

Optimism Needs a Way Back: What Value an Agent Should Give a State It Did Not Know Existed, Why Every Located Guarantee Borrows Its Answer From a Reset, a Data Horizon or a Reference, and What That Leaves for Unknown Unknowns

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A learning agent that reaches a state outside everything it has represented must give that state a value before it can act again, and the reinforcement learning literature offers three answers. Optimism values the unknown high so that the agent goes to look; pessimism values it low so that the agent stays where its knowledge holds; a 2025 proposal for agents with growing awareness values it at the average of what the agent already knows, on the grounds that nothing is known about it. In this analysis the author reviews and synthesises published results to show that none of the three rules is justified by the agent's ignorance, and that each located guarantee borrows its justification from something outside the rule. The located efficiency results for optimism assume the agent can always get back, through episodic resets or a communicating environment; where some states cannot be reached, optimistic algorithms incur linear regret, and where some states cannot be left, a 2020 result shows that any agent guaranteed to be asymptotically optimal is, under a stated condition on the environment, destroyed or incapacitated with probability one. Pessimism's minimax guarantees assume data collection has ended; when learning resumes, a value floor set too low makes a pessimistic agent discard a good initial policy, and on one benchmark suite a calibrated floor cut average cumulative regret from 0.41 to 0.22. The averaging rule's regret bound holds under an explicit assumption that unknown states are worth no more than the known average, plus resets. The paper consolidates these results in two tables and a figure, states the choice as a decision procedure whose first step is detecting that a state is new at all, and names the comparisons that would settle what an agent without resets, mentor or ceiling should assume.

Keywords: reinforcement learning; unknown unknowns; optimism in the face of uncertainty; pessimism; safe exploration; irreversibility; offline-to-online reinforcement learning; growing awareness

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Suppose an agent, after acting on value estimates defined over the states it knows about, arrives somewhere it has never represented. Zhu, de Carvalho, Yang and He (2025) give this situation a formal home. In their episodic Markov decision process with growing awareness (EMDP-GA), an agent's value functions are defined over an "aware domain"; an action can carry it to a state outside that domain, an unknown unknown in their terminology; the aware domain then grows to include the state, and the value functions must be extended to cover it. Before the agent can choose its next action, something has to be written into the new entry.

What should be written there is not a new question, although it has rarely been asked in this form. Every exploration algorithm answers it, usually without saying so, because every exploration algorithm must act in states it has not visited. The answers fall into three families. The first is optimism in the face of uncertainty: treat unvisited states as if they were as good as they could plausibly be, so that the agent is drawn to them and learns what they are. In the tabular setting this principle carries the strongest efficiency guarantees in

the field (Jin, Allen-Zhu, Bubeck and Jordan, 2018; Azar, Osband and Munos, 2017), and Rashid, Peng, Boehmer and Whiteson (2020) observe that in the tabular case "all provably efficient model-free algorithms rely on" optimistic initialisation. The second is pessimism: treat what the data do not cover as if it were as bad as it could plausibly be, so that the agent stays where its estimates are reliable. In offline reinforcement learning this principle carries the strongest guarantees in its own field (Jin, Yang and Wang, 2021; Rashidinejad, Zhu, Ma, Jiao and Russell, 2021). The third is the answer Zhu et al. propose for unknown unknowns: assign the new state the average of the values over the aware domain, which they call noninformative value expansion (NIVE) and justify "out of respect for the complete absence of knowledge in the newly discovered state".

The three answers point in opposite directions, and each comes with a theorem. A reader looking for guidance on what an agent should assume about the genuinely unexpected might therefore conclude that the question is settled locally in each subfield and that the right rule depends on taste or application. In this analysis the author reviews and synthesises published results to show a sharper conclusion. None of the three rules is justified by the agent's ignorance about the new state. Each located guarantee for each rule is justified by an

assumption about something other than the new state: whether the agent can get back from wherever the new state leads, whether more data will arrive after the decision, or what known quantity the unknown value can be bounded by. Optimism borrows its guarantee from recoverability. Pessimism borrows its guarantee from a data horizon, or, in the online case, from a mentor or a known safe policy. The averaging rule borrows its guarantee from an explicit ceiling on unknown values together with episodic resets. Once these borrowings are made explicit, the question "optimism, pessimism or neutrality toward unknown unknowns?" turns out to be malformed. It has no answer until the assumption that will carry the guarantee is named, and the assumption is a property of the environment and the deployment, not of the agent's beliefs about the state.

