A HYPOTHESIS ON BLACK SWAN IN UNCHANGING ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Black swan events are statistically rare occurrences that carry extremely high risks. A standard view of black swans assumes that they originate from an unpredictable and changing environment; however, the community lacks a comprehensive definition of black swan events. To this end, this paper challenges that the standard view is *incomplete* and claims that high-risk, statistically rare events can also occur in unchanging environments due to human misperception of events' values and likelihoods, which we refer to as S-BLACK SWAN . We first carefully categorize black swan events, focusing on S-BLACK SWAN , and mathematically formalize the definition of black swan events. We hope these definitions can pave the way for the development of algorithms to prevent such events by rationally correcting limitations in perception.

1 INTRODUCTION

To successfully deploy machine learning (ML) systems in open-ended environments, these systems must exhibit robustness against *rare and high-risk events*, often referred to as *black swans* (Taleb, 2010). Achieving this robustness requires a deep and precise understanding of the origins of such events, which has been increasingly recognized as a critical factor for enabling ML algorithms to attain full control and make optimal decisions (Chollet, 2019; Silva & Najafirad, 2020; He et al., 2021; Li et al., 2023; Yang et al., 2024). Nevertheless, many contemporary ML systems remain vulnerable to black swans in real-world scenarios, as evidenced by automated trading systems that overreact to market anomalies (Kirilenko et al., 2017; Phillips, 2021; Stafford, 2022), unexpected bankruptcies (Wiggins et al., 2014; Akhtaruzzaman et al., 2023), the Covid pandemic (Antipova, 2020), and autonomous vehicles encountering unforeseen road or weather conditions (Tesla, 2021; Witman et al., 2023; Nordhoff et al., 2023).

In this paper, we argue that ML systems remain susceptible to black swan events, regardless of an algorithm's representation capacity or scalability, due to an AI community's *incomplete* understanding of the origins of these events. The prevailing belief in most algorithmic approaches to preventing black swan events (Prestwich, 2019; Artemenko et al., 2020; Devarajan et al., 2021; Wabartha et al., 2021; Bhanja & Das, 2024; Jin, 2024) is that such events primarily arise from *dynamic, time-varying* environments. We contend, however, that black swans can also emerge from *static, stationary* environments. To this end, we propose a new hypothesis on their origins:

> **Hypothesis 1.** Black swans can originate from misperceptions of an event's reward and likelihood, even within static environments.

To warmly introduce our new hypothesis, consider the bankruptcy of Lehman Brothers, widely recognized as the most significant black swan event in the financial industry (Wiggins et al., 2014). A strong explanation points to the investors making rational decisions on the false market perception which appeared rational at the time but proved irrational by correcting their perception in hindsight . The firm declared bankruptcy within 72 hours without any precursor (McDonald & Robinson, 2009), and the only factor that changed during those three days was investors' perception of the company (Housel, 2023; Mawutor, 2014; Fleming & Sarkar, 2014) [1]. Investors made optimal decisions based on this perception, which turned out to be suboptimal once the perception was revealed to be false during those 72 hours.

Contribution. We refer to black swan events in stationary environments as **S-BLACK SWAN** and define them in the context of a Markov Decision Process (MDP) as follows:

> **(Informal)** *An* S-BLACK SWAN *event is a state-action pair where humans misperceive both its likelihood and reward. It is perceived as impossible, despite occurring with small probability, while its reward is overestimated relative to its true value in a stationary environment.*

Our work begins with a case study on how S-BLACK SWAN emerge and cause suboptimality gaps in various MDP settings, such as bandit (Theorem 1), small state spaces (Theorem 2), and large state spaces (Theorem 3). We introduced three MDPs to define S-BLACK SWAN : the ground truth MDP (GMDP), the Human MDP (HMDP), and the Human-Estimation MDP (HEMDP). The GMDP represents the real world, while the HMDP reflects humans' biased perceptions (Definitions 1 and 2). S-BLACK SWAN (Definitions 4 and 5) are state-action pairs perceived as impossible in the HMDP but occur with small probability and higher rewards in the GMDP. Our main finding (Theorem 4) shows that while the HEMDP value function asymptotically converges to that of the HMDP over longer horizons, the gap between HMDP and GMDP has a lower bound, influenced by reward distortion, the size of the S-BLACK SWAN set, and their minimum probability of occurrence. Finally, Theorem 5 examines S-BLACK SWAN hitting time, showing that larger reward distortion and higher S-BLACK SWAN probability necessitate more frequent updates to human perception functions.

2 PRELIMINARY

Notations. The sets of natural, real, nonnegative, and nonpositive real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, and $\mathbb{R}_{\leq 0}$ respectively. For a finite set Z, the notation $|Z|$ represents its cardinality, and $\Delta(Z)$ denotes the probability simplex on Z. Given $X, Y \in \mathbb{N}$ with $X < Y$, we define $[X] := \{1, 2, \ldots, X\}$, the closed interval $[X, Y] := \{X, X+1, \ldots, Y\}$. For $x \in \mathbb{R}_{\geq 0}$, the floor function $\lfloor x \rfloor$ is defined as $\max\{n \in \mathbb{N} \cup \{0\} \mid n \leq x\}$[2].

Markov Decision Process. We consider a finite-horizon MDP denoted as $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma, T \rangle$, where $P = \{P_t\}_{t=0}^T$ and $R = \{R_t\}_{t=0}^T$ for $t \in \mathbb{N}$. Here, $\mathcal{S}$ represents the state space, $\mathcal{A}$ denotes the action space, $P_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability function at time t, $R_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function at time t, γ is the discount factor, and $T \in \mathbb{N}$ is the horizon length. We define $\mathcal{M}$ as a stationary MDP if $P_t(s' \mid s, a) = P_{t+1}(s' \mid s, a)$ and $R_t(s, a) = R_{t+1}(s, a)$ for all $(s', s, a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$ and for all $t \in [T-1]$. Otherwise, we define $\mathcal{M}$ as a non-stationary MDP. In the stationary case, we denote P and R as the single transition probability function and reward function, respectively. A policy is denoted as $\pi \in \Pi$, where $\Pi : \mathcal{S} \to \Delta(\mathcal{A})$ is the set of policies. We denote a T-length trajectory from $\mathcal{M}$ under policy π as $\{s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, s_T\}$, where $s_t \sim P_t(\cdot \mid s_{t-1}, a_{t-1})$ and $r_t = R_t(s_t, a_t)$. Assume that all rewards are bounded, i.e., $r_t \in [-R_{\max}, R_{\max}]$ for all t. The agent's goal is to compute the optimal policy $\pi^\star \in \Pi$ that maximizes the value function: $V_{\mathcal{M}}^{\pi}(s) := \mathbb{E}_\pi[\sum_{t=0}^T \gamma^t R_t(s_t, a_t) \,|\, P, s_0 = s]$. We further define the normalized visitation probability as $P^\pi(s, a) := \frac{1-\gamma^T}{1-\gamma} \sum_{t=0}^{T-1} \gamma^t \mathbb{P}((s_t, a_t) = (s, a) | s_0, \pi, P)$, where

[1] The bank's loss endurance, evaluated at 11.7% by the U.S. government, stayed *stationary* over the 72 hours.

