is insufficient for most decisions under risk, as it falls silent whenever no action dominates (e.g., choosing between two probabilistic lotteries with overlapping outcome profiles).

When dominance is combined with expected utility—using one when the other is silent—the agent remains vulnerable to HULP exploitation. In a HULP problem, neither the HULP action nor the walk-away action dominates the other, so dominance offers no guidance. The agent then resorts to expected utility and chooses the HULP action. Thus, hybrid dominance-expected-utility theories fail to block the vulnerability.

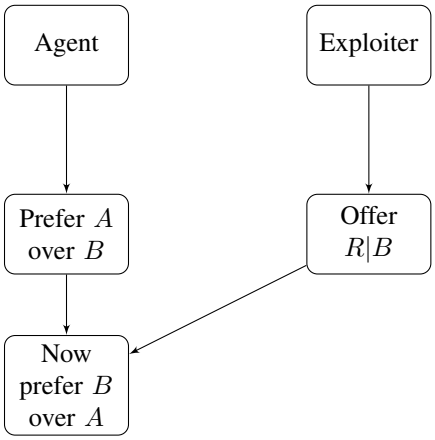

Fig. 1. HULP exploitation flips agent preferences via a conditional promise R given action B. The exploiter incurs no cost, and the agent's original ordering is subverted.

B. Bounded Utility Functions

Another classical response to infinite-value paradoxes is to assume that agents have bounded utility functions, i.e., there exists some finite maximum utility that any outcome can provide [8]. If utility is bounded, then even a St. Petersburg-type gamble converges to a finite expected value, eliminating the overvaluation. Bounded utility also blocks HULP exploitation, because no promised reward can exceed the bound.

Nevertheless, requiring bounded utility as a norm of rationality is problematic. First, it is not obviously irrational for an agent to have an unbounded utility function; an AI designed to maximize paperclip production, for example, might assign ever-increasing value to each additional paperclip without intrinsic saturation. Second, even bounded utility can be circumvented if the reward includes factors that extend the bound itself, such as increased lifespan or computational capacity [5]. Third, imposing a bound for theoretical convenience feels ad hoc—it "cuts the utility function to fit the decision theory" rather than developing a decision theory that works for natural agent preferences [13].

C. Low-Probability Cutoffs

A more direct approach is to ignore outcomes whose probability falls below a fixed threshold ϵ. Arrow [1] considered this idea but rejected it as arbitrary, noting that a universal cutoff leads to paradoxes when many low-probability outcomes collectively exhaust the probability mass. However, Smith's Rationally Negligible Probabilities (RNP) proposal refines the concept by allowing ϵ to be context-dependent and requiring that it be no larger than the highest probability assigned to any outcome in the lottery [13]. This avoids Arrow's aggregation paradox.

TABLE I
COMPARISON OF DECISION-THEORETIC DEFENSES AGAINST HULP EXPLOITATION.

Defense	**Blocks HULP?**	**Limitations**	**Tractability**
Pure dominance	Yes	Fails in non-dominant choices	Low
Bounded utility	Yes	Ad-hoc, may be circumvented	Medium
Fixed prob. cutoff	Partially	Arbitrary, aggregation issues	High
RNP with ϵ	Yes	Needs principled ϵ choice	High

RNP effectively truncates HULP gambles: outcomes with probability below ϵ are treated as having zero probability, rendering the expected utility finite. Consequently, an RNP-agent cannot be HULP-exploited, because the exploiter's promised reward will be disregarded if its likelihood is beneath the threshold. Moreover, RNP does not require tampering with the utility function, making it compatible with both bounded and unbounded preferences. The remaining challenge is to choose ϵ in a principled, non-arbitrary manner—a task taken up in the next section.

IV. A PRINCIPLED EPSILON FROM COGNITIVE SKEPTICISM

A. The Cognitive Skepticism Hypothesis (CSH)

To select ϵ without circular appeal to expected utility, we introduce an epistemic bound based on the possibility of catastrophic cognitive failure. Let CSH denote the hypothesis that an agent's reasoning faculties are systematically defective, or that its perception of the decision problem is fundamentally mistaken. CSH encompasses scenarios such as hardware faults, sensory hallucinations, malicious deception, or undetected software corruption. While exceedingly unlikely, CSH has a non-zero probability for any physically embodied agent.

Formally, let p_{CSH} be the agent's prior probability that CSH holds. This probability can be estimated from failure rates of underlying components, historical error frequencies, or theoretical models of reliability. For a well-designed AI system, p_{CSH} might be extremely small (e.g., 10^{-10} or less), but it remains positive.