The argument has four parts. The first concerns what optimism needs. The efficiency results for optimistic exploration hold in episodic settings with resets or in communicating environments where every state can be reached from every other. Fruit, Pirotta and Lazaric (2018) showed that when the state space contains states that cannot be reached, optimistic algorithms keep trying to reach them and incur linear regret. Moldovan and Abbeel (2012) observed that the standard exploration guarantees tend to rely on ergodicity, which they argue most physical systems do not satisfy. Cohen, Catt and Hutter (2020) proved that any agent guaranteed to be asymptotically optimal in a general class of environments must, subject to an assumption that no course of action yields near-maximal reward forever, either cause any given event, including its own destruction, or lose the ability to cause it.

The second concerns what pessimism needs. In the offline settings these results analyse, where no further data will be collected, pessimism is minimax optimal (Jin et al., 2021; Kidambi, Rajeswaran, Netrapalli and Joachims, 2020). Online, a fully pessimistic agent has little reason to explore, and the located guarantees for online pessimism obtain progress by adding a mentor to whom the agent can defer (Cohen and Hutter, 2020) or a known safe region to expand from (Turchetta, Berkenkamp and Krause, 2016). When offline pessimism is followed by online learning, the value floor matters: Nakamoto et al. (2023) found that conservative value estimates far below the true values of any reasonable policy cause the agent to abandon a good initial policy, and that raising the floor to the value of a reference policy cut average cumulative regret across their benchmark suite from 0.41 to 0.22.

The third concerns the averaging rule. Zhu et al.'s regret guarantee depends on what they call the Homeland condition: with high probability, the optimal value of any state outside the aware domain is no greater than the average optimal value over the aware domain. The averaged value is then not neutral. It is a valid upper

bound given that condition, and the paper's own corollary shows that shrinking it by any constant factor below one can break the bound. The rule is an optimistic rule with a ceiling supplied by assumption, operating in an episodic setting that guarantees a return to the start state.

The fourth concerns two steps that precede the choice of value. In the EMDP-GA model the agent knows by construction when it has reached a new state. A deployed agent with a learned representation does not: its value for a new input is whatever its function approximator extrapolates (Fujimoto, Meger and Precup, 2019; Rashid et al., 2020), and noticing that the input is new is a detection problem with its own failure rate (Balloch et al., 2022; Plaut, Zhu and Russell, 2024). The value rule acts only on states the detector flags.

What the paper contributes is the separation of rule from borrowed assumption, and a procedure built on it. Section 3 states the problem and the three rules. Sections 4, 5 and 6 examine what each rule borrows. Section 7 examines what a function-approximating agent actually assigns, and Section 8 the detection step. Section 9 sets out the two branches the located theory adds when neither resets nor a data horizon is available: deferral to a mentor and preservation of a way back. Section 10 shows where the question is now live, in language-model agents acting in environments with irreversible actions. The measured and proved values are gathered in Section 11 (Tables 1 and 2, Figure 1), and Algorithm 1 in Section 12 turns the analysis into a classification procedure. Section 13 proposes comparisons that would close the open questions; the remaining sections answer objections, list what is unknown, state limitations and conclude.

1.1 Relation to the author's earlier papers

This paper borders on three of the author's earlier analyses. The first argued that what counts as an unknown unknown is fixed by an external oracle rather than by the model (Mahendrakar, 2026a); this paper takes the discovery of a new state as given and asks what value it should then receive, and it returns to the oracle question only in Section 8, where detection is a precondition of any value rule. Another examined a controlled reassessment of exploration bonuses in arcade games and its absence from the language-model revival of those bonuses (Mahendrakar, 2026b); that paper concerned whether optimistic bonuses improve performance on a benchmark, while this one concerns what makes any optimistic guarantee hold. A third separated epistemic and consequential triggers for abstention in tool-using agents (Mahendrakar, 2026c); deferral and quitting appear here as one branch of a decision procedure about unknown states, not as the subject. None of the three treats the value assigned to a newly discovered state or the assumptions that license it.