[2] For clarity and readability, all notations used throughout the entire paper are elaborated in Appendix A

$\mathbb{P}(s, a|s_0, \pi, P)$ is the probability of visiting (s, a) at time t under policy π and transition probability P starting from s_0 .

The following three theorems, drawn from existing work, lay the groundwork for mathematically formulating *misperception* of the Hypothesis 1.

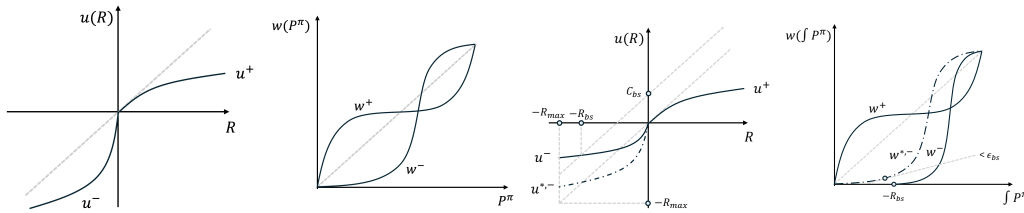
Expected Utility Theory. Given an outcome space $\mathcal{O} = \{o_1, \ldots, o_K\}$, we define a utility function $g : \mathcal{O} \to \mathbb{R}$ that quantifies the gain or loss associated with each outcome o_i. An individual agent is faced with choices, where each choice represents a scenario in which the outcomes o_i occur with given probabilities p_i, summing to one. The set of all choices is denoted by $\mathcal{C}$. Each choice $c \in \mathcal{C}$ returns $\mathcal{O}$ with a probability distribution $\boldsymbol{p}_c = (p_1^{(c)}, \ldots, p_K^{(c)})$. Under a given choice c, *Expected Utility Theory (EUT)* evaluates the riskiness of that choice as $V(c) = \sum_{i=1}^{K} g(o_i) p_i^{(c)}$ (von Neumann, 1944; Rabin, 2013). To illustrate, consider a stock market investment scenario where $\mathcal{O} = \{\text{Economic Boom (EB)}, \text{Economic Recession (ER)}\}$. Here, $g(EB)$ represents a gain, while $g(ER)$ represents a loss. The set of choices $\mathcal{C} = \{\text{invest in stocks}, \text{invest in bonds}, \text{keep cash}\}$ corresponds to different probability distributions $\boldsymbol{p}_c = (p_1^{(c)}, p_2^{(c)})$ of outcomes.

Prospect Theory. However, *Expected Utility Theory (EUT)* fails to account for empirical observations from psychological experiments (Drakopoulos & Theodossiou, 2016; Pandit et al., 2019; Wahlberg & Sjoberg, 2000; Vasterman et al., 2005; van der Meer et al., 2022) and economic cases (Rogers, 1998; Wheeler & Wheeler, 2007; BetterUp, 2022) that demonstrate human irrationality. Specifically, humans tend to exhibit internal distortions when perceiving event probabilities $\boldsymbol{p}_c$ and evaluating outcome values $g(\mathcal{O})$ for any choice c (Opaluch & Segerson, 1989). To address these discrepancies, *Prospect Theory (PT)* introduces a probability distortion function $w : [0, 1] \to [0, 1]$ and a value distortion function $u : \mathbb{R} \to \mathbb{R}$, which modify the expected utility calculation to $V(c) = \sum_{i=1}^{K} u(g(o_i)) w(p_i^{(c)})$ (Kahneman & Tversky, 2013; Fennema & Wakker, 1997). The motivation for introducing *PT* is not only to acknowledge human irrationality but also to provide a more accurate mathematical framework for how people actually perceive probabilities and outcomes. *PT* describes the characteristics of the functions u and w based on empirical case studies. The function u represents *value distortion*, capturing how individuals assess gains and losses (x-axis of Figure 1a represents the true value, and the y-axis represents the perceived value). The function w represents *probability distortion*, reflecting how individuals tend to overestimate the likelihood of rare events and underestimate the likelihood of more probable events. (x-axis of Figure 1b represents the true probability, and the y-axis represents the perceived probability.)

Cumulative Prospect Theory. To enhance mathematical rigor—specifically, to ensure that distorted probabilities still sum to one—*Prospect Theory (PT)* was further revised into *Cumulative Prospect Theory (CPT)*. In *CPT*, the expected value is defined as $V(c) = \sum_{i=1}^{K} u(g(o_i)) \left(w\left(\sum_{j=1}^{i} p_j^{(c)}\right) - w\left(\sum_{j=1}^{i-1} p_j^{(c)}\right)\right)$, where the function w distorts the cumulative probability of an event o_i. The following insurance example illustrates *CPT* in action.

Example 1 (Insurance policies)**.** *Consider an example where the probability of an insured risk is* 1%*, the potential loss is* $1,000$*, and the insurance premium is* 15*. According to CPT, most would opt to pay the* 15 *premium to avoid the larger loss.*

Example 1 shows how a simple decision can be modeled as a two-step Markov Decision Process with states $\mathcal{S} = \{s_{base}, s_{premium}, s_{risk}\}$ representing utility value of 0, −15, and −1000, and actions (or choice set $\mathcal{C}$) $\mathcal{A} = \{a_p, a_{np}\}$ for paying or not paying the premium. At $t = 0$, humans choose between a_p (leading to $s_{premium}$) and a_{np}, which could result in s_{base} with 99% probability or s_{risk} with 1% probability. Expected utility theory suggests a_{np} is optimal since its expected value ($V(a_{np}) = -1000 \cdot 0.01 = -10$) is lower than that of a_p ($V(a_p) = -15 \cdot 1 = -15$), but real-world decisions often favor a_p, highlighting a divergence from theoretical rationality.



(a) Value distortion (b) Probability distortion (c) u with black swans (d) w with black swans

Figure 1: Value distortion function u and probability distortion function w. The gray line in Figures 1a and 1b represents $y = x$.

Therefore, we begin by formalizing the key empirical observations from *CPT* into the following definitions.

