ChatGPT

Consequence Minimization: Formal Foundations and Models

Introduction

Consequence Minimization (CM) is the principle that an agent should **lexicographically** prioritize avoiding catastrophic outcomes before pursuing gains [1]. In other words, the agent first minimizes the probability (or severity) of entering a designated catastrophe set S_c (an absorbing "ruin" state such as death, bankruptcy, or system failure), and only once this risk is minimized does it maximize expected rewards or utility. This "safety-first" philosophy has deep roots: in ethics, Popper's negative utilitarianism argues we must minimize suffering before maximizing happiness [2], and in finance, Roy's safety-first criterion directs investors to minimize the probability of disaster before seeking profit [3]. CM formalizes this intuition in decision-theoretic terms as a lexicographic preference order. In this paper, we develop a rigorous foundation for CM across five objectives. We present axiomatic definitions and representation theorems distinguishing CM from traditional risk-aversion, formulate CM in constrained Markov decision processes (CMDPs) using viability kernels to characterize safe policies, analyze state-dependent policy inversions where CM prescribes **risk-seeking** behavior under unavoidable ruin, compare CM to minimax strategies in games, and situate CM within the broader literature on risk and uncertainty. The goal is a comprehensive, formal treatment showing that CM is a well-defined decision principle—one that is falsifiable, distinct from existing models, and explanatory of behaviors observed in high-stakes environments.

1. Axiomatic Foundations and Representation Theorem

Definition (Lexicographic Catastrophe-First Preference): Let outcomes include a *catastrophe set* S_c (undesirable absorbing outcomes) and *normal outcomes* $S_{\neg c}$. A CM agent's preference over lotteries (probability distributions on outcomes) is defined lexicographically [4]: for any two lotteries L and M, L is preferred to M if and only if (i) $P_L(S_c) < P_M(S_c)$ – i.e. L has a strictly smaller probability of yielding a catastrophic outcome than M – or (ii) $P_L(S_c) = P_M(S_c)$ and $\mathbb{E}[U(L)\mid \neg S_c] > \mathbb{E}[U(M)\mid \neg S_c]$, where $U(\cdot)$ is a baseline utility function evaluating non-catastrophic outcomes (e.g. monetary reward). In essence, the agent **first minimizes** the probability of catastrophe, and **only if catastrophe risks are equal** does it maximize the expected utility of the remaining outcomes. This induces a **lexicographic preference order** that is non-Archimedean (there is no finite trade-off that the agent will accept in exchange for an incremental increase in catastrophic risk) [2]. Formally, an infinitesimal ϵ increase in $P(S_c)$ cannot be compensated by any finite gain in expected non-catastrophic utility.

Representation Theorem: An agent's preferences satisfy Consequence Minimization if and only if there exists a lexicographic utility representation $(U_{\text{cat}}, U_{\text{norm}})$ such that $U_{\text{cat}}(x)\in\{0,1\}$ flags catastrophic outcomes ($U_{\text{cat}}=0$ for non-catastrophic, $U_{\text{cat}}=1$ for catastrophic), and $U_{\text{norm}}(x)$ is a standard utility function for outcomes, with the preference ordering defined by $L \prec M$ if and only if $\mathbb{E}[U_{\text{cat}}(L)] < \mathbb{E}[U_{\text{cat}}(M)]$, or equal and $\mathbb{E}[U_{\text{norm}}(L)] < \mathbb{E}[U_{\text{norm}}(M)]$. In other words, *there is no single real-valued utility that*

represents CM preferences, but they can be represented by a two-tier lexicographic utility vector. This reflects the failure of the Archimedean axiom in lexicographic orders [5] [6]. Intuitively, any attempt to encode CM into a one-dimensional concave utility (for example, by assigning an extremely large disutility to catastrophe) either breaks the expected utility framework or collapses to ordinary risk aversion if the penalty is finite. Indeed, lexicographic preferences can only be captured in an expected utility model by allowing *infinitely negative* utility for catastrophic outcomes – effectively treating them as having lexicographically higher importance than any finite reward [7]. This aligns with Popper's observation that from an ethical point of view, *"pain cannot be outweighed by pleasure"* and thus preventing the worst outcomes must come first [7].

CM vs. Risk-Aversion and Regret: It is critical to distinguish CM from traditional risk aversion. A *risk-averse* decision-maker typically maximizes **expected utility** under a concave utility function, placing extra weight on worst-case outcomes but still allowing trade-offs [8]. Such an agent has a *continuous* preference: a sufficiently large potential gain can justify a small increase in risk. In contrast, a CM agent has a *discontinuous* aversion to catastrophe: no finite gain is worth even a tiny increase in the probability of ruin. For example, a risk-averse individual might accept a 1-in-1,000,000 chance of death for a billion-dollar reward if their utility for money grows unbounded, whereas a strict CM agent would refuse **any** chance of death if a zero-risk alternative exists, no matter the reward. This stark difference means that CM choices generally **cannot be replicated by any concave utility model** unless that utility effectively assigns an infinite penalty to catastrophic outcomes (which standard expected utility theory disallows). To illustrate, consider two options: (A) a sure \$1 million, and (B) a 0.1% chance of \$1 **billion** with a 99.9% chance of \$0. An expected-value maximizer or even a moderately risk-averse agent might choose (B) because its expected value is \$1 million, equal to (A), and higher potential upside might tempt them. But if we define "catastrophe" as ending up with \$0 (impoverishment), a lexicographic CM agent will *never* choose (B) because (A) guarantees safety from ruin (no chance of \$0) whereas (B) carries a 99.9% chance of that disastrous outcome. No concave utility function $u(w)$ that is finite for $u(0)$ can reproduce this strict preference for (A) in the face of arbitrarily large rewards from (B); a sufficiently steep utility might *approximate* it, but if $u(0)$ is finite, then for a large enough gain G, $0.001 \times u(G)$ will eventually exceed $u(1{,}000{,}000)$, leading a risk-averse expected utility optimizer to pick (B). Thus, **a falsifiability test for CM** is to find choices where an individual consistently refuses trades that any finite-penalty expected utility model would accept. If an agent's choices in extreme trade-offs can always be explained by some concave utility curve (perhaps extremely risk-averse but still trading at some point), then CM as a separate principle collapses to conventional risk aversion. However, observing a **lexicographic threshold** – a point at which an agent stops trading risk for reward entirely – would support CM. Empirically, one could offer a decision-maker increasing reward multiples for a fixed tiny probability of ruin; a purely risk-averse (but not lexicographic) individual would eventually relent and take the gamble as the reward grows, whereas a true CM individual would *flat-out refuse any positive ruin probability*. Such behavior would indicate an underlying lexicographic preference that **cannot** be reduced to concave utility.

CM is also distinct from **minimax regret** (Savage's criterion). Minimax regret asks the decision-maker to minimize the maximum possible *regret* (the difference between the outcome and the best that could have been done in hindsight for each state) [9]. While this often leads to conservative choices, its logic is different: it focuses on avoiding the pain of having made the "wrong" decision relative to a counterfactual, rather than directly avoiding material catastrophe. For instance, a minimax-regret policy might tolerate a catastrophic outcome if in that state all options were catastrophic (hence no regret) – an outcome the CM agent would still abhor absolutely. Conversely, the CM agent cares nothing about foregone gains or comparative regret; it cares only about the **objective consequence** of catastrophe [10]. A concrete counterexample: suppose an investor has two strategies in an uncertain market. Strategy X can either yield

\$100 profit or \$0 (loss) depending on the market, while Strategy Y can yield \$1000 or \$0. If in the worst-case both X and Y result in \$0 (say, a market crash causes ruin for either choice), minimax regret will look at that state and see zero regret (since even the best strategy still got \$0). It might then choose the strategy with the better upside (Y) because in every state the regret is manageable. A lexicographic CM investor, however, might identify "\$0 = ruin" as the catastrophic outcome and ask which strategy has a lower probability of hitting \$0. If, for example, X has a 1% chance of yielding \$0 while Y has a 5% chance of \$0 (with their higher payoffs occurring the rest of the time), CM will **strictly prefer** X, even though in the worst-case state both lose everything (and hence Y's *regret* in that state is no worse). This shows CM and minimax regret can recommend different actions. In summary, CM cares about absolute outcomes (survival vs. ruin) [10], whereas regret-based criteria care about relative comparison between actions' outcomes.