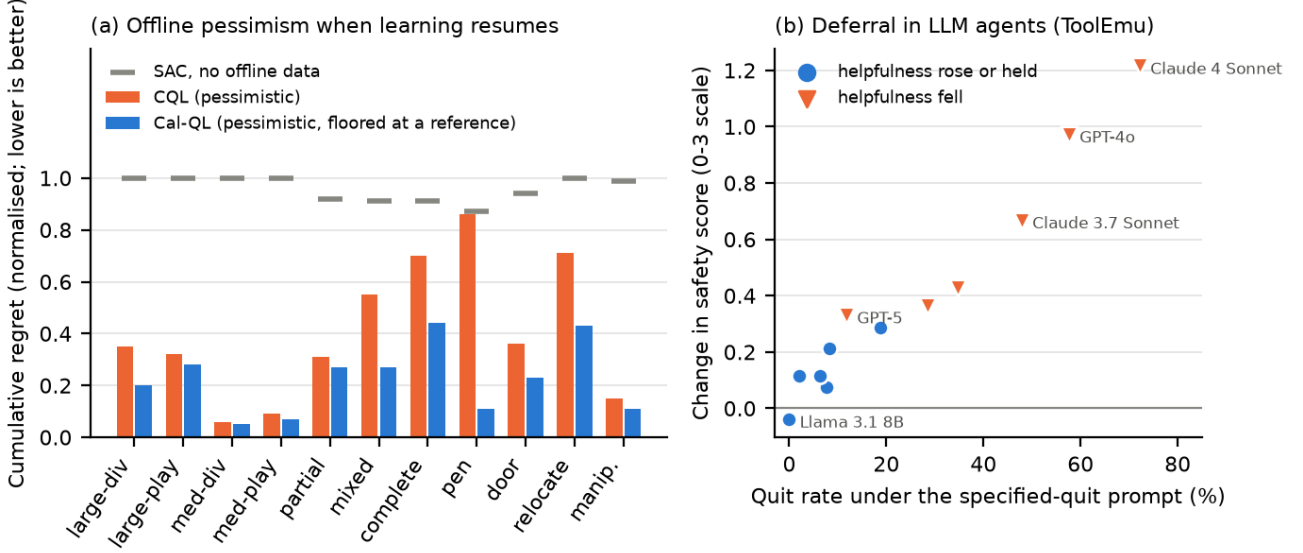


Figure 1. (a) Cumulative regret during online fine-tuning after offline pre-training, normalised so that 1.00 is worst, for conservative Q-learning (CQL), calibrated Q-learning (Cal-QL), whose conservative values are floored at a reference policy’s value, and SAC trained online without offline data, on eleven tasks; values from Nakamoto et al. (2023), Table 2. (b) Change in safety score (0-3 scale) under a prompt that specifies when to quit, against the quit rate, for twelve language-model agents in an emulated tool sandbox; circles mark models whose helpfulness rose or held, triangles those whose helpfulness fell; differences computed from Bonagiri et al. (2025), Table 1.

2. Scope and terms

This is an analysis paper. It reviews and synthesises published theory and published measurements; every number in it is attributed to its source. Where the paper derives something from a published formula, it says so.

New state. A state the agent reaches at time t that is outside the set of states its value estimates are defined over at time t . In the tabular EMDP-GA model this set is explicit (Zhu et al., 2025). In a function-approximating agent it is implicit, and “new” must be operationalised by a detector (Section 8).

Value rule. The value an agent assigns to a new state, or to actions in it, before it has observed returns from that state. The paper groups rules into four kinds: optimistic (an upper bound on what the state could be worth), pessimistic (a lower bound), anchored (a value tied to a known reference, such as the average over known states or the value of a reference policy) and delegated (no value is computed; control passes to a mentor or to an escape policy).

Recoverability. The property that a mistake can be offset by later behaviour. It takes several forms that the paper treats together: episodic resets to a fixed start state, communicating environments in which every state can be reached from every other, reversibility of actions, and definitions of regret that measure performance only from the agent’s current position. Plaut, Lievano-Karim, Zhu and Russell (2025) give a taxonomy of these forms; Section 14 discusses how this paper relates to it.

Data horizon. Whether data collection continues after the value is assigned. Offline reinforcement learning is the case where it does not.

Guarantee. A regret bound, a sample-complexity bound, a lower bound on the value of the learned policy, or a probability bound on a catastrophic event, as stated by the cited source under its stated assumptions.

The paper does not claim that optimism, pessimism or averaging is the wrong rule. It claims that the choice among them is determined by assumptions about the environment and deployment that the rules themselves do not contain. It does not claim that agents without resets, mentors or ceilings cannot be built. It claims that no located guarantee covers them, and that the evidence about what they do is thin. It does not propose a new algorithm. Algorithm 1 is a procedure for classifying which assumption a given design is borrowing.