Definition 1 (Value Distortion Function)**.** *The value distortion function* u *is defined as:*

$$u(x) = \begin{cases} u^+(x) & \text{if } x \geq 0, \\ u^-(x) & \text{if } x < 0, \end{cases}$$

where $u^+ : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *is non-decreasing, concave with* $\lim_{h \to 0^+} (u^+)'(h) \leq 1$*, and* $u^- : \mathbb{R}_{\leq 0} \to \mathbb{R}_{\leq 0}$ *is non-decreasing, convex with* $\lim_{h \to 0^-} (u^-)'(h) > 1$*.*

Definition 2 (Probability Distortion Function)**.** *The probability distortion function* w *is defined as:*

$$w(p_i) = \begin{cases} w^+(p_i) & \text{if } g(x_i) \geq 0, \\ w^-(p_i) & \text{if } g(x_i) < 0, \end{cases}$$

where $w^+, w^- : [0,1] \to [0,1]$ *satisfy:* $w^+(0) = w^-(0) = 0$*,* $w^+(1) = w^-(1) = 1$*;* $w^+(a) = a$ *and* $w^-(b) = b$ *for some* $a, b \in (0,1)$*;* $(w^+)'(x)$ *is decreasing on* $[0,a)$ *and increasing on* $(a,1]$*;* $(w^-)'(x)$ *is increasing on* $[0,b)$ *and decreasing on* $(b,1]$*.*

The derivative constraints encapsulate the core observations of *CPT*. Specifically, the conditions on $(u^-)'$ and $(u^+)'$ in Definition 1 formalize the tendency for individuals to value losses more heavily than equivalent gains (see Figure 1a). The constraints on $(w^-)'$ and $(w^+)'$ in Definition 2 describe the tendency to overweight (or underweight) the probabilities of rare events and underweight (or overweight) those of average events where the outcome results in a gain (or a loss) (see Figure 1b).

3 Black Swan in stationary and non-stationary environments

Hypothesis 1 concerns the feasibility of black swan events existing in stationary environments. We next illustrate how black swans can originate from both stationary and non-stationary environments. We begin by defining the black swan event dimension as follows.

Definition 3 (Black Swan Event Dimension)**.** *For a given MDP* $\mathcal{M}$*, we define the dimension of a black swan event as the set* $\mathcal{S} \times \mathcal{A} \times [T]$*.*

Then, we informally refer to $(s, a, t_{bs}) \in \mathcal{S} \times \mathcal{A} \times [T]$ as a black swan event if it represents a rare, high-risk occurrence that significantly deviates from expected outcomes based on prior experience in the real world $\mathcal{M}$. This could involve an unexpected transition or an anomalous reward signal. We then introduce a classification rule that distinguishes black swan events based on whether they occur in non-stationary environments or arise within stationary environments, as follows.

Algorithm 1 (Black Swan Classification: S-BLACK SWAN)**.** *For a given (possibly non-stationary)* $\mathcal{M}$*, suppose* (s, a, t_{bs}) *is a black swan event. If* (s, a, t) *is a black swan event for* $\forall t \in [T]$*, then we classify* (s, a, t_{bs}) *as a black swan that originates from environment's stationarity (***S-BLACK SWAN** *).*

Based on Algorithm 1, one can always identify a unit time interval that classifies any black swan event as an S-BLACK SWAN , as stated in the following proposition.

Proposition 1. *If (s, a, t_{bs}) is a black swan event, then there exists a time interval $[t_1, t_2] \subseteq [T]$ such that for every $t \in [t_1, t_2]$, the (s, a, t) is classified as* S-BLACK SWAN *.*

We provide an intuitive interpretation of Proposition 1 through the following example.

Example 2. *Suppose (s, a, t_{bs}) is a black swan event.*

- *Case 1. Consider $\mathcal{M}$ as a non-stationary MDP where P_t and R_t change at each time step, i.e., $P_t \neq P_{t+1}$ and $R_t \neq R_{t+1}$. If $t_1 = t_2 = t_{bs}$, then (s, a, t_{bs}) is classified as an* S-BLACK SWAN *. However, if $t_1 \neq t_2$ and $t_{bs} \in [t_1, t_2]$, then (s, a, t_{bs}) cannot be definitively classified as an* S-BLACK SWAN *.*
- *Case 2. Consider $\mathcal{M}$ as a piecewise non-stationary MDP where P_t and R_t change every $\lfloor T/k \rfloor$ time steps, i.e., $P_t = P_{t+1}$ and $R_t = R_{t+1}$ for $t \in [kj, kj + (k-1)]$ where $j = 0, 1, \dots, \lfloor T/k \rfloor$. If $t_1 = kj_{bs}$ and $t_2 = kj_{bs} + (k-1)$, then (s, a, t_{bs}) is classified as an* S-BLACK SWAN *where j_{bs} satisfies $t_{bs} \in [kj_{bs}, kj_{bs} + (k-1)]$.*
- *Case 3. Consider $\mathcal{M}$ as a stationary MDP where $P_t = P_{t+1}$ and $R_t = R_{t+1}$ for all $t \in [T-1]$. In this case, (s, a, t_{bs}) is always classified as an* S-BLACK SWAN *, regardless of the interval $[t_1, t_2]$.*

We then present Case 3 of Example 2 as the following main remark:

Remark 1. *If $\mathcal{M}$ is stationary, then any black swan event (s, a, t) is classified as an* S-BLACK SWAN *. In this case, we omit t and denote the* S-BLACK SWAN *simply as (s, a).*

Our main goal for the remainder of the paper is to explore Remark 1, with a focus on mathematically defining S-BLACK SWAN within a *stationary* MDP $\mathcal{M}$. We will retain the notation for stationary transition probabilities and reward functions as P and R, respectively, omitting the subscript t.

4 THE EMERGENCE OF S-BLACK SWAN IN SEQUENTIAL DECISION MAKING

We next present a case study to substantiate Hypothesis 1 before formally defining S-BLACK SWAN . We begin by examining how S-BLACK SWANS emerge in sequential decision-making within a *stationary environment*, starting with the bandit case. For a given $(s, a) \in \mathcal{S} \times \mathcal{A}$, let us assume that the function u distorts the reward $R(s, a)$, and the function w distorts the transition probabilities $\{P(s'|s, a)\}_{\forall s' \in \mathcal{S}}$ where s' is the next state. In this Section, we refer to the MDP distorted by functions u and w as the distorted MDP $\mathcal{M}_d \coloneqq \langle \mathcal{S}, \mathcal{A}, w(P), u(R), \gamma \rangle$, with this notation being used exclusively within this section.

4.1 CASE 1. CONTEXTUAL BANDIT ($T = 1$)

We begin with a simple case where the horizon length is $T = 1$, commonly referred to as a contextual bandit (Lattimore & Szepesvári, 2020). Surprisingly, in this setting, the optimal policy of a distorted world coincides with the real world optimal policy as a following Theorem.

Theorem 1 (One-Step Optimality Deviation)**.** *If $T = 1$, then the optimal policy in the MDP $\mathcal{M}$ is identical to the optimal policy in the distorted MDP $\mathcal{M}_d$.*

Theorem 1 may seem counterintuitive, as Example 1 illustrates that human decision-making often exhibits irrationality. In single-step decision-making, distortions in perception do not significantly affect the optimal policy. For clarification, as shown in Example 1, the perceived reward order remains $u^-(r(s_{loss})) < u^-(r(s_{premium})) < u^-(r(s_{base}))$ because u^- is a non-decreasing convex function. This further implies that a *short* decision horizon may *reduce* the influence of human irrationality.