Remark: If all choices available to an agent can be rationalized by some concave utility function – that is, the agent never insists on avoiding a risk that a sufficiently risk-averse utility could be indifferent to – then one cannot empirically distinguish CM from extreme risk aversion. Therefore, a *crucial test* for the CM hypothesis is the existence of decisions where the agent's indifference curves exhibit a sharp corner: no slope no matter the incentive. The absence of any such lexicographic behavior in practice would indicate that CM might not be a fundamentally new principle but rather an approximation to very strong risk aversion. On the other hand, widespread evidence of **no-compromise safety rules** (e.g. "we never fly if the aircraft has *any* chance of catastrophic failure above 10^{-9} per hour") supports the notion of lexicographic safety preferences beyond standard utility calculus.

2. Consequence MDPs and Viability Kernel Modeling

To examine CM in dynamic environments, we formalize it in the context of **constrained Markov Decision Processes (MDPs)** with an absorbing catastrophe state. Consider an MDP defined by (S, A, P, R) where S is the state space, A the action space, $P(s'|s,a)$ the transition probabilities, and $R(s,a)$ the reward function. Let $S_c \subset S$ be a set of *catastrophic (absorbing) states* such that once $s \in S_c$, the process terminates (ruin). All other states $S_{\neg c} = S\setminus S_c$ are non-terminal. We impose a *viability constraint*: the agent's primary objective is to **avoid ever entering** S_c. This naturally leads to the concept of a **viability kernel** from viability theory [11].

Definition (Viability Kernel): The viability kernel S_V is the set of all safe states from which there exists at least one policy π that can keep the state trajectory *forever* within $S_{\neg c}$ (never reaching S_c) [11]. Equivalently, S_V is the **maximal subset** of $S_{\neg c}$ such that for every state $s_0 \in S_V$, $\exists \pi$ with $P_\pi(\forall t: s_t \notin S_c \mid s_0)=1$. If the system starts in S_V, the agent has some strategy to survive indefinitely without catastrophe; if it starts outside S_V, ruin is inevitable under any policy [12]. This set can be characterized by a dynamic programming or fixed-point condition: S_V contains those states s for which there is at least one action a with $P(s' \in S_V \cup S_{\neg c} \mid s,a) = 1$ and $P(s' \in S_c \mid s,a) = 0$ – intuitively, from s the agent can choose an action that stays in $S_{\neg c}$ and lands in a state that remains viable. In practice, S_V is computed via *backward reachability*: start from the safe set and include states that have some action leading back into the set [11]. By definition, S_V enjoys a key property: it is **closed under the dynamics given a suitable policy**. If $s_0 \in S_V$, the agent can ensure $s_t \in S_V$ for all $t\ge0$ by following a viability policy. In contrast, any state $s \notin S_V$ is "doomed": even optimal control cannot avoid eventually hitting S_c (this is sometimes termed *inevitable failure* or an *unviable state* [13] [14]).

CM Policy in a CMDP: A CM-oriented agent in this MDP operates with a lexicographic objective: (1) maximize the probability of *survival* (never reaching S_c over an infinite horizon), and (2) subject to survival, maximize expected cumulative reward $E[\sum_{t=0}^{\infty} \gamma^t R(s_t,a_t)]$ (or another performance criterion) conditional on survival. The resulting optimal behavior is clear: **if the initial state $s_0 \in S_V$ (survival is possible), the agent will adopt a *survival-first policy*** that keeps the state within S_V at all costs, even if it sacrifices some reward. Only among equally survival-preserving strategies would it then compare rewards. If $s_0 \notin S_V$ (survival impossible), the lexicographic criterion cannot get probability of ruin to zero; all policies have $P(\text{eventual }S_c)=1$. In that case, as we explore in Section 3, the CM agent will choose the policy that **minimizes** $P(S_c)$ (maximizes survival time or probability of very long survival) and then optimizes rewards as a second priority. Notably, if $s_0 \in S_V$, a CM-optimal policy ensures *almost sure* avoidance of catastrophe (by definition of viability kernel) [15], whereas a standard expected-value (EV) maximizing policy might accept a small probability of visiting S_c if it yields higher reward. Thus **CM and EV policies diverge whenever avoiding a catastrophic tail risk requires sacrificing expected reward**.

Hypothesis: In environments with **heavy-tailed loss distributions**, a CM policy achieves higher *conditional* performance and longer survival than an EV-maximizing policy. Heavy-tailed risks are those with low-probability but extreme losses (e.g. power-law tailed outcomes). An EV-maximizer might chase high returns at the cost of a small probability of an enormous loss, because the expected value calculation underweights extremely rare events or even fails if the mean is infinite. A CM agent effectively places an infinite disutility on those tail losses, thus avoiding strategies that court disaster even if they have high average yield. Over repeated trials or a long time horizon, the EV agent is likely to eventually hit a catastrophic loss (risk of ruin accumulates over time [16]), whereas the CM agent survives by **staying in the viability region** and forgoing ruinous bets. This leads to a higher *conditional* reward (reward given that the agent is still alive) and a greater expected lifespan. As Taleb succinctly put it, *"never cross a river if it is on average four feet deep"* [17] – the EV optimizer might cross for speed (high average gain) ignoring that 1% of the time the river is 8 feet deep (catastrophe by drowning), whereas the CM optimizer will take a slower safe route, maximizing the probability of reaching the other side. Over many repetitions, the *time-average outcome* for the CM policy can far exceed that of the EV policy, because the latter eventually hits a ruin that resets its gains to zero [18] [16]. In finance terms, **survivorship** is a precondition to enjoy long-term expected returns [19]. Investors like Warren Buffett explicitly prioritize avoiding ruin ("ensure survival first") before seeking profit [20], reflecting a CM philosophy that "the presence of ruin invalidates cost-benefit analysis" [17] in the long run.

We formalize this intuition with a simple model: imagine an i.i.d. sequence of investments each yielding +\$1 with probability p and a catastrophic $-\$W$ (large loss) with probability $1-p$. An EV-maximizer would invest if $p \cdot 1 + (1-p)\cdot(-W) > 0$, i.e. if $p > \frac{W}{W+1}$, even when W is enormous. A CM agent sets a threshold on W relative to the agent's tolerance – effectively W is "infinite" in disutility if it causes ruin. If W represents bankruptcy (say W equals your entire capital), then no matter how small $(1-p)$ is, repetition will eventually bankrupt the EV investor with probability 1 [21] [16]. The CM investor would reject the investment unless $1-p$ is truly zero (or below some lexicographic epsilon representing negligible risk). Thus, **under heavy-tailed or repeated-risk scenarios, CM policies dominate in terms of long-term survival and realized reward conditional on survival** [22] [20]. We expect that if losses follow a power-law distribution, an EV policy might maximize short-term gain but suffer a massive loss within a finite horizon, whereas a CM policy avoids the tail and accumulates moderate gains indefinitely, leading to a higher median outcome and higher expected *survival-conditioned* reward. This can be tested via simulation.