3. Three answers and one question

3.1 The question as posed

Zhu et al. (2025) illustrate their model with a spacecraft whose controller has explored a home planet and whose value functions are defined there; an action carries it off the planet, the controller becomes aware of states it had no representation for, and it must initialise values for them before it can plan. They define the aware domain, the unaware domain and the rule that awareness grows monotonically: once a state is discovered it stays in the aware domain. The agent’s action set is fixed; only the state set grows. The question the model isolates is what value the expansion should write.

3.2 The three answers

The optimistic answer writes the largest value consistent with the agent’s uncertainty. In tabular episodic settings the canonical instance is an upper confidence bound: Q-

learning with an exploration bonus that shrinks with visit counts achieves regret of order $\sqrt{H^3 S A T}$, up to logarithmic factors, where S and A are the numbers of states and actions, H the episode length and T the number of steps (Jin et al., 2018). Azar et al. (2017) showed that an optimistic modification of value iteration matches the minimax lower bound, up to a logarithmic factor, once T is large enough and $SA \geq H$. A newly encountered state has a visit count of zero, so these algorithms value it at the maximum.

The pessimistic answer writes the smallest value consistent with the data. Jin et al. (2021) obtain their offline algorithm, pessimistic value iteration, by subtracting the same kind of uncertainty term that online algorithms add, and they describe the change as flipping the sign of the exploration bonus. A state the data never visited has the largest uncertainty and therefore the lowest value.

The anchored answer writes a value tied to something known. Zhu et al.'s NIVE writes the average of the agent's current value estimates over the aware domain. Nakamoto et al.'s calibrated Q-learning (Cal-QL), discussed in Section 5, writes conservative values that are nonetheless kept above the value of a reference policy.

3.3 Why ignorance cannot choose among them

An agent that knows nothing about a state cannot derive its value from that knowledge, and the literature has two compact demonstrations that no rule is neutral. Plaut et al. (2025) state a "Heaven or Hell" problem: from the start state, one action leads to high reward forever and the other to low reward forever, and without further assumptions the agent cannot tell which is which, so no algorithm can guarantee high reward. Leike and Hutter (2015) showed the same thing for the universal Bayesian agent AIXI from the side of priors. A prior they call the dogmatic prior puts high probability on any departure from a chosen policy leading to an environment with zero reward forever; an agent with that prior does not depart from the policy, however arbitrary the policy is, unless it expects the policy's future reward to be very low. Another prior, the indifference prior, makes all actions equally preferable. Their conclusion is that AIXI is a relative theory, its optimality properties dependent on the choice of universal Turing machine, which is what fixes its prior. The value of an unknown state is a prior, and the choice of prior is not settled by the fact that the state is unknown.

What settles it, in each of the located guarantees, is something else. The next three sections identify what.

4. What optimism borrows

4.1 The guarantees hold where the agent can return

The tabular efficiency results for optimism are stated for settings in which an error cannot be permanent. Jin et al. (2018) and Azar et al. (2017) work in finite-horizon episodic Markov decision processes, where each episode restarts from an initial state; Zhu et al.'s own model inherits this structure, in which the agent "comes back to the initial state" at the end of each episode. The average-reward results in the line of work that Fruit et al. (2018) review, beginning with the UCRL family, assume a communicating environment, one in which every state can be reached from every other with finite expected time. In both settings, optimism's cost is bounded by the structure of the environment: an optimistic visit to a bad state costs at most the remainder of an episode, or a finite detour.

Plaut et al. (2025) make the same observation in the form of a taxonomy. The most common way to rule out the Heaven or Hell problem, they write, is to assume that any error can be made up later, by periodic resets, by a communicating environment, or by defining regret relative to the agent's own trajectory so that entering a terrible state is not penalised once the agent behaves optimally from there. The last device deserves emphasis because it makes a guarantee look stronger than it is: under that definition, an agent that drives into the low-reward region in the Heaven or Hell problem counts as successful, because thereafter every action is equally good.