4.2 Case 2. $|\mathcal{S}| = 2$ when $T > 1$

Now, let us consider the simplest case where $T > 1$ and $|\mathcal{S}| = 2$. Surprisingly, the result that optimality does not deviate still holds similarly to Theorem 1.

Theorem 2 (Multi-step Optimality Deviation with $|\mathcal{S}| = 2$)**.** *If $|\mathcal{S}| = 2$, then the optimal policy from the MDP $\mathcal{M}$ is also identical to the optimal policy of the distorted MDP $\mathcal{M}_d$ for all $t \in [T]$.*

Theorem 2 may initially seem counterintuitive, given that model errors propagate through distorted transition probabilities and rewards as time t progresses (Janner et al., 2019). However, a straightforward explanation is that for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, the function w preserves the order of probabilities. Specifically, if $P(s_1|s,a) > P(s_2|s,a)$, then $w(P(s_1|s,a)) > w(P(s_2|s,a))$ still holds, where $\mathcal{S} = \{s_1, s_2\}$. This suggests that when the state space $|\mathcal{S}|$ is small, the informational complexity required to determine the real-world optimal action remains relatively *low*.

4.3 Case 3. $|S| = 3$ with unbiased reward perception

We now consider a general setting with arbitrary $\mathcal{S}$, $\mathcal{A}$, and T, but under the assumption that $u(R(s,a)) = R(s,a)$ for all (s,a), indicating that humans have an unbiased perception of their rewards.

Theorem 3 (Two-step Optimality Deviation with $|\mathcal{S}| = 3$)**.** *If $|\mathcal{S}| = 3$ and $T = 2$, there exists a transition probability function P and a reward function R such that the optimal policy of the MDP $\mathcal{M}$ differs from that of the distorted MDP $\mathcal{M}_d$.*

The optimality deviation in Theorem 3 now aligns with the empirical observation in model-based reinforcement learning; increasing suboptimality is caused by model error propagation (Janner et al., 2019). In summary, Theorems 1, 2, and 3 demonstrate that the discrepancy between the optimal policy derived from human perception and the real-world optimal policy increases as the complexity of the environment ($\mathcal{S}$) grows or as the horizon length (T) extends, regardless of the w function.

5 Agent- Environment framework : perception as intersection

To explore Hypothesis 1, we propose a novel agent-environment framework that treats misperception as information loss in an agent's understanding of the real world [3] (See Figure 2). This framework introduces two *stationary* MDPs: the Human MDP and the Human-Estimation MDP. We begin by defining the *stationary* ground MDP (GMDP) $\mathcal{M}$ as an abstraction of real-world environments without information loss. The following subsections detail the Human MDP (HMDP) and the Human-Estimation MDP (HEMDP).

5.1 Human MDP

We define the Human MDP $\mathcal{M}^\dagger = \langle \mathcal{S}, \mathcal{A}, P^\dagger, R^\dagger, \gamma, T \rangle$, where the human (agent) misperceives the visitation probability $P^\pi(s,a)$ through the function w, denoted as $P^{\dagger,\pi}(s,a)$, and the reward function $R(s,a)$ through the function u, denoted as $R^\dagger(s,a)$. An internal assumption in the HMDP is that its state and action spaces are identical to those of the GMDP $\mathcal{M}$, i.e., $\mathcal{S}^\dagger = \mathcal{S}$ and $\mathcal{A}^\dagger = \mathcal{A}$. Although this assumption may seem unrealistic, especially given that insufficient exploration in large discrete state and action spaces may violate it, the following method shows how the human (agent) can approximate $\mathcal{S}^\dagger$ and $\mathcal{A}^\dagger$ to $\mathcal{S}$ and $\mathcal{A}$, thus supporting this assumption.

Remark 2. *If the human (agent) cannot perceive a state $s \in \mathcal{S}$, the state space $\mathcal{S}^\dagger$ can be updated to $\mathcal{S}^\dagger \leftarrow \mathcal{S}^\dagger \cup \{s\}$, then set $R^\dagger(s,a) = R(s,a)$ and $P^\dagger(s' \mid s,a) = P(s' \mid s,a)$ while ensuring $P(s \mid s',a) = 0$*

[3] We detail how misperception reflects information loss from the agent's perspective in Appendix B. .

for all $s \in \mathcal{S}^\dagger$ and $a \in \mathcal{A}^\dagger$. As a result, the new state s does not influence decision-making in the HMDP, since the probability of the trajectory visiting s remains zero.

For discrete $\mathcal{S}$ and $\mathcal{A}$, the order statistics of P^π can be defined over the sequence $[|\mathcal{S}||\mathcal{A}|]$, with each (s,a) corresponding to an order index in $[|\mathcal{S}||\mathcal{A}|]$, enabling the subsequent definition of the cumulative distribution. For brevity, we denote the cumulative distribution of $P^\pi(s,a)$ as $\int P^\pi(s,a)$. The distortions are then defined by the following relationships:

$$\int P^{\dagger,\pi}(s,a) = \begin{cases} w^+(\int P^\pi(s,a)) \text{ if } R(s,a) \geq 0 \\ w^-(\int P^\pi(s,a)) \text{ if } R(s,a) < 0 \end{cases} \quad, \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{1}$$

$$R^\dagger(s,a) = \begin{cases} u^+(R(s,a)) \text{ if } R(s,a) \geq 0 \\ u^-(R(s,a)) \text{ if } R(s,a) < 0 \end{cases} \quad, \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{2}$$

We introduce the concept of the *perception gap*: if $\max_{(s,a)} |R(s,a) - R^\dagger(s,a)| < \epsilon_r$, then $R^\dagger(s,a)$ is referred to as an ϵ_r-perceived reward. Similarly, if $\max_{(s,a)} |P^\pi(s,a) - P^{\pi,\dagger}(s,a)| < \epsilon_d$, then $P^{\dagger,\pi}(s,a)$ is called an ϵ_d-perceived visitation probability, where $\epsilon_r, \epsilon_d \in \mathbb{R}_+$. The case where $\epsilon_r = \epsilon_d = 0$ represents an *unbiased perception*. Once the agent perceives $\mathcal{M}$ as $\mathcal{M}^\dagger$, it executes the policy π in $\mathcal{M}^\dagger$ and collects a trajectory. Finally, the value function of $\mathcal{M}^\dagger$ is given by $V^\pi_{\mathcal{M}^\dagger}(s) \coloneqq \mathbb{E}_\pi \left[\gamma^t R^\dagger(s_t,a_t) \middle| P^\dagger, s_0 = s\right]$.

A key challenge in understanding $\mathcal{M}^\dagger$ is why distortions occur in visitation probability rather than transition probability, as discussed in Section 5. This distinction arises because (s,a) is the fundamental event unit (see Remark 1), and a distortion in transition probability implies a distortion in the state itself. The central question, then, is how distortions in visitation probability relate directly to data collection. The following lemma partially addresses this question.