Simulation Framework (Pseudo-code): To empirically compare CM and EV strategies, one can simulate a CMDP with an absorbing failure state. For example, consider a gambler's wealth process with two actions each turn: a *safe* action yielding a small certain reward, and a *risky* action yielding a higher reward with probability p but catastrophic loss (ruin) with probability $q=1-p$. The EV agent picks the action with higher $p \times \text{gain} + q \times \text{loss}$, while the CM agent picks the action that minimizes q (if q differs) and only if q is equal does it consider the gain. Pseudocode:

```
initialize wealth = W0
for t in 1..T:
    if agent == EV:
        choose action = argmax_a [p(a)*gain(a) + q(a)*loss(a)]  # maximize
expected value
    if agent == CM:
        choose action = argmin_a [q(a)]  # minimize failure probability first
        if multiple actions tie in q, choose among those by highest expected
gain
    simulate outcome; update wealth or enter failure state; if failure, break
```

By running many simulations, we can estimate metrics: (i) frequency of ruin, (ii) average reward accumulated by time T conditional on not having been ruined, and (iii) distribution of wealth over time. We hypothesize the CM policy will show **lower ruin rates** and **higher median/conditional wealth** than the EV policy, especially as T grows or as the loss tail becomes heavier. Preliminary toy experiments support this: for instance, with $p=0.99$, gain = +1, loss = -100 (heavy-tailed downside), an EV agent always takes the risky bet (expected value $0.99(1) + 0.01(-100) = -0.01$, oops – a rational EV agent would actually *avoid* a negative-EV bet in this simple case, so let's choose a slightly different example). Let's adjust to $p=0.999$, gain=+1, loss=-1000; EV sees expected value $\approx +0.999 - 0.001*1000 = -0.001$, so actually still negative. We need an example where EV would take the bet: say $p=0.9999$, gain=+1, loss=-10000, then EV sees $0.9999(1) - 0.0001(10000) = +0$ (break-even). If $p=0.99991$, EV sees positive +0.0001 expected gain and will take it. Over 10000 rounds, the EV gambler will almost surely go broke (each round 0.01% chance, but over many independent rounds ruin probability $\approx 1 - 0.9999^{10000} \approx 63\%$). The CM gambler* never* takes that bet, so they steadily earn 0 each round (foregoing the tiny edge) but survive with certainty. By round 10000, the EV gambler has a significant chance of ruin and an equal or lower wealth on average (since one catastrophic loss wipes out all the small gains). This simplistic simulation encapsulates how a long-term perspective (many repetitions) flips the desirability of heavy-tail risks and shows CM's advantage in survival and conditional returns [16]. In more complex MDPs, one would observe similarly that CM policies avoid entering states from which ruin is likely (staying within the viability kernel), whereas EV policies might venture out for reward and occasionally crash. Quantitative evaluation would compare survival curves and reward totals.

Viability Kernel Properties: We formally state a few properties of the viability kernel in CMDPs, which underpin the above reasoning:

- *Infinite-horizon survival:* If $s_0 \in S_V$, there exists a policy π such that for all $t\ge0$, $P_\pi(s_t \in S_c) = 0$. Conversely, if $s_0 \notin S_V$, then $\forall \pi,\; P_\pi(\exists t: s_t \in S_c) = 1$. In words, S_V is the set of states with **survival probability 1**, and its complement has survival

probability 0 under any policy [15]. This dichotomy justifies treating S_V as the "safe set" in lexicographic optimization.

- *Maximality:* S_V is maximal w.r.t. set inclusion: if any state outside S_V had a policy to avoid S_c, it would by definition be included. Thus S_V can be computed as the **fixed point** of an operator $\mathcal{F}(X) = {s: \exists a \text{ with } P(s' \notin X \cup S_c \mid s,a)=0 \text{ and } P(s' \in S_c \mid s,a)=0}$; starting with $X=S_{\neg c}$ and iterating $\mathcal{F}$ yields S_V. This is analogous to backward induction on an absorbing set.

- *Policy uniqueness:* Generally, there may be many policies that keep the process in S_V. However, a **reward-maximizing CM policy** will choose among them those actions that also maximize long-term reward. If the reward structure is such that a *deterministic* policy can be optimal, one can characterize an **optimal viability policy** via a Bellman-like equation on S_V: e.g., $V(s) = \max_{a: P(s' \in S_c| s,a)=0} \sum_{s'\in S_V}P(s'|s,a)V(s') + R(s,a)$, which is just a standard optimality equation constrained to S_V. This highlights that within the viability kernel the agent still optimizes performance, but never at the expense of leaving S_V.

- *Heavy-tail implication:* If transitions or rewards have heavy-tailed distributions (e.g., a small probability of a very large loss that leads to S_c), the viability kernel might exclude states that an EV optimizer would willingly occupy. For instance, a state that offers a high reward action but with a tiny probability of immediate ruin might be *excluded* from S_V if no alternative action in that state can avoid the ruin chance. The EV agent might go to that state for the high reward, but a CM agent will avoid it entirely. In this sense, S_V acts as a shield against tail risks: only states that allow *risk-free control* (or risk below some threshold if we extend to ϵ-viability) are considered viable.

In summary, modeling CM in CMDPs via the viability kernel and lexicographic criteria yields policies focused on **survival**. Our hypothesis that CM achieves better survival and conditional rewards under heavy-tailed risk can be tested by simulations or analysis of toy models, and it underscores a key message: **maximizing expected value is not always optimal for long-term welfare when rare disasters loom** [22] [20].

3. State-Dependent Inversion and Risk-Seeking Behavior

A striking implication of Consequence Minimization is that it can recommend **risk-seeking behavior in desperate situations**. This runs counter to the usual association of CM with extreme caution. The logic is: *when all "safe" policies guarantee eventual ruin, the lexicographic rule advises choosing the policy that offers even a sliver of hope of avoiding ruin*. In other words, if catastrophe is certain under cautious actions, a CM agent will **invert** its attitude and take on risk in order to *create a path to survival with nonzero probability*. Formally, if the initial state s_0 lies outside the viability kernel ($s_0 \notin S_V$), then $\forall \pi,\; P_\pi(\text{ruin})=1$. Among such states, define the *survival probability* of a policy π as $P_\pi(\text{avoid }S_c \text{ up to time } T)$ as $T\to\infty$ (which will tend to 0 for all π, but the rate differs). A CM agent in this scenario chooses $\pi^ = *arg\max*\pi P*\pi(\text{survival for as long as possible})$. Often, this is the policy that maximizes the probability of* eventually *reaching some region S_V' where conditional survival is higher (e.g. reaching a "lifeline" state even if it requires a risky leap). Such a policy may involve deliberately increasing short-term risk or outcome variance – essentially gambling – because a riskier strategy might by luck reach a safe harbor that cautious play would never attain. This is analogous to a gambler's "double-or-nothing" strategy*: if going slow guarantees losing (e.g. your chips will dwindle to zero), sometimes the optimal choice is to bet big and pray, since cautious betting has probability 1 of ruin [21].