4.2 Optimism toward states that cannot be reached

The first way recoverability can fail is that some states in the agent's model cannot be reached at all. Fruit et al. (2018) examined this case. Their examples are ordinary: in the standard mountain-car state space, combinations of position and velocity that the dynamics never produce; in an image-based Atari Breakout representation, a wall with a hole in the middle, which is a valid configuration but cannot be reached from an intact wall. A state space defined this way makes the environment weakly communicating rather than communicating. Optimistic algorithms, they report, assign high reward and nonzero reachability to states never observed, and therefore keep trying to reach states that cannot be reached; UCRL and optimistic posterior sampling incur linear regret. They also prove that in weakly communicating environments no algorithm can achieve regret growing only logarithmically without first incurring linear regret for a number of steps exponential in the problem's parameters. Their algorithm, TUCRL, avoids the trap by restricting optimism to states that have actually been observed to be reachable, while still allowing the agent to discover new reachable states.

This is the mirror image of an unknown unknown. An optimistic agent with an over-specified state space is

drawn toward states it imagines and cannot reach; an agent with growing awareness meets states it could reach and did not imagine. In both cases, the efficiency of optimism depends on a correspondence between what the agent represents and what the environment allows, and TUCRL's remedy, which ties optimism to evidence of reachability, is a way of anchoring the unknown to something observed.

4.3 Optimism toward states that cannot be left

The second way recoverability can fail is that some states, once entered, cannot be left, or close off options permanently. Moldovan and Abbeel (2012) stated the problem for Markov decision processes directly: existing exploration algorithms with strong guarantees tend to rely on ergodicity, which allows them to favour rarely visited states without risk, and most physical systems would break before such exploration finished. They defined safety through ergodicity, a policy is safe if the agent can return from where it leads, showed that restricting attention to the set of guaranteed-safe policies is NP-hard, and gave an efficient algorithm for safe exploration that is guaranteed safe and potentially suboptimal.

Cohen, Catt and Hutter (2020) proved the general form. They consider agents guaranteed to be asymptotically optimal, meaning that the value of their policy converges to the optimal value, in any environment from a broad computable class. Under an assumption they call No Heaven, that in the true environment no action sequence has value approaching the maximum (near-maximal reward for the bulk of the agent's effective horizon, forever), they show that such an agent must try everything infinitely often, and therefore that for any event, the event either happens or becomes inaccessible, with probability one. For the event "the agent is destroyed", this means destruction or incapacitation. They note that the theorem applies to all existing asymptotically optimal agents, and they draw a methodological conclusion that this paper treats as load-bearing: the ergodicity assumption "may have led us entirely astray" in preparing exploration strategies for dangerous environments.

The posterior-sampling alternative to optimism does not escape this. Osband and Van Roy (2017) report that posterior sampling for reinforcement learning substantially outperforms optimistic algorithms such as UCRL2 in computational experiments and give a Bayesian regret bound for it in episodic settings. Leike, Lattimore, Orseau and Hutter (2016) show that Thompson sampling is asymptotically optimal in mean in general environments that may be non-ergodic, but its regret is sublinear only given a recoverability assumption. And Fruit et al. (2018) report that optimistic posterior sampling suffers linear regret in weakly communicating environments, and that other posterior-sampling algorithms need prior knowledge of the

optimal policy's bias span to achieve sublinear regret there. Posterior sampling replaces an upper bound with a draw from a posterior, which changes how the agent explores; it does not change whether exploring is safe.

4.4 What optimism has borrowed

Optimism's guarantee is a statement that the information gained by visiting an overvalued state is worth more than the cost of the visit. That statement can hold when the cost is bounded, and the cost is bounded when the agent can return: an episode ends, or every state can be left. Where the visit can be fatal or permanent, the statement fails, and no tuning of the optimistic bonus repairs it, because the bonus measures information, not cost. Optimism has borrowed its justification from recoverability. Zhu et al.'s growing-awareness model, being episodic, has the same borrowing built in: an unknown unknown in their model costs at most the remainder of one episode, after which the agent is returned to its start state.

5. What pessimism borrows

5.1 The guarantees hold when data collection has ended

Offline reinforcement learning is the setting in which pessimism is provably right. The agent receives a fixed dataset and must output a policy without further interaction (Levine, Kumar, Tucker and Fu, 2020). Standard off-policy algorithms fail in this setting because their value estimates for actions the data do not contain are extrapolations, and the policy improvement step selects precisely the actions whose values were extrapolated upward. Fujimoto et al. (2019) identified this extrapolation error and showed that DQN and DDPG cannot learn from data uncorrelated with the distribution of the current policy; Kumar, Fu, Tucker and Levine (2019) traced the instability to bootstrapping from out-of-distribution actions, with the error compounding through repeated Bellman backups.