Lemma 1. *For a given $\mathcal{M}$, there always exists a function $h : \mathcal{S} \to \mathcal{S}$ such that $w(\int P^\pi(s,a)) = \int P^\pi(h(s),a)$ holds for any function w.*

Our perspective is that distortions in the probability distribution, state space, or other factors lead to distortions in visitation probabilities. With unbiased perception, the agent collects a trajectory $\tau = \{s_0, a_0, r_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T\}$. However, when the agent perceives $\mathcal{M}$ as $\mathcal{M}^\dagger$, it observes a distorted trajectory $\tau^\dagger = \{h(s_0), a_0, u(r_0), h(s_1), a_1, \ldots, h(s_{T-1}), a_{T-1}, h(s_T)\}$, where function h distorts the states. Lemma 1 demonstrates that visitation probability distortion arises from state distortion via h.

5.2 Human-Estimation MDP

After the agent have perceived world as $\mathcal{M}^\dagger$, it *estimates* the perceived reward $R^\dagger(s,a)$ as $\widehat{R}^\dagger(s,a)$ and visitation probability $P^{\dagger,\pi}(s,a)$ as $\widehat{P}^{\dagger,\pi}(s,a)$ from its trajectory $\tau^\dagger$. We define a Human-Estimation MDP as $\widehat{\mathcal{M}}^\dagger = \langle \mathcal{S}, \mathcal{A}, \widehat{P}^\dagger, \widehat{R}^\dagger, \gamma, T \rangle$. Note that this estimation process is the same as estimation of generative model in model-based reinforcement learning (Gheshlaghi Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020; Kakade, 2003). We also introduce *estimation gap*, that is if $\max_{(s,a)} |R^\dagger(s,a) - \widehat{R}^\dagger(s,a)| \leq \kappa_r$ holds, then $\widehat{R}^\dagger(s,a)$ is κ_r-estimated reward, and if $\max_{(s,a)} |P^{\pi,\dagger}(s,a) - \widehat{P}^{\pi,\dagger}(s,a)| \leq \kappa_d$ holds, then $\widehat{P}^{\pi,\dagger}(s,a)$ is κ_d-estimated visitation probability for constant $\kappa_r, \kappa_d \in \mathbb{R}_+$. Finally, the value function of $\widehat{\mathcal{M}}^\dagger$ is given as $V^\pi_{\widehat{\mathcal{M}}^\dagger}(s) \coloneqq \mathbb{E}_\pi \left[\gamma^t \widehat{R}^\dagger(s_t,a_t) \middle| \widehat{P}^\dagger, s_0 = s\right]$.

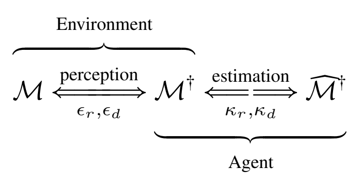


Figure 2: The agent and environment intersect with perception.

We use the perception and estimation gaps to illustrate the novel agent-environment framework in Figure 2.

6 S-BLACK SWAN

Finally, Section 6 provides a definition of S-BLACK SWAN and presents a theoretical analysis aimed at guiding the design of safer ML algorithms in the future.

6.1 A DEFINITION OF S-BLACK SWAN

Assume that the rewards for all state-action pairs are ordered as $R_{[1]} \leq \cdots \leq R_{[l]} \leq 0 \leq R_{[l+1]} \leq \cdots \leq R_{[|\mathcal{S}||\mathcal{A}|]}$, and the visitation probabilities are ordered as $P^{\pi}_{[1]} \leq P^{\pi}_{[2]} \leq \cdots \leq P^{\pi}_{[|\mathcal{S}||\mathcal{A}|]}$. We denote the order index of $R(s,a)$ as $I_r(s,a) \in [|\mathcal{S}||\mathcal{A}|]$ and the order index of $P^{\pi}(s,a)$ as $I_p(s,a) \in [|\mathcal{S}||\mathcal{A}|]$, such that $R_{[I_r(s,a)]} = R(s,a)$ and $P^{\pi}_{[I_p(s,a)]} = P^{\pi}(s,a)$. We first provide the definition of S-BLACK SWAN in case of discrete state and action space.

Definition 4 (S-BLACK SWAN - Discrete State and Action Space)**.** *Given distortion functions u, w and constants $C_{bs} \gg 0$ and $\epsilon_{bs} > 0$, if (s,a) satisfies:*

1. *(High-risk):* $R_{[I_r(s,a)]} - u^{-}(R_{[I_r(s,a)]}) < -C_{bs}$.
2. *(Rare):* $w^{-}\left(\sum_{j=1}^{I_p(s,a)} P^{\pi}_{[j]}\right) = w^{-}\left(\sum_{j=1}^{I_p(s,a)-1} P^{\pi}_{[j]}\right)$, *yet* $0 < P^{\pi}_{[I_p(s,a)]} < \epsilon_{bs}$.

then we define (s,a) as S-BLACK SWAN .

Definition 4 finally formalizes the informal concept of black swan events introduced in Section 3. The first property of Definition 4 identifies a *high-risk event* through value distortion. Specifically, if the agent perceives R optimistically, such that $R \ll u^{-}(R) < 0$, it is classified as a high-risk event (see Figure 1c). The second property characterizes a *rare event* through probability distortion, describing an S-BLACK SWAN event that occurs with a small probability in the real world $\left(0 < P^{\pi}_{[I_p(s,a)]} < \epsilon_{bs}\right)$, but is perceived by the agent as infeasible $\left(w^{-}\left(\sum_{j=1}^{I_p(s,a)} P^{\pi}_{[j]}\right) = w^{-}\left(\sum_{j=1}^{I_p(s,a)-1} P^{\pi}_{[j]}\right)\right)$ (See Figure 1d).

The constants C_{bs} and ϵ_{bs} in Definition 4 quantify the extent of distortion in the functions u and w, respectively. Intuitively, C_{bs} and ϵ_{bs} are directly related to the magnitude of the misperception gap between $\mathcal{M}$ and $\mathcal{M}^{\dagger}$, denoted by ϵ_r and ϵ_p. This relationship will be further formalized in Theorem 4. We now extend the definition of S-BLACK SWAN to continuous state and action spaces. Suppose the reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is bijective. Then, the probability $R^{-1} \circ P^{\pi} : \mathbb{R} \to [0,1]$ denotes the probability of a feasible reward induced by policy π, denoted as $\mathbb{P}_r$. We then have the following definition.

Definition 5 (S-BLACK SWAN - Continuous State and Action Space)**.** *Given distortion functions u, w and constants $C_{bs} \gg 0$ and $\epsilon_{bs} > 0$, if (s,a) satisfies:*

1. $R(s,a) - u^{-}(R(s,a)) < -C_{bs}$.
2. $\frac{dw^{-}(x)}{dx}\Big|_{x=F(R(s,a))} \cdot \mathbb{P}_r(r = R(s,a)) = 0$, *yet* $0 < \mathbb{P}_r(r = R(s,a)) < \epsilon_{bs}$,

where $F(r) := \int_{-\infty}^{r} d\mathbb{P}_r$ is the cumulative distribution of $\mathbb{P}_r$, then we define (s,a) as S-BLACK SWAN .