B. Disregard Sub-Skepticism Hypotheses (DSH)

We propose the following norm: agents should disregard any hypothesis H whose probability is lower than p_{CSH}. The rationale is that if H is less likely than a complete failure of cognitive reliability, then even evidence apparently supporting H is more plausibly explained by CSH. In Bayesian terms, CSH acts as a "dead hypothesis" [7] with respect to any such H: because CSH has higher prior and can explain the same observations, no amount of data can raise the posterior of H above that of CSH.

Consequently, considering H cannot yield useful decision-theoretic guidance. If the agent treats H as live, it must also take CSH more seriously, which undermines confidence in its own decision-making process. This self-defeating dynamic justifies ignoring probabilities below the CSH threshold. Thus, we set $\epsilon \geq p_{\text{CSH}}$ as a lower bound for RNP.

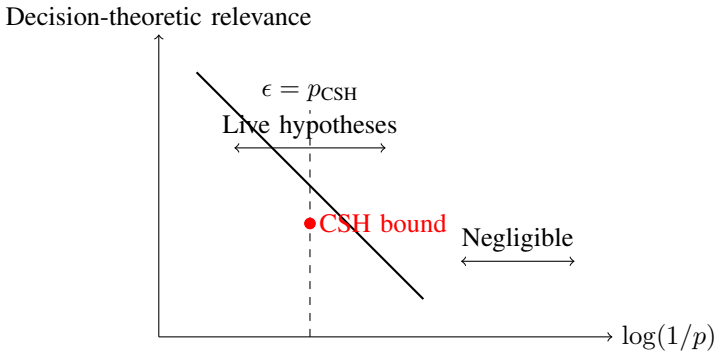

Fig. 2. Schematic of the probability relevance threshold. Hypotheses with probability below ϵ (right of dashed line) are dominated by the Cognitive Skepticism Hypothesis and can be disregarded without loss of rational guidance.

C. Microscope and Event-Horizon Analogies

The CSH bound can be understood through two physical analogies. First, consider a microscope: it extends human vision but has a diffraction limit beyond which finer details cannot be resolved. Similarly, decision theory extends rational choice but has a "probability resolution limit" given by p_{CSH}; probabilities finer than this limit cannot be meaningfully utilized because they are overshadowed by the risk of cognitive failure.

Second, akin to a black-hole event horizon, p_{CSH} forms a probability threshold: once a hypothesis drops below it, no evidence can lift it back above, because any confirming data is more likely due to CSH. This does not indicate a flaw in probability theory, just as an event horizon does not indicate a flaw in general relativity; it is a natural consequence of applying the theory to extreme regimes.

D. Implementing ϵ in AI Systems

For practical AI design, ϵ can be set as a configurable safety parameter. The lower bound p_{CSH} can be estimated via reliability engineering (e.g., mean time between failure, error rates in perception modules, historical data on deception attempts). The actual ϵ may be set higher for additional safety margins, or adjusted dynamically based on context (e.g., stricter thresholds in high-stakes settings). Importantly, ϵ is not a universal constant but an agent- and environment-specific value, aligning with Smith's requirement that ϵ may vary across decision problems and agents [13].

This approach avoids arbitrariness: the threshold is tied to the agent's own epistemic limitations, a consideration that is already relevant for any bounded rational entity. It also avoids circularity, as the choice of ϵ is grounded in reliability estimates, not in the utilities at stake in the decision itself.

V. Design Norms for Unexploitable AI Agents

A. Core Architectural Principles

To embed HULP-resistance into autonomous systems, we propose three design norms:

1. Epsilon-screening: The agent's decision module should automatically discard outcomes with probability below a configured ϵ. This can be implemented as a pre-processing step: any conditional promise or lottery offering with likelihood $< \epsilon$ is treated as having zero probability. The value of ϵ should be set conservatively, using the CSH bound as a baseline.

2. Utility bounding (optional but recommended): While not strictly required, bounding the utility function provides an additional layer of protection. If the system's goal can be expressed with a natural saturation point (e.g., "maximize human welfare up to a feasible maximum"), a bounded utility function simplifies analysis and closes potential loopholes where rewards might extend the agent's capacity.

3. Calibrated priors for deception: The agent should maintain a prior over "malicious offer" hypotheses, updated by experience. This prior can inform dynamic adjustments to ϵ in environments with frequent deception attempts. Bayesian filtering can help distinguish between genuinely low-probability opportunities and systematic adversarial patterns.