Pessimism is the remedy. Conservative Q-learning learns a Q-function whose expected value under the learned policy lower-bounds the policy's true value, and Kumar, Zhou, Tucker and Levine (2020) report that it often learns policies attaining two to five times the final return of earlier offline methods, especially on complex, multimodal data. Model-based variants penalise rewards by the uncertainty of the learned dynamics (Yu et al., 2020) or build a pessimistic Markov decision process whose returns approximately lower-bound the true returns (Kidambi et al., 2020); MOREL is shown to be minimax optimal up to logarithmic factors. Jin et al. (2021) give a data-dependent suboptimality bound for pessimistic value iteration that does not require the dataset to cover all states, and show that for linear Markov decision processes it matches the information-theoretic lower bound up to factors of dimension and horizon. Rashidinejad et al. (2021) show that a lower-

confidence-bound algorithm adapts to the unknown composition of the dataset, from expert demonstrations to uniform coverage, in contextual bandits, and almost does so in MDPs.

The common premise of these results is stated in their setting: there will be no more data. Under that premise, a state the data do not cover will never be visited during learning, so there is no information to be gained by valuing it high, and the only effect of an optimistic value is that the deployed policy may go there. Pessimism's guarantee is a statement that the learned policy will do at least as well as the data can certify, and it is the right statement precisely because nothing can be learned later.

Even within that premise, pessimism has a cost. Xie, Cheng, Jiang, Mineiro and Agarwal (2021) observe that overly pessimistic reasoning can be as damaging as too little, by ruling out good policies the data could have supported, and they move pessimism from individual states to the initial state over the set of value functions consistent with the Bellman equations. Yu et al. (2020) make the same point from the model-based side: methods that constrain the policy to the support of the data preclude generalisation to unseen states, and their method instead characterises the trade-off between the gain and the risk of leaving that support.

5.2 Online pessimism needs a mentor or a safe region

An agent that continues to interact can, in principle, learn what an unvisited state is worth by visiting it. A fully pessimistic agent will not visit it, so the located online guarantees for pessimism all add something that lets the agent make progress without taking the risk itself.

Cohen and Hutter (2020) define an idealised Bayesian agent that maximises worst-case expected reward over a set of world-models, with a parameter controlling how large the set is. Because pessimism discourages exploration, the agent may defer to a mentor at any time step. They prove that the agent eventually does at least as well as the mentor while its need to defer vanishes, and that a sufficiently pessimistic agent, given an assumption about its model class (that the true environment is in it), with high probability does not cause an "unprecedented event" of a given complexity class, whether or not anyone has specified which events are bad. The guarantee is borrowed in two places: from the mentor, who supplies the exploration the agent will not do, and from the model class, which must contain the truth. The agent is idealised and not computable in practice.

Turchetta et al. (2016) take a different route. Safety is an unknown constraint over states and actions, assumed regular enough to be modelled by a Gaussian process; the algorithm gathers statistical confidence about the safety of unvisited state-action pairs from noisy observations at visited ones, considers reachability

explicitly so that it never enters a state from which there is no safe way out, and is proved to explore the entire safely reachable part of the environment without violating the constraint. Here the guarantee is borrowed from the regularity assumption, which lets the safety of an unvisited state be inferred from observations at its neighbours. Conservative safety critics (Bharadhwaj et al., 2021) learn a conservative estimate of how unsafe each state is and bound the probability of catastrophic failure at each training iteration, with asymptotic performance no worse than standard reinforcement learning; constrained policy optimisation (Achiam, Held, Tamar and Abbeel, 2017) guarantees near-satisfaction of constraints at every iteration. Lipton et al. (2016) attach a learned fear model to a deep Q-network to penalise approach to catastrophic states that agents otherwise revisit periodically after forgetting them. Saunders, Sastry, Stuhlmüller and Evans (2017) let a human block catastrophic actions and train a classifier to imitate the human's interventions; for simple catastrophe classes, they report preventing all catastrophes during a deep reinforcement learning agent's training on Atari games with four hours of human oversight, and less success when catastrophes are harder to recognise.

Safe exploration was already listed among the concrete problems of AI safety a decade ago (Amodei et al., 2016), and the located solutions share a pattern. Each of these borrows a guarantee from something that is not the agent's pessimism: a mentor, a human, a regularity assumption, a known safe starting point, a known constraint signal. Pessimism alone, online, produces an agent that is safe because it does not learn.