Theoretical Construction: Consider a simple CMDP with states ${s_0, s_{\text{win}}, s_{\text{lose}}}$. $s_{\text{lose}} \in S_c$ is ruin (absorbing failure). From s_0 the agent has two actions: (A) a **safe** action that transitions to s_{lose} with probability 1 (a slow guaranteed decline), and (B) a **risky** action that with probability p jumps to s_{win} (a permanently safe state with no further risk) and with probability $(1-p)$ goes to s_{lose} immediately. Imagine p is small, say 10%. Thus, action A yields certain ruin; action B yields ruin 90% of the time but a 10% chance of eternal safety. A risk-averse or timid agent might erroneously view A as "safer" because it has no variability (it doesn't appreciate that certain ruin is worse than 90% ruin). But a lexicographic CM agent will compare the *probabilities of catastrophe*: action A has 100% chance of catastrophe, action B has 90%. **CM thus strictly prefers the risky action B** because it offers a non-zero chance of avoiding S_c entirely. This is a case of *risk-seeking under duress*. Once the agent takes action B, if it reaches s_{win}, it will then revert to risk-averse mode (since now survival is assured, it will just maximize reward). This toy example shows how **state-dependent** a CM policy is – the same agent can be extremely cautious in one scenario and extremely bold in another, purely based on whether caution can achieve survival. Indeed, psychologists have observed exactly this pattern in humans: people become risk-seeking when faced with a sure loss or disaster, a phenomenon predicted by prospect theory as well (the "domain of losses" risk-seeking) [23] . CM provides a normative rationale for that behavior: when **the status quo is catastrophic**, the only rational choice is to gamble on a long-shot that might escape the catastrophe [23] . As a proverb says, *"a drowning man will clutch at a straw."*

We can formalize a **theorem** in this vein:

Theorem: *If a state s is such that $s \notin S_V$ but there exists an action a leading to a successor state s' that has a higher survival probability than s (for example $P(s' \in S_V) > 0$), then the CM-optimal policy at s will choose that action a even if it entails a lower immediate reward or a higher variance of outcomes than other actions.* In particular, if one action guarantees immediate absorption in S_c (certain ruin) and another action yields S_c with probability $q < 1$ (some chance of avoiding ruin), the latter is strictly preferred by CM regardless of rewards. This can be proven directly from the lexicographic definition: minimizing $P(S_c)$ dominates all other criteria, so any action offering a smaller $P(S_c)$ in the long run is preferred.

Proof Sketch: Assume action a_1 results in ruin with probability 1 (either immediately or eventually) and action a_2 results in ruin with probability $q<1$. Let π_1, π_2 be policies that start by taking a_1 or a_2 from s and then follow some continuation. For any continuation, the probability of eventual ruin starting with a_1 is 1, while for a_2 it is q (possibly plus whatever risk in the continuation, but assuming continuation policies are chosen optimally, if s' reached by a_2 is in or closer to viability, that continuation will minimize further risk). Thus $P_{\pi_1}(S_c) - P_{\pi_2}(S_c) = 1 - q > 0$. By lexicographic preference, $\pi_2 \succ \pi_1$ regardless of expected rewards. More generally, if all policies lead to ruin a.s., pick the one that maximizes the *survival function* $P(\text{survive at least } t)$ for all t. A policy that delays ruin longer (stochastically) is preferred, since at any finite time horizon T, it has lower probability of having hit S_c by that time than any other policy. In the limit $T \to \infty$ all reach ruin, but a CM agent effectively cares about pushing that event as far out as possible (since it lexicographically minimizes the *hazard rate* at each instant). This criterion can induce risk-seeking: often the way to *delay* ruin is to take risky maneuvers that might either rescue the situation or crash immediately, rather than slow decline. For example, in a resource depletion scenario, spending resources on a high-risk/high-reward gamble that could replenish reserves is preferred to simply consuming the remaining resources until they run out.

Variance-Seeking Optimality under Boundary Conditions: We highlight that in boundary conditions (when an agent is at the brink of ruin), **higher variance actions can dominate lower variance ones** under CM – the opposite of what a risk-averse utility maximizer would prefer. This is reminiscent of the well-known result in gambler's ruin problems: if a gambler is behind and will surely lose by betting small, an optimal strategy to maximize probability of winning is to bet as boldly as possible (maximize variance of outcomes) [24]. In technical terms, if all policies have expected value below zero (a losing game) so ruin is a matter of time, the one that **maximizes the probability of ever hitting a big win before ruin** is lexicographically best. This often corresponds to **"variance-increasing" strategies** – e.g. betting one's entire bankroll in one shot to either double it (and possibly reach a safer zone) or lose it, can yield a higher survival probability (however defined, say reaching a target wealth) than slowly bleeding away with many small bets [25]. The CM agent facing such a scenario will essentially exhibit **reversal of risk attitude**: extremely risk-seeking in the loss domain (because only a lucky upside can save it) and risk-averse in the gain/safe domain. This aligns with patterns noted in behavioral studies: people take wild risks when they feel they have "nothing to lose" or when failure is certain otherwise [23]. Our framework provides a rational interpretation: those risks are *instrumental* to possibly avoid the catastrophe.

Falsification Test – When Risk-Seeking Fails: Is it possible that a risk-averse policy could ever outperform a risk-seeking one in terms of survival, even when all policies eventually lead to ruin? If so, that would falsify a simplistic reading of CM's prescription. One could construct environments where taking big risks actually shortens the horizon more than a slow decline. For example, suppose there is *no possible escape* and the only difference between policies is how quickly they lead to ruin. If a high-variance policy has a chance to end things immediately (with no chance of rescue) while a low-variance policy guarantees, say, 100 time steps before the end, a CM agent might actually favor the policy that *maximizes the expected time until ruin*, not just the probability of eventual escape (since escape probability is zero for all). In such a scenario, risk-seeking that increases the *volatility* of the time of ruin without increasing the chance of avoiding it is not beneficial. To test CM's predictions, one could identify scenarios where an agent faces a series of losses and only risky bets can potentially recover losses. **CM predicts** the agent will take the risky bets as long as there is any hope of recovery. However, if we find environments where human or rational agents still avoid gambles even when doomed (perhaps because they prefer a slow demise to a fast one, valuing short-term outcomes or utility along the way), that might contradict pure CM. For instance, an agent might prefer to "die gracefully" later rather than thrash wildly and die now – which implies some utility for the *manner* or *time* of downfall, outside CM's lexicographic scope. In formal terms, one could extend the lexicographic preference to include a second-level criterion of *maximizing expected life span* if survival probability is zero. A CM agent so extended would choose the policy that maximizes $E[\text{time to ruin}]$ when ruin is inevitable. In our simple model, that would mean if a risky policy yields ruin in either 1 or ∞ steps (with some probabilities) and a safe policy yields ruin deterministically in 100 steps, it depends on whether the risky policy's small chance of ∞ (escape) outweighs the guaranteed 100 steps. Pure CM says any nonzero ∞ is lexicographically better than a sure finite life – hence risky. But if escape is truly impossible (both have 0% escape), then it would compare expected horizons. Thus, a falsifying scenario for naive CM would be if agents consistently choose longer certain horizons over shorter probabilistic ones *even when both have 0 survival probability*. However, this may not falsify CM so much as refine it: CM could be formulated to lexicographically prioritize (1) survival probability, then (2) conditional expected life span or *viability time*. Indeed, real-world prudent decision-makers often do exactly this: if ultimate catastrophe is unavoidable, they choose options that delay it the longest (e.g. terminally ill patients choosing treatments that extend life by some time versus high-risk surgeries that could either cure or kill immediately, depending on their outlook).