We then define the minimum probability of S-BLACK SWAN as $\epsilon^{\min}_{bs}$, denoted as $\epsilon^{\min}_{bs} := \min_{(s,a)} \mathbb{P}_r(r = R(s,a))$. Let $\mathcal{B}$ denote the collection of all S-BLACK SWAN . For given constants C_{bs} and ϵ_{bs}, we define the distortion functions w^{-} and u^{-} that result in $\mathcal{B} = \varnothing$ as $w^{-}_{\star}$ and $u^{-}_{\star}$, respectively. Intuitively, $w^{-}_{\star}$ and $u^{-}_{\star}$ represent a *safe* perception, meaning that if an agent perceives the world through those, then $\mathcal{B} = \varnothing$. However, it is important to note that $w^{-}_{\star}$ and $u^{-}_{\star}$ are not unique functions (see Figure 1d).

6.2 Theoretical Analysis of S-black swan

Subsection 6.2 explores the properties of S-BLACK SWAN , focusing on how their presence establishes a lower bound on policy performance (Theorem 4) and the timing of their occurrences (Theorem 5), laying the groundwork for future algorithm design. For further analysis, we assume the following.

Assumption 1 (Relative convexity)**.** *Assume* $u_\star^-(r) \leq u^-(r)$ *holds for* $r < 0$.

Assumption 1 ensures that a human (agent) with u^- perceives rewards more optimistically than one with $u_\star^-$ across all (s, a) pairs. This concept is well illustrated in Figure 1c, where the function $u^-(r) = r$ represents an *unbiased perception*, and deviations from this line indicate increasing reward distortion. In conjunction with Assumption 1, we introduce a proposition regarding S-BLACK SWAN , enabling interpretation within the reward space $[-R_{\max}, R_{\max}]$.

Proposition 2 (S-BLACK SWAN)**.** *Let the intersection of the functions* $r + C_{bs}$ *and* $u^-(r)$ *occur at* $r = -R_{bs}$ *(see Figure 1c). Under Assumption 1, if* $r(s,a) \in [-R_{\max}, -R_{bs}]$ *satisfies:*

1. $r - u^-(r) < -C_{bs}$,
2. $w^-(F(r)) = 0$, *with* $0 < F(r) < \epsilon_{bs}$,

then the (s, a) *is* S-BLACK SWAN .

A key insight from Proposition 2 is that as $u^-(r)$ approaches $u_\star^-(r)$, the approximation $-R_{bs} \to -R_{\max}$ occurs, finally leading to $|\mathcal{B}| \to 0$ since $|[-R_{\max}, -R_{bs}]| \to 0$ (see Figures 1c). In other words, Proposition 2 demonstrates that reducing the perception gap directly correlates with a decrease in $|\mathcal{B}|$.

Now, to provide an guideline for designing safe learning algorithms to prevent S-BLACK SWAN , it is crucial to quantify how the existence of S-BLACK SWAN leads to an inevitable deviation from the real-world optimal policy. We address this by analyzing how the misperception gap establishes a lower bound on the value function gap between the HMDP $\mathcal{M}^\dagger$ and the GMDP $\mathcal{M}$, as presented in the following theorem.

Theorem 4 (Convergence of value estimation gap but lower bound on value perception gap)**.** *Under Assumption 1, the asymptotic convergence of the value function estimation holds as follows,*

$$V^\pi_{\widehat{\mathcal{M}}^\dagger}(s) \to V^\pi_{\mathcal{M}^\dagger}(s) \quad a.s. \quad as \quad T \to \infty,\ \forall s, \pi \in \mathcal{S} \times \Pi. \tag{3}$$

However, under specific conditions on $\epsilon_{bs}, \epsilon_{bs}^{\min}, R_{bs}$*, the lower bound of value perception gap as follows.*

$$|V^\pi_{\mathcal{M}^\dagger}(s) - V^\pi_{\mathcal{M}}(s)| = \Omega\left(\frac{\left((R_{\max} - R_{bs})\,\epsilon_{bs}^{\min} - R_{bs}\epsilon_{bs}\right)(R_{\max} - R_{bs})C_{bs}}{R_{\max}^2}\right) \tag{4}$$

There are two key consequences of Theorem 4. First, Equation (3) demonstrates that the value estimation error converges to zero as the agent rolls out longer trajectories. However, Equation (4) reveals that the value perception gap has a non-zero lower bound, regardless of the horizon length. Equation (4) further indicates that if $u^-(x) \to u_\star^-(x)$ and $w^-(x) \to w_\star^-(x)$, then $R_{bs} \to R_{\max}$ and $\epsilon_{bs} \to 0$ (see Figures 1c and 1d), leading to the convergence of this lower bound to zero. Second, Equation (4) aligns with the intuition that greater distortion in reward perception (i.e., larger C_{bs}) and an increased number of S-BLACK SWAN (i.e., larger $(R_{\max} - R_{bs})$) coupled with a higher minimum probability of S-BLACK SWAN occurrence (i.e., larger $\epsilon_{bs}^{\min}$) result in a higher lower bound. Therefore, Theorem 4 concludes that even with zero estimation error, a lower bound on approximating the true value function remains, and this lower bound increases as C_{bs} and $\epsilon_{bs}^{\min}$ become more pronounced.

Then, the next natural question is *how to decrease that lower bound*, specifically, how can an agent can learn to self-correct toward a safe perception, i.e., $u^- \to u_\star^-$ and $w^- \to w_\star^-$. This question can be further refined to: *What is the probability of encountering* S-BLACK SWAN *if the agent takes t steps?* We address this under the assumption of non-zero one-step reachability, as follows.

Theorem 5 (S-BLACK SWAN hitting time). *Assume $\mathbb{P}_{\pi^\star}(s' \mid s) > 0$ for any $s, s' \in \mathcal{S}$, indicating that the one-step state reachability equipped with optimal policy is non-zero, and consider that one step corresponds to a unit time. Then, if the agent takes t steps such that $t \geq \log\left(\frac{\delta}{p_{\min}}\right) / \log(1 - p_{\max}) + 1$, where $p_{\min} = \frac{R_{\max} - R_{bs}}{2R_{\max}} \epsilon_{bs}^{\min}$ and $p_{\max} = \frac{R_{\max} - R_{bs}}{2R_{\max}} \epsilon_{bs}$, it will encounter* S-BLACK SWAN *with at least probability $\delta \in (0, 1]$.*

A key takeaway of Theorem 5 is determining how often a human should correct their internal perception. A large perception gap ($R_{\max} - R_{bs}$) and frequent occurrence of black swan events ($\epsilon_{bs}^{\min}$) require more frequent execution of the self-perception correction algorithm.