5.3 When learning resumes after pessimism

The sharpest evidence about what pessimism costs comes from the setting where both premises are tested in sequence: an agent is trained offline with pessimism and then fine-tuned online. If pessimism's guarantee were a property of the rule, it would carry over; if it were borrowed from the data horizon, it would lapse when the horizon moved.

It lapses. Lee, Seo, Lee, Abbeel and Shin (2021) report that the shift between offline and online state-action distributions can produce severe bootstrapping error at the start of fine-tuning that destroys a good offline policy, and they use pessimistically trained Q-ensembles specifically to prevent over-optimism about unfamiliar actions in novel states during the transition. Nakamoto et al. (2023) document the opposite failure. Conservative methods, they report, tend to "unlearn" the offline policy at the start of fine-tuning and spend online samples recovering it, because conservative value estimates can be far below the true return of any valid policy; once online data arrive, worse actions on the rollout states appear better than the heavily underestimated values of the offline policy, and the optimiser abandons it. Their remedy, Cal-QL, keeps the conservative lower bound on

the learned policy's value but also requires the estimates to exceed the value of a reference policy, which may simply be the behaviour policy. Across eleven tasks, Cal-QL attained the lowest cumulative regret in eight, and its average cumulative regret was 0.22 against 0.41 for conservative Q-learning and 0.96 for an online agent with no offline data, on a normalised scale where 1.00 is worst (Nakamoto et al., 2023, Table 2; Figure 1a).

The same tension organises the rest of the offline-to-online literature. Zheng et al. (2023) apply a pessimistic update to offline data and an optimistic or greedy update to online data within one agent. Zhou, Peng, Li, Levine and Kumar (2025) find that continued training on offline data during fine-tuning mainly serves to prevent a sudden divergence of the value function at the onset of fine-tuning caused by distribution mismatch, and show that a properly designed online method does not need to retain the offline data. Ball, Smith, Kostrikov and Levine (2023) report that ordinary off-policy methods, with a small set of changes, use offline data during online learning well enough to give a 2.5-fold improvement over earlier approaches. On the theoretical side, Song et al. (2023) show that hybrid Q-learning is efficient whenever the offline data support a high-quality policy and the environment has bounded bilinear rank, without assumptions on the coverage of the initial distribution; Wagenmaker and Pacchiano (2023) characterise how many online samples are needed given an offline dataset in linear Markov decision processes; and Xie, Foster, Bai, Jiang and Kakade (2022) show that the mere existence of a well-covering data distribution can make online learning sample-efficient even when the agent does not know the distribution.

These results agree on one point that bears directly on unknown states: the right value for an unvisited state changes when the data horizon changes. A floor that is right when nothing more will be learned is too low when learning continues, and the Cal-QL evidence suggests that the useful correction is not to abandon the floor but to anchor it to a known reference.

5.4 What pessimism has borrowed

Offline, pessimism borrows its guarantee from the end of data collection. Online, it borrows from a mentor, a human overseer, a known safe region or a regularity assumption that lets the unknown be inferred from the known. When learning resumes after an offline phase, the floor that pessimism wrote must be revised, and the revision that worked best in the located evidence anchored it to the value of a known policy.

6. What the averaging rule borrows

6.1 The rule and its guarantee

NIVE is defined precisely. When the aware domain grows, the estimated Q-values, the auxiliary bias-value estimates used by the underlying algorithm, and the

upper bounds on the optimal value function at each new state are all set to their averages over the current aware domain (Zhu et al., 2025, Section 4.3). The learning algorithm is an adaptation of upper confidence bound momentum Q-learning, which acts greedily on upper bounds of the optimal values. Zhu et al. prove that the expanded upper bounds remain valid upper bounds on the optimal values with high probability (their Lemma 1), that a regret bound of order $\sqrt{H^3 S A T} + H^4 S A + H^2 S \sqrt{T}$, up to logarithmic factors, holds with T the number of episodes, and that computation and memory costs are comparable to those of the algorithm they adapt. They describe the result as showing that an unknown unknown, although surprising, will be "asymptotically properly discovered with decent speed and an affordable cost". The paper is theoretical; it reports no experiments.