In summary, **CM prescribes risk-seeking in no-win scenarios** as a rational strategy to *"roll the dice" for a chance of survival*, a principle that differentiates it sharply from traditional risk aversion. This state-dependent inversion can be empirically observed and aligns with known behavior patterns [23]. It underscores that CM is not simply "always play safe" – it is "play safe when safety is attainable, but if ruin is assured, take the gamble that might just avert it." This nuanced guideline is part of what makes CM a compelling principle across domains (from evolutionary biology – e.g. an animal cornered by a predator will take extreme actions – to finance and AI).

4. Connections to Minimax/Maximin and Limits in Games

Consequence Minimization bears a philosophical resemblance to the **maximin (minimax)** principle in game theory and robust decision-making: both prioritize the worst-case outcome. However, CM is not equivalent to maximin in general, except under specific adversarial conditions. Here we explore when CM *collapses* to a standard minimax strategy (and when it doesn't), by comparing their prescriptions in different environments.

In a **zero-sum adversarial game**, the maximin criterion advises choosing a strategy that *maximizes the minimum payoff* the opponent can force. If we interpret "catastrophic outcome" as any payoff below some disaster threshold, a maximin player will avoid strategies that allow the opponent to push the payoff below that threshold. In effect, against a purely adversarial nature (worst-case thinking), a CM agent and a maximin agent behave similarly: they both evaluate actions by their worst possible consequence. For example, consider a two-player zero-sum game with payoff matrix for our player. If one column (opponent's action) yields a catastrophic loss for one of our strategies, a maximin player will never choose that strategy because the opponent can exploit it to give us catastrophe. A CM player likewise won't choose it because it has a nonzero probability (in fact, certainty if the opponent plays that column) of catastrophe. In such strictly adversarial settings, **any probability of catastrophe effectively becomes a worst-case outcome** since a rational adversary will steer the outcome there. Thus, **CM reduces to a maximin strategy**: minimize the probability of catastrophe translates to "assume the adversary will hit you with catastrophe if possible, so avoid any strategy that allows it" – which is exactly what minimax (with appropriate payoff coding) would do. In game-theoretic terms, if the opponent is malicious and all uncertainty is adversarial, the lexicographic probability criterion is equivalent to treating catastrophe as a payoff of $-\infty$ and using maximin. This means that in zero-sum games, CM offers no new strategic insight beyond the classical minimax solution: both will pick a "safe" equilibrium strategy that guarantees the highest payoff in the worst case [26].

However, **in stochastic or cooperative environments, CM and minimax diverge**. The minimax strategy is often overly conservative when probabilities are involved. Minimax imagines an adversary even when none exists – it is blind to the likelihood of outcomes, focusing only on worst-case *payoffs*. CM, by contrast, focuses on worst-case *probabilities* of disaster (lexicographically minimizing that) but not necessarily worst-case payoffs otherwise. Consider a game against nature where nature is not adversarial but random with known probabilities. A minimax policy would treat nature as if it chooses the worst outcome for the agent each time; a CM policy would instead minimize *probability* of disaster but might accept a small probability if it's sufficiently small and there is compensating expected reward (if no zero-risk strategy exists). In fact, if no strategy can avoid catastrophe with probability 1, minimax would consider all strategies "worst-case = catastrophe" and be indifferent (or use some secondary criterion like maximizing the payoff in the catastrophic state perhaps). CM in that situation differentiates between strategies by their *degree* of risk (probability of catastrophe). For a concrete example, suppose an agent has two options in a probabilistic

scenario (not adversarial): Option X yields a catastrophic loss with probability 0.01 and a good outcome otherwise; Option Y yields a catastrophic loss with probability 0.001 but a moderate outcome otherwise (slightly lower expected value than X). A minimax decision-maker who only sees that both have a possible catastrophic outcome might view the worst-case payoff of both as the same (catastrophe) and be indifferent or require a tie-break by expected payoff – possibly choosing X if its good outcome is much better. A CM decision-maker will clearly prefer Y, because it has an order-of-magnitude lower chance of catastrophe (0.1% vs 1%). This illustrates how **CM "breaks ties" that minimax would not**: by incorporating probabilities lexicographically, it distinguishes degrees of risk whereas pure minimax does not.

We can construct a simple game matrix to see divergence:

Strategy (Player) vs Nature's move	Nature: Good (90%)	Nature: Bad (10%)
Action A (cautious)	+5 (gain)	-50 (catastrophic loss)
Action B (risky)	+20 (gain)	-50 (catastrophic loss)

Nature's "moves" Good or Bad occur with indicated probabilities (not adversarially chosen but by chance). Both actions A and B have a worst-case payoff of -50 (disaster). A minimax agent treats this like an adversarial game where Nature could choose Bad intentionally: thus the worst-case outcome of either action is -50, and the minimax value is -50. The minimax criterion would consider both A and B equally (both allow the worst-case -50). If forced to pick, a maximin agent might be indifferent or might incorporate a secondary criterion like maximize the payoff in the worst case (which are equal) or something arbitrary. Meanwhile, a CM agent looks at **probability** of hitting -50: for Action A, $P(\text{catastrophe}) = 0.1$ (10% if Bad happens), for Action B, presumably also 10% here since both have -50 in Bad scenario. So in this toy setup, both have equal probability of catastrophe (since both yield catastrophe only if the 10% Bad state occurs). If that's the case, then CM would go to the second criterion – expected gain – and prefer B (higher expected value). So in this particular example, CM would end up choosing the riskier high-payoff action B, which coincides with expected value choice, whereas minimax was indifferent. This shows that *when probabilities are exogenous and not under an adversary's control*, CM is *less conservative* than minimax: it doesn't punish a strategy for a bad outcome unless that outcome has higher probability. If we tweak the example so that actions differ in catastrophe probabilities, say:

- Action A: 5% chance of -50, otherwise +5.
- Action B: 10% chance of -50, otherwise +20.

Now, a CM agent prefers A (because 5% < 10% chance of catastrophe, lexicographically), whereas an EV agent might prefer B for higher average (EV_A = 0.95*5* + *0.05*(-50) = +4.75 - 2.5 = +2.25, EV_B = 0.90*20* + *0.10*(-50) = +18 - 5 = +13, so EV loves B). A minimax agent again only sees that both have -50 worst-case, thus no distinction on worst-case outcome; it might then consider something like "maximin regret" or just remain indifferent. In practice, minimax would require a criterion for mixed strategies perhaps, but since Nature isn't really an adversary here, minimax is an ill-defined approach. This highlights that **CM diverges strongly from minimax in non-zero-sum, stochastic environments**. Minimax is generally too pessimistic (it ignores the probability that the worst case is rare), whereas CM allows extremely low-probability catastrophes if they are necessary for higher reward – but only up to the point that no alternative has a lower probability.

In **cooperative or semi-cooperative games**, where another player might not be trying to hurt you, a pure minimax strategy would be unnecessarily cautious and possibly suboptimal. CM would instead focus on negotiating or planning to ensure no catastrophic outcomes occur (e.g. through trust-building or safeguards), but it wouldn't treat the other player as an adversary if evidence suggests otherwise. This means in games where some outcomes are catastrophic to one player, CM might encourage strategies that avoid mutual disaster (in a way aligning with Pareto optimal safety), whereas minimax might unilaterally avoid any vulnerability even if the other player is well-intentioned.