7 Related works: Necessity of S-black swan

This section discusses safe reinforcement learning (RL) algorithms, emphasizing the limitations of existing approaches in addressing black swan events and highlighting the need for a new perspective[4].

Safe RL algorithms are generally classified into three approaches: worst-case criterion, risk-sensitive criterion, and constrained criterion (Garcıa & Fernández, 2015). However, these approaches face significant limitations when dealing with black swan events. The worst-case criterion, which optimizes policy performance under the least favorable scenarios by maximizing the minimum return, becomes overly conservative when black swan events are considered, as they expand the uncertainty set $\mathcal{W}$, leading to impractical decisions such as avoiding all risky activities or adopting extreme safety measures (Heger, 1994; Coraluppi, 1997; Coraluppi & Marcus, 1999; 2000). Similarly, risk-sensitive algorithms, which incorporate a sensitivity factor to balance return maximization and risk management (Howard & Matheson, 1972; Chung & Sobel, 1987; Patek, 2001), are inadequate for handling black swan events because return variance, a commonly used risk measure, fails to account for the fat tails in distributions (Huisman et al., 1998; Bradley & Taqqu, 2003; Bubeck et al., 2013; Agrawal et al., 2021). Additionally, log-exponential utility functions, often associated with robust MDPs, do not effectively address the risks posed by black swans (Osogami, 2012; Moldovan & Abbeel, 2012; Leqi et al., 2019). The constrained criterion, which maximizes expected returns while meeting multiple utility constraints such as return variance or minimum thresholds (Geibel, 2006; Delage & Mannor, 2010; Ponda et al., 2013; Di Castro et al., 2012), also faces challenges with black swan events. These events complicate threshold selection, often necessitating more conservative policies, and suggest that constraints should be redefined to focus on state and action-specific risks rather than overall returns (Bagnell et al., 2001; Iyengar, 2005; Nilim & El Ghaoui, 2005; Wiesemann et al., 2013; Xu & Mannor, 2010). Furthermore, distributional RL is vulnerable to black swans, as extreme outliers in the reward distribution slow the convergence of the Bellman operator and provide a large suboptimality gap due to biased return expectations (Bellemare et al., 2017).

In summary, traditional risk criteria in RL are insufficient for managing the unique risks associated with black swan events, highlighting the need for novel approaches.

8 Conclusion

In conclusion, this paper redefines black swan events by introducing S-BLACK SWAN , highlighting that such high-risk, rare events can occur even in unchanging environments due to human misperception. We categorized and mathematically formalized these events, aiming to guide the development of algorithms that correct human perception to prevent such occurrences. This work opens the door for future research to enhance decision-making systems and reduce the impact of black swan events.

[4]Further details are in Appendix C, along with a discussion of CPT's application in risk analysis in Appendix D.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pp. 26–62. PMLR, 2021.

Md Akhtaruzzaman, Sabri Boubaker, and John W Goodell. Did the collapse of silicon valley bank catalyze financial contagion? *Finance Research Letters*, 56:104082, 2023.

Tatiana Antipova. Coronavirus pandemic as black swan event. In *International conference on integrated science*, pp. 356–366. Springer, 2020.

Michail Artemenko, Vladimir Budanov, and Nicoly Korenevskiy. Self-organizing algorithm for pilot modeling the reaction of society to the phenomenon of the black swan. In *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–7. IEEE, 2020.

J Andrew Bagnell, Andrew Y Ng, and Jeff G Schneider. Solving uncertain markov decision processes. 2001.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017.

BetterUp. The availability heuristic. `https://www.betterup.com/blog/the-availability-heuristic`, 2022. Accessed: 2024-05-12.

Samit Bhanja and Abhishek Das. A black swan event-based hybrid model for indian stock markets' trends prediction. *Innovations in Systems and Software Engineering*, 20(2):121–135, 2024.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, pp. 3003–3020. PMLR, 2023.

Brendan O Bradley and Murad S Taqqu. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pp. 35–103. Elsevier, 2003.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Jonathan Colaço Carr, Prakash Panangaden, and Doina Precup. Conditions on preference relations that guarantee the existence of optimal policies. In *International Conference on Artificial Intelligence and Statistics*, pp. 3916–3924. PMLR, 2024.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Kun-Jen Chung and Matthew J. Sobel. Discounted mdp's: distribution functions and exponential utility maximization. *Siam Journal on Control and Optimization*, 25:49–62, 1987. URL `https://api.semanticscholar.org/CorpusID:119760011`.

Stefano P Coraluppi and Steven I Marcus. Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes. *Automatica*, 35(2):301–309, 1999.

Stefano P Coraluppi and Steven I Marcus. Mixed risk-neutral/minimax control of discrete-time, finite-state markov decision processes. *IEEE Transactions on Automatic Control*, 45(3):528–532, 2000.

Stefano Paolo Coraluppi. *Optimal control of Markov decision processes for performance and robustness*. University of Maryland, College Park, 1997.

Dominic Danis, Parker Parmacek, David Dunajsky, and Bhaskar Ramasubramanian. Multi-agent reinforcement learning with prospect theory. In *2023 Proceedings of the Conference on Control and its Applications (CT)*, pp. 9–16. SIAM, 2023.

Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

Jinil Persis Devarajan, Arunmozhi Manimuthu, and V Raja Sreedharan. Healthcare operations and black swan event for covid-19 pandemic: A predictive analytics. *IEEE Transactions on Engineering Management*, 70(9):3229–3243, 2021.

Dotan Di Castro, Aviv Tamar, and Shie Mannor. Policy gradients with variance related risk criteria. *arXiv preprint arXiv:1206.6404*, 2012.

Stavros A Drakopoulos and Ioannis Theodossiou. Workers' risk underestimation and occupational health and safety regulation. *European Journal of Law and Economics*, 41:641–656, 2016.

Hein Fennema and Peter Wakker. Original and cumulative prospect theory: A discussion of empirical differences. *Journal of Behavioral Decision Making*, 10(1):53–64, 1997.

Michael J Fleming and Asani Sarkar. The failure resolution of lehman brothers. *Economic Policy Review, Forthcoming*, 2014.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Peter Geibel. Reinforcement learning for mdps with constraints. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*, pp. 646–653. Springer, 2006.

Peter Geibel and Fritz Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349, 2013.

Thomas Gilovich, Dale Griffin, and Daniel Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.

Abhijit Gosavi. Reinforcement learning for model building and variance-penalized control. In *Proceedings of the 2009 winter simulation conference (wsc)*, pp. 373–379. IEEE, 2009.

Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-based systems*, 212:106622, 2021.

Matthias Heger. Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 105–111. Elsevier, 1994.

Morgan Housel. Penguin, 2023.

Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

Ronald Huisman, Kees G Koedijk, and Rachel A Pownall. Var-x: Fat tails in financial risk management. *Journal of risk*, 1(1):47–61, 1998.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Nan Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

Cheng Jie, LA Prashanth, Michael Fu, Steve Marcus, and Csaba Szepesvári. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63(9):2867–2882, 2018.