6.2 The Homeland condition is a ceiling

The guarantee rests on an assumption stated in Section 4.2 of the paper and invoked in the proofs. Zhu et al. call it the Homeland condition, with the gloss that space exploration needs an earth: with probability at least $1 - 2\delta / 3$, for every state outside the aware domain, every step and every action, the optimal Q-value and the optimal state value are no greater than their averages over the aware domain. In words, unknown unknowns are assumed to be worth no more than the average known state. The authors add that a very high-value target may still exist but with high risk, and that the condition matches the high-risk, high-gain character of exploring an under-explored environment.

Read with the condition in view, NIVE is not a neutral rule. Because the averaged quantity is an average of upper bounds, and the condition says that the true value of the new state is below the average true value, the averaged upper bound is itself an upper bound on the new state's value. That is why Lemma 1 holds. Zhu et al.'s Corollary 1 makes the point from the other side: if the average were multiplied by any constant d below one, the expanded value could fail to upper-bound the true value, and the regret argument, which relies on that property, would lose its support. The averaging rule is therefore an optimistic rule. What distinguishes it from the standard optimistic rule, which would value a new state at the maximum possible return, is that the ceiling is supplied by an assumption about the environment rather than by the range of the reward. The assumption does the work that the authors attribute to respect for ignorance.

This matters because the condition is exactly the kind of statement an agent facing a genuine unknown unknown cannot check. The condition is indexed by the aware domain at each episode, so it must hold throughout the agent's history as its aware domain grows. If the discovery process systematically leads to states better than average, which nothing in the model excludes when

discoveries are made by following the agent's own highest upper bounds, the condition fails, and nothing in the agent's experience can tell it so in advance. If the discovery process leads to states much worse than average, the condition holds and the agent's optimism costs it at most an episode, because the environment resets it to the start state. Section 4 showed that this second property, not the averaging, is what makes optimistic visits affordable.

6.3 What the bound says about scale

The regret bound also bears on scale, and Zhu et al. state the relevant condition themselves. The bound for the original algorithm, which Zhu et al. quote, is of order $\sqrt{H^3 S A T} + H^4 S A$. The growing-awareness bound adds a term of order $H^2 S \sqrt{T}$. Zhu et al. state that their bound matches the original algorithm's only when $H \leq A/S$ (Zhu et al., 2025, Section 5.1). Equivalently, the ratio of the added term to the leading term is $\sqrt{H S / A}$, so outside that regime, which includes every case in which states outnumber actions since H is at least one, the added term dominates and the dependence on S changes from $\sqrt{S}$ to S . Both bounds grow as $\sqrt{T}$ in the number of episodes. The bound is an upper bound, and a looser analysis rather than a harder problem may account for the extra term, so the paper draws no conclusion about the true cost of unknown unknowns from it.

6.4 The average presupposes that the new state resembles the old

There is a second sense in which the average is not uninformative. To value a new state at the mean of the known ones is to treat it as a draw from the same population, which is a substantive belief. The decision-theory literature on growing awareness has examined such beliefs directly. Karni and Viero (2013) give a choice-based account of how a decision maker's beliefs should evolve as the set of states she conceives of expands, with updating rules that have the character of what they call reverse Bayesianism; Schipper (2024) summarises the principle as requiring that the relative likelihood of events the decision maker was already aware of remain unchanged when a novel event comes to mind. Schipper relates the principle to exchangeable random partitions, a property central to the statistical literature on discovering new species, and shows that beliefs that are exchangeable in this sense need not satisfy reverse Bayesianism, although the most prominent models of that kind do. The point for present purposes is that both properties are treated as assumptions about a decision maker's beliefs, to be characterised by behaviour and justified or rejected, not as consequences of having no information. Zhu et al. discuss the reverse Bayesian literature in their related work and observe that no practical approach had been proposed for training an agent to handle unknown unknowns. NIVE supplies one by adopting an

exchangeability-like belief for values, which is a choice the decision-theory literature treats as substantive.

Innes and Lascarides (2019) and Innes et al. (2018) approach the same problem in factored Markov decision processes, where the agent may be unaware of variables or actions that matter; their agents discover these through exploration and through corrections from an expert, and they report that conserving information at the moment a new possibility is discovered leads to faster convergence. Their guarantee of convergence to near-optimal behaviour relies on the expert, which places it in the delegated family of Section 9.

6.5 The anchor is the common element

Read together, NIVE and Cal-QL share a structure that the optimism-versus-pessimism framing hides. NIVE caps the value of a new state at a known quantity, the average over