Constructing Divergent Example: Imagine a *coordination game* between two players where one particular outcome is disastrous for both (say, a mis-coordination). If both play conservatively (minimax style), they avoid disaster but also miss out on high payoff coordination. A CM player might reason: the other player likely also wants to avoid disaster, so as long as we choose a strategy that gives a tiny risk of disaster but a huge gain, the probability of disaster might actually be low (assuming some level of trust or rational behavior from the other side). If that probability is sufficiently low, CM could endorse the risk for the chance of mutual gain. Minimax would not – it would stick to the safe equilibrium that yields lower payoffs but zero chance of disaster. Thus in a non-adversarial context, CM can support more cooperative, higher-yield strategies than minimax, provided the catastrophic outcomes are sufficiently unlikely. In summary, **minimax is a special extreme of CM** that corresponds to the case where *any* possibility of catastrophe is treated as certainty (the worst-case will happen). CM generalizes this by acknowledging probabilities: it seeks to drive that possibility to zero if it can, but doesn't assume an adversary actively ensuring the worst happens if there is some probability otherwise.

One can also connect CM to concepts in **robust optimization** and **maxmin expected utility**. For instance, Gilboa and Schmeidler's maxmin expected utility (1989) involves an agent who maximizes the minimum expected utility across a set of possible probability distributions (often interpreted as ambiguity aversion). If the set of distributions is large and includes some that concentrate on catastrophic outcomes, a maxmin-EU decision might avoid those acts altogether. CM is somewhat different: rather than multiple priors, it's a single known probabilistic model but with a lexicographic utility (infinite disutility for catastrophe). It will behave similarly to a maxmin expected utility optimizer in that it won't accept any strategy that in *any* scenario leads to catastrophe with significant probability. However, if there is uncertainty about probabilities (ambiguity), a CM agent might effectively adopt a robust approach (assuming nature "chooses" the worst distribution for catastrophe probability unless proven otherwise). This shades into ambiguity aversion: an ambiguity-averse agent might overweight the probability of worst outcomes, similar to CM's focus on worst-case. But ambiguity aversion typically involves an attitude of preferring known risks over unknown [27], whereas CM is about the outcomes themselves rather than uncertainty about probabilities.

To pinpoint **limits of CM**: If applied naively in multi-agent contexts, CM could be *overly* cautious or even pathological. For example, if two players both follow CM and each is extremely fearful of catastrophic retaliation by the other, they may end up in a mutually bad equilibrium (think of two countries each so afraid of worst-case war that they engage in extreme arms races or preemptive strikes — ironically creating the catastrophic risk they sought to avoid). Real-world strategic interaction may require nuances beyond lexicographic safety (like signaling and commitment to reassure that certain catastrophic actions are off the table). Thus, CM must be contextualized: it provides a rational baseline ("never allow outcomes that spell doom"), but in some interactive settings the perception of what is "catastrophe" or how likely it is can itself be strategic.

In summary, **CM aligns with minimax in adversarial zero-sum settings**, essentially reproducing the same safe strategy as classical game theory (since any nonzero catastrophe chance will be exploited by an adversary). But **in stochastic or mixed-motive situations, CM and minimax can recommend different strategies**: CM will sometimes take calculated tiny risks for large gains (something minimax forbids), and conversely CM will distinguish between 0.1% vs 1% vs 5% catastrophe probabilities whereas pure minimax treats anything >0 as equally bad. This makes CM a more flexible and often less overly-conservative criterion than maximin, while still fundamentally prioritizing safety. It stands in contrast to minimax regret as well – CM cares about absolute outcomes, not regret, and to pure risk-neutral EV optimization – CM cares about probabilities, not just averages.

5. Comparative Literature Review and Conceptual Distinctions

Consequence Minimization is rooted in a rich intellectual tradition spanning philosophy, economics, psychology, and decision theory. To appreciate CM's unique stance, it is helpful to compare it against several related concepts:

- **Traditional Risk Aversion (Concave Utility):** Risk aversion describes preferring a sure outcome over a gamble of equal or even higher expected value [8]. Technically, it's modeled by concave utility functions (diminishing marginal utility of wealth) which cause decision-makers to weigh uncertain losses more heavily than gains [8]. While both risk-averse agents and CM agents avoid risk, the crucial difference is one of **degree and trade-offs**. A concave-utility maximizer will accept some small probability of a large loss if the compensating potential gain is sufficient (since utility, though concave, is finite). A CM agent will *not* – it effectively behaves as if the disutility of a "ruin" loss is infinite or lexicographically dominant. In economic terms, standard risk aversion smooths outcomes but *does not* impose an absolute lexicographic priority on avoiding the worst-case. For instance, an Arrow-Pratt risk-averse investor might buy insurance or diversify to reduce risk, but could still invest in a venture with 1% chance of bankruptcy if the other 99% outcomes are lucrative enough. A CM-oriented investor would not, as long as any safer investment exists. Thus CM can be seen as an extreme form of risk-aversion that violates the continuity/Archimedean axiom of expected utility [5]. It is "safety-first" rather than "risk-smoothing." Historically, the idea of lexicographic safety was foreshadowed by **A.D. Roy's Safety-First Criterion (1952)** in portfolio selection, which explicitly said investors first minimize the probability of disaster (portfolio return below some threshold) and only then consider returns [3]. Roy's criterion and CM share this lexicographic structure, whereas the classical Markowitz mean-variance approach or utility theory trade off risk and reward continuously. Another related concept is the **Precautionary Principle** in policy-making: when an action could lead to catastrophic outcomes, even with low probability, a precautionary (CM-like) approach says to avoid that action unless the risk can be reduced – effectively giving infinite weight to worst-case scenarios in critical domains (like nuclear safety or biosecurity). Risk aversion alone might not justify extremely costly prevention of ultra-rare events, but CM reasoning would, because even an ultra-rare existential catastrophe is lexicographically worse than any benefit.

- **Ambiguity Aversion:** Ambiguity aversion (pioneered by Daniel Ellsberg) is the preference for known probabilities over unknown probabilities [27]. For example, many people prefer a 50% chance of \$100 (known risk) over a draw from an urn with an unknown probability of winning \$100, even if the latter might also be 50% – they dislike the ambiguity itself. This is formalized in models like **maxmin expected utility** (Gilboa-Schmeidler 1989) where a person considers the worst-case probability distribution. At first glance, ambiguity aversion and CM both lead to conservative choices: an

ambiguity-averse person might avoid an option because *it could* conceal a bad probability, somewhat like a CM person avoids it because it could lead to disaster. However, ambiguity aversion is about **uncertainty in probabilities** (second-order uncertainty), whereas CM is about **severity of outcomes**. An ambiguity-averse decision-maker with potential catastrophic outcomes might behave like a CM agent if they suspect the "worst" probability of catastrophe. For instance, if one doesn't know the failure rate of a new technology, ambiguity aversion might lead them to assume it's high and avoid using it (a cautious approach), consistent with CM. But if probabilities are known and fixed, an ambiguity-averse and a risk-neutral but CM-focused agent would differ: CM cares not whether probability is known, only that it's minimized. Ambiguity aversion alone wouldn't always prioritize worst outcomes lexicographically; it would, however, amplify the weighting of unknown risks. In sum, CM is *orthogonal* to ambiguity concerns: one could be ambiguity-neutral but still lexicographically avoid low-probability catastrophes because of outcome severity, or one could be ambiguity-averse yet still not lexicographic about outcomes. Models like **α-maxmin** or **lexicographic probabilities** [28] can combine these attitudes, but conceptually CM is about the utility of outcomes, not the knowledge of probabilities. Ambiguity-averse people might be thought of as hedging against unknown catastrophes (robustness), which complements CM's directive to avoid catastrophe at all.

- **Minimax and Maximin (Worst-Case Optimizing):** In classical decision theory under complete uncertainty (no probabilities), the **maximin** rule says to evaluate each action by its worst possible outcome and choose the action with the best worst-case. This is essentially a *deterministic* analog of CM if we label "catastrophe" any very bad outcome: maximin would avoid any action whose worst-case is worse than another's. CM can be seen as a probabilistic refinement: it agrees with maximin when probabilities cannot be quantified (treating any possible catastrophe as something to avoid), but once probabilities are introduced, CM does not ignore them – it still tries to minimize that probability rather than assuming it will happen. Another related criterion is **minimax regret**, which we discussed earlier: minimizing the worst-case regret [9] is different because regret is measured against the best that could have been in each state, rather than absolute outcomes. CM does not consider regret at all; it is only concerned with actual consequences, not how one might feel or compare outcomes post hoc. Maximin often yields very conservative decisions (sometimes criticized as too pessimistic if probabilities are known and small for worst-cases). CM is similarly conservative but more nuanced, as described in Section 4. Notably, in **robust control theory**, a minimax (H-infinity) controller is designed to handle the worst disturbance. A CM-based controller would be designed to never fail (if possible), even if it means suboptimal performance, which is analogous in that domain. But robust/minimax control doesn't typically involve a lexicographic *two-tier* goal; it's a single worst-case optimization. CM's two-tier nature is closer to what in multi-objective optimization is called a **lexicographic objective**: primary objective survival, secondary performance. This concept appears in some engineering contexts (e.g., first ensure stability, then optimize efficiency).

- **Regret Minimization:** Although already touched on, from a *behavioral perspective*, regret is a subjective emotional criterion – *"choose the option you'll regret least"*. People often say they want to avoid actions that could lead to terrible regret even if they might have good outcomes [29]. This can align with avoiding catastrophe: a catastrophic outcome usually also implies huge regret if one had an alternative. The formal minimax regret rule [30], however, is, as noted, not the same as minimizing catastrophe probability – it's minimizing the difference between what you get and what you *could have gotten* in each state. CM doesn't compute such differences; it cares about absolute survival. One interesting link: a person who is following CM might *rationalize* it in terms of regret ("I'd never forgive myself if I took a risk and it led to someone dying, so I won't take that risk"). That is essentially

converting the objective consequence into a psychological cost. But regret as a theory can sometimes favor risk-seeking (to avoid regret of missing out on a huge gain) or other paradoxical behaviors, whereas CM would not if no catastrophic downside is present. Thus, regret minimization and CM can coincide in some choices but diverge in others. For example, a doctor choosing a treatment might think in CM terms ("I must avoid the treatment option that could kill the patient, above all"), or in regret terms ("If I choose a risky surgery and the patient dies, I'll regret it terribly, but if I don't do it and they die slowly, I might also regret not trying – which regret is worse?"). These frames can lead to different decisions because regret weighs counterfactuals, whereas CM looks only at actual outcomes.

- **Loss Aversion (Prospect Theory) vs CM:** In behavioral economics, **loss aversion** means losses loom larger than gains – e.g. people need about twice the gain to compensate for a given loss (Kahneman & Tversky, 1992). Loss aversion is often cited as an explanation for cautious behavior: people avoid bets with positive expectation if there's a chance of loss, due to the psychological extra weight of losses. This is related to CM in spirit – prioritizing avoiding negative outcomes. However, prospect theory's value function is still finite for any loss; it's just steeper for losses than gains. There is no lexicographic cutoff where an outcome is infinitely bad. So extreme loss aversion would approach CM, but not reach it. Interestingly, some evolutionary arguments for loss aversion invoke a CM-like logic: organisms near survival thresholds who treat losses as more serious than equivalent gains may out-survive those that don't, because preserving what you have (to avoid death) is crucial [31] [32]. In fact, research has found that an optimal loss aversion coefficient around 2 might relate to minimizing extinction risk in uncertain environments [31]. That hints that *biologically*, humans' loss aversion could be an *approximation* to CM developed through evolution: overweighting potential losses helps avoid ruin in the long run [31]. But again, loss aversion is a one-system utility adjustment, whereas CM is a two-system lexicographic rule. If humans truly had lexicographic preferences, classic experiments (like trading off a tiny risk of death for a large reward) would show absolute refusal beyond some point. Generally, people's behavior is better described by very steep but not vertical trade-offs (e.g. they might skydive for fun, accepting a 0.001% chance of death for enjoyment – something a pure CM agent would never do unless that person considered that risk literally zero or not catastrophic). So psychologically, CM is an idealized limit that real behavior approaches in high-stakes cases (people become extremely loss-averse or risk-averse when stakes are life-and-death, as expected).

- **Negative Utilitarianism:** Philosophically, CM echoes **negative utilitarianism**, which holds that minimizing suffering (or bad outcomes) is morally more urgent than maximizing happiness [33]. Karl Popper's quote encapsulates this: we should prioritize the elimination of suffering over the promotion of pleasure [33]. This is virtually a moral analog of CM: treat extreme suffering (which we can liken to catastrophe in a moral sense) as lexicographically worse than any absence of extra happiness. Negative utilitarians would, for instance, argue that preventing hellish outcomes for some people is more important than providing mild pleasures to many others. CM in decision-making likewise says preventing catastrophic loss (to oneself or one's system) outweighs chasing extra gains. A criticism raised against negative utilitarianism is the so-called "world destruction" argument (R. N. Smart, 1958) – if one took minimizing suffering literally, one might justify painlessly annihilating humanity to eliminate the possibility of suffering [34] [35]. That's an *extreme* consequence of pure lexicographic weighting of suffering over all else. CM as applied to individual or organizational decision-making usually has a bounded scope (e.g. an AI should minimize probability of catastrophic error). It doesn't inherently say "if existence itself has some suffering, eliminate

existence" because the agent's decision scope isn't that broad (and also because CM is typically applied with respect to that agent's own catastrophic outcomes, not aggregating across individuals as in ethics). Nonetheless, the parallel is thought-provoking: a strict CM attitude could lead to very *conservative or drastic* measures if taken to an extreme. In AI safety, for example, if an AI is given a lexicographic goal to never cause harm, it might choose to stop operating or confine itself extremely to avoid any chance of harm – analogous to the negative utilitarian's paralysis or extreme acts to avoid worst outcomes. So in design, one might temper lexicographic rules with some pragmatic considerations to avoid pathological solutions (this is related to the idea of "lexicographic but with epsilon" or considering very small probabilities as effectively zero beyond some threshold).

The following table summarizes key distinctions between CM and related decision frameworks:

Concept	Core Focus	Trade-off Stance	Formal Model	Attitude to Catastrophe
Consequence Minimization (CM)	Avoiding *catastrophic outcomes* first, then maximizing gain [4].	Lexicographic (no trade-off accepted if it increases $P(S_c)$ above minimum attainable).	Two-tier utility: primary = survival (0/1), secondary = expected utility [2] [5].	Treats catastrophe as infinitely (lexicographically) worse than any non-catastrophic outcome.
Risk Aversion (Concave EU)	Reducing *outcome uncertainty* (variance); dislike of losses [8].	Continuous trade-off (will accept small risk for large reward).	Concave utility $u(x)$; maximize $E[u(x)]$.	Bad outcomes heavily dispreferred but *finite* disutility; small risk can be outweighed by enough benefit.
Ambiguity Aversion	Avoiding *uncertainty in probabilities* (unknown odds) [27].	Prefers known risks to unknown; effectively assumes worse probabilities for ambiguous options.	Maxmin over priors, e.g. maximize $ \min_{P\in\mathcal{P}} E_P[u(x)]$.	Indirect: may overweight probability of catastrophe if its likelihood is ambiguous, thus often cautious about potential catastrophes but not lexicographically (except under worst-case prior).

Concept	Core Focus	Trade-off Stance	Formal Model	Attitude to Catastrophe
Minimax (Maximin)	Optimizing *worst-case payoff* (zero-sum mindset).	No trade-off: evaluate only worst outcome of each action.	$\max_a \min_{\omega} \text{Payoff}(a,\omega) $.	If catastrophe is one possible outcome, that action's value = catastrophic payoff (treated as certain if possible). Avoids any act with worse worst-case than alternatives [26].
Minimax Regret	Minimizing *worst-case regret* (hindsight comparison) [9].	No direct trade-off on outcomes; trade-off on *regret values*.	$\min_a \max_{\omega} [\text{opt payoff in } \omega - \text{payoff} (a,\omega)]$.	Not directly about catastrophe; if all actions equally bad in a state, regret is low, so criterion might not avoid catastrophe if it's unavoidable in that state (focuses on comparative loss, not absolute).
Loss/ Disappointment Aversion (Prospect Theory)	Avoiding *losses relative to a reference*; losses felt more strongly [23].	Diminishing sensitivity; some trade-off (will take risk if gains huge, but needs extra incentive for risky losses).	Value function $v(x)$ kinked at 0 (loss side steeper); probability weighting.	Big losses carry high disutility (often ~2x of equal gain [31]), but extreme loss still finite negative value. No strict lexicographic cutoff.
Negative Utilitarianism (Ethics)	Reducing *suffering/ harm* in aggregate [2].	Lexicographic (in strong form): no amount of good can outweigh extreme suffering [7].	Lexical priority to negative utility (suffering) in social welfare function.	Preventing catastrophic suffering is the paramount moral duty; could justify extreme measures to avoid worst-case for anyone.

Sources: Key characteristics compiled from theoretical definitions and examples [8] [26] [27] [2].

This comparison clarifies that **Consequence Minimization stands out in its absolute prioritization of avoiding existential or catastrophic failures**. It is more rigid than risk aversion (which still allows trades),

more outcome-focused than ambiguity aversion or regret (which involve subjective factors), and more probability-sensitive than pure minimax (which ignores how likely worst-cases are). CM aligns with certain ethical intuitions (e.g. "first, do no harm" in medicine could be seen as a CM principle) and with some evolved behaviors (e.g. organisms strongly avoiding deadly risks), but it also challenges the conventional utility-maximizing paradigm by introducing a discontinuity in preferences.

Conclusion

We have formally defined the principle of Consequence Minimization (CM) as a lexicographic preference structure that treats the prevention of catastrophic outcomes as lexicographically prior to the pursuit of gains. Through representation theorems and examples, we distinguished CM from standard risk aversion – notably by its refusal to trade off even tiny increases in ruin-risk for outsized rewards, a stance not captured by any finite concave utility [7]. We showed how CM can be embedded in constrained MDP models via the viability kernel, yielding policies that ensure survival (if possible) at the expense of some expected reward, and we hypothesized that in heavy-tailed environments such policies achieve greater long-term welfare than naive expected-value maximization [20]. A perhaps counterintuitive aspect of CM is its state-dependent risk posture: when an agent is on the brink of inevitable failure, CM prescribes risk-seeking "gambling for resurrection" as the only rational chance for survival [23], a result we formalized with simple CMDP constructions. We explored connections to minimax strategies, finding that CM generalizes minimax: it agrees in adversarial worst-case scenarios but is less conservative when probabilities allow hope, thereby diverging in stochastic and cooperative contexts. Finally, we situated CM among related ideas – from economic safety-first rules [3] to psychological loss aversion and philosophical negative utilitarianism [2] – underscoring that while all share a focus on bad outcomes, CM's lexicographic structure is unique in its uncompromising hierarchy of objectives.

The development of CM as a decision principle has both normative and practical implications. Normatively, it provides a coherent framework for preferences in high-stakes situations (e.g. AI systems that must never cause catastrophic harm can be designed with CM-like utility). It is falsifiable by observing whether decision-makers ever exhibit such lexicographic indifference to trade-offs – a question for empirical research in human and organizational behavior. Practically, CM approaches may improve safety in fields like engineering (fail-safe design), finance (preventing ruinous investment strategies [22]), and public policy (risk management for extreme events). However, adopting CM must be tempered by awareness of its potential downsides: it can lead to overly cautious or paradoxical choices if applied inappropriately (e.g. paralysis in the face of any risk). One must also define carefully what constitutes "catastrophe" (S_c) – a non-trivial ethical and technical decision.

In conclusion, Consequence Minimization offers a powerful lens to analyze decisions where catastrophic outcomes are possible. By modeling it rigorously, we highlight how it departs from classical decision theory and when it aligns with or contradicts human behavior. This work lays an axiomatic foundation for CM and opens avenues for further research: exploring hybrid models (e.g. allowing ϵ probabilities of catastrophe), dynamic consistency of CM plans, multi-agent considerations, and computational methods to solve CM-optimized policies in complex systems. Ultimately, CM formalizes the age-old wisdom that **survival is the precondition of success** [17] – a principle that, once encoded into our decision models, could help safeguard against the low-probability, high-impact risks that our increasingly complex world faces.

References: (Selected inline source citations) [8] [3] [2] [9] [23] [20]

1 4 29 30 33 ChatGPT - Consequence Minimization Across Disciplines.pdf
file://file-UH2TCa15vbz619GKs2Uss7

2 7 34 35 Negative utilitarianism - Wikipedia
https://en.wikipedia.org/wiki/Negative_utilitarianism

3 Roy's safety-first criterion - Wikipedia
https://en.wikipedia.org/wiki/Roy%27s_safety-first_criterion

5 6 28 Lexicographic Probabilities and Choice Under Uncertainty
https://cet.econ.northwestern.edu/dekel/pdf/lexicographic-probabilities-and-choice-under-uncertainty.pdf

8 Risk aversion (psychology) - Wikipedia
https://en.wikipedia.org/wiki/Risk_aversion_(psychology)

9 26 Regret (decision theory) - Wikipedia
https://en.wikipedia.org/wiki/Regret_(decision_theory)

10 Gemini - Consequence Minimization.pdf
file://file-WpB75pbPLkHKhJincNPnZW

11 12 13 14 15 proceedings.mlr.press
http://proceedings.mlr.press/v100/heim20a/heim20a.pdf

16 20 22 fooledbyrandomness.com
https://fooledbyrandomness.com/DarwinCollege.pdf

17 18 19 Medium: The Logic of Risk Taking
https://nassimtaleb.org/2017/08/medium-logic-risk-taking/

21 Gambler's ruin - Wikipedia
https://en.wikipedia.org/wiki/Gambler%27s_ruin

23 Understanding Decision-Making: Inherent Risk Preferences | Darden Ideas to Action
https://ideas.darden.virginia.edu/-effective-decision-making-risk-preferences

24 25 Upon proving that the best betting strategy for “Gambler's Ruin” was ...
https://math.stackexchange.com/questions/3327687/upon-proving-that-the-best-betting-strategy-for-gamblers-ruin-was-to-bet-all

27 Ambiguity and Ambiguity Aversion - ScienceDirect
https://www.sciencedirect.com/science/article/abs/pii/B9780444536853000131

31 32 Perplexity - Consequence Minimization - A Universal Principle.pdf
file://file-EvGZcTVqZsCuMV8DRzqdVa