Ming Jin. Preparing for black swans: The antifragility imperative for machine learning. *arXiv preprint arXiv:2405.11397*, 2024.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Liu Leqi, Adarsh Prasad, and Pradeep K Ravikumar. On human-aligned risk minimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

Benedetto De Martino, Dharshan Kumaran, Ben Seymour, and Raymond J. Dolan. Frames, biases, and rational decision-making in the human brain. *Science*, 313:684 – 687, 2006.

John Kwaku Mensah Mawutor. The failure of lehman brothers: causes, preventive measures and recommendations. *Research Journal of Finance and Accounting*, 5(4), 2014.

Larry McDonald and Patrick Robinson. *A colossal failure of common sense: The incredible inside story of the collapse of Lehman Brothers*. Random House, 2009.

Teodor Moldovan and Pieter Abbeel. Risk aversion in markov decision processes via near optimal chernoff bounds. *Advances in neural information processing systems*, 25, 2012.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Sina Nordhoff, John D. Lee, Simeon C. Calvert, Siri Berge, Marjan Hagenzieker, and Riender Happee. (mis-)use of standard autopilot and full self-driving (fsd) beta: Results from interviews with users of tesla's fsd beta. *Frontiers in Psychology*, 14, 2023. ISSN 1664-1078. URL `https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1101520`.

James J Opaluch and Kathleen Segerson. Rational roots of "irrational" behavior: new theories of economic decision-making. *Northeastern Journal of Agricultural and Resource Economics*, 18(2):81–95, 1989.

Takayuki Osogami. Robustness and risk-sensitivity in markov decision processes. *Advances in Neural Information Processing Systems*, 25, 2012.

Bhavana Pandit, Alex Albert, Yashwardhan Patil, and Ahmed Jalil Al-Bayati. Impact of safety climate on hazard recognition and safety risk perception. *Safety science*, 113:44–53, 2019.

Stephen D Patek. On terminating markov decision processes with a risk-averse objective function. *Automatica*, 37(9):1379–1386, 2001.

Matt Phillips. Gamestop's wild stock ride: How ai and social media drove a short squeeze. The New York Times Business, 2021. URL `https://www.nytimes.com/2021/01/29/business/gamestop-stock.html`. Accessed: 2024-08-19.

Silviu Pitis. Consistent aggregation of objectives with diverse time preferences requires non-markovian rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

Sameera S Ponda, Luke B Johnson, and Jonathan P How. Risk allocation strategies for distributed chance-constrained task allocation. In *2013 American Control Conference*, pp. 3230–3236. IEEE, 2013.

LA Prashanth, Cheng Jie, Michael Fu, Steve Marcus, and Csaba Szepesvári. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, pp. 1406–1415. PMLR, 2016.

SD Prestwich. Tuning forecasting algorithms for black swans. *IFAC-PapersOnLine*, 52(13):1496–1501, 2019.

Matihew Rabin. Risk aversion and expected-utility theory: A calibration theorem. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 241–252. World Scientific, 2013.

Lillian J Ratliff and Eric Mazumdar. Inverse risk-sensitive reinforcement learning. *IEEE Transactions on Automatic Control*, 65(3):1256–1263, 2019.

Paul Rogers. The cognitive psychology of lottery gambling: A theoretical review. *Journal of gambling studies*, 14(2):111–134, 1998.

Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

Mehran Shakerinava and Siamak Ravanbakhsh. Utility theory for sequential decision making. In *International Conference on Machine Learning*, pp. 19616–19625. PMLR, 2022.

Yun Shen, Michael J Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328, 2014.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. *arXiv preprint arXiv:1806.01492*, 2018.

Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020.

Herbert A Simon. Decision making: Rational, nonrational, and irrational. *Educational administration quarterly*, 29(3):392–411, 1993.

Philip Stafford. Citadel securities trading algorithm triggers market volatility. Financial Times Online, 2022. URL `https://www.ft.com/content/f53e3159-ab98-4926-ab41-63a577355825`. Accessed: 2024-08-19.

Robert Sugden. Rational choice: a survey of contributions from economics and philosophy. *The economic journal*, 101(407):751–785, 1991.

Peter Sunehag and Marcus Hutter. Axioms for rational reinforcement learning. In *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings 22*, pp. 338–352. Springer, 2011.

Peter Sunehag and Marcus Hutter. Rationality, optimism and guarantees in general reinforcement learning. *The Journal of Machine Learning Research*, 16(1):1345–1390, 2015.

Richard S Sutton. The reward hypothesis, 2004. *URL http://incompleteideas. net/rlai. cs. ualberta. ca/RLAI/rewardhypothesis. html*.

Richard S Sutton. The quest for a common model of the intelligent decision maker. *arXiv preprint arXiv:2202.13252*, 2022.

Hamdy A Taha. Operations research an introduction. 2007.

Nassim Nicholas Taleb. *The Black Swan:: The Impact of the Highly Improbable: With a new section:" On Robustness and Fragility"*, volume 2. Random house trade paperbacks, 2010.

Tesla. *Tesla AI Day*. 2021. URL `https://www.youtube.com/watch?v=j0z4FweCy4M`.

Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

Toni GLA van der Meer, Anne C Kroon, and Rens Vliegenthart. Do news media kill? how a biased news reality can overshadow real societal risks, the case of aviation and road traffic accidents. *Social forces*, 101(1):506–530, 2022.

Peter Vasterman, C Joris Yzermans, and Anja JE Dirkzwager. The role of the media and media hypes in the aftermath of disasters. *Epidemiologic reviews*, 27(1):107–114, 2005.

Morgenstern von Neumann. Theory of games and economic behaviour, 1944.

Maxime Wabartha, Audrey Durand, Vincent Francois-Lavet, and Joelle Pineau. Handling black swan events in deep learning with diversely extrapolated neural networks. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 2140–2147, 2021.

Anders AF Wahlberg and Lennart Sjoberg. Risk perception and the media. *Journal of risk research*, 3(1): 31–50, 2000.

Gregory Wheeler and G Wheeler. A review of the lottery paradox. *Probability and inference: Essays in honour of Henry E. Kyburg, Jr*, pp. 1–31, 2007.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Rosalind Wiggins, Thomas Piontek, and Andrew Metrick. The lehman brothers bankruptcy a: overview. *Yale program on financial stability case study*, 2014.

Paul D Witman, Jim Prior, Tracy Nickl, and Scott Mackelprang. Southwest airlines didn't crash, but it nearly fell apart.... In *Proceedings of the ISCAP Conference ISSN*, volume 2473, pp. 4901, 2023.

Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pp. 1–28, 2024.

Yuanyuan Zhang, Xiang Li, and Sini Guo. Portfolio selection problems with markowitz's mean–variance framework: a review of literature. *Fuzzy Optimization and Decision Making*, 17:125–158, 2018.