B. Group Decision-Making and Robustness

In multi-agent systems, the CSH bound can be strengthened. If N independent agents each have individual failure probability p_{CSH}, the probability that a majority suffer simultaneous cognitive failure drops exponentially with N. Hence, a committee or consensus mechanism can effectively lower the operative ϵ, allowing the collective to safely consider probabilities that would be negligible for a single agent. This mirrors engineering practices where redundancy increases overall reliability.

C. Handling Edge Cases and Objections

A common objection is that ignoring probabilities below ϵ could lead to disregarding genuinely grave risks, such as a nuclear meltdown with extremely low probability. In practice, most catastrophic risks have probabilities far above realistic CSH estimates (which are exceedingly small). Moreover, for risks that are both severe and very unlikely, specialized safety analyses (e.g., fault-tree analysis, probabilistic risk assessment) operate outside the agent's everyday decision loop and can employ different thresholds.

Another concern is that RNP might conflict with Bayesian updating. However, as Jaynes [7] clarifies, dead-hypothesis phenomena are a consequence of Bayesian reasoning, not a contradiction. If a hypothesis is below the CSH bound, Bayesian updating will never raise it above that bound, justifying its exclusion from decision-theoretic consideration.

VI. Conclusion and Future Work

This paper has reframed a classic decision-theory vulnerability as a concrete AI-safety problem. HULP exploitation—where an adversary uses low-probability, high-utility promises to hijack an agent's preferences—poses a real threat to autonomous systems that rigidly maximize expected utility. We have shown that existing defenses (dominance, bounded utility) are either incomplete or overly restrictive, whereas Rationally Negligible Probabilities (RNP) offers a balanced solution.

By grounding the RNP threshold ϵ in the probability of cognitive failure (CSH), we provide a principled, non-arbitrary method for selecting the cutoff. This approach respects the normative force of expected utility in ordinary decisions while blocking adversarial gambles. We further translate the theory into practical design norms for AI agents: epsilon-screening, optional utility bounding, and calibrated deception priors.

Future work should explore several directions. First, empirical studies could test how different ϵ settings affect agent performance in simulated environments containing deceptive actors. Second, the CSH bound could be refined using more detailed models of hardware/software reliability and adversarial manipulation. Third, the framework could be extended to sequential decision problems (MDPs/POMDPs) where low-probability threats might arise over long horizons. Finally, integration with existing AI safety techniques (e.g., debate, interpretability, reward modeling) could create multi-layered protection against preference corruption.

For AI designers, the lesson is clear: rational choice under uncertainty must include safeguards against extreme-tail exploitation. By adopting a negligibility threshold informed by the agent's own epistemic limits, we can build systems that are both rational and robust—systems that pursue their goals without falling for infinite mirages.

References

[1] K. J. Arrow, "Alternative approaches to the theory of choice in risk-taking situations," *Econometrica*, pp. 404–437, 1951.

[2] D. Bernoulli, "Exposition of a new theory on the measurement of risk," *Commentaries of the Imperial Academy of Science of Saint Petersburg*, 1738 (reprinted 1954).

[3] N. Bostrom, "Pascal's mugging," *Analysis*, pp. 443–445, 2009.

[4] N. Bostrom, "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents," *Minds and Machines*, vol. 22, no. 2, pp. 71–85, 2012.

[5] T. Cowen and J. High, "Time, bounded utility, and the St. Petersburg paradox," *Theory and Decision*, vol. 25, no. 3, pp. 219–223, 1988.

[6] A. Hájek, "Waging war on Pascal's Wager," *The Philosophical Review*, vol. 112, no. 1, pp. 27–56, 2003.

[7] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[8] K. Menger, "Das Unsicherheitsmoment in der Wertlehr," *Zeitschrift für Nationalökonomie*, vol. 51, pp. 459–485, 1934.

[9] H. Nover and A. Hájek, "Vexing expectations," *Mind*, vol. 113, no. 450, pp. 237–249, 2004.

[10] S. M. Omohundro, "The basic AI drives," in *Proc. AGI*, vol. 171, 2008, pp. 483–492.

[11] B. Pascal and E. Havet, *Pensées*. Dezobry et E. Magdeleine, 1852.

[12] M. D. Resnik, *Choices: An Introduction to Decision Theory*. University of Minnesota Press, 1987.

[13] N. J. Smith, "Is Evaluative Compositionality a Requirement of Rationality?" *Mind*, vol. 123, no. 490, pp. 457–502, 2014.

[14] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 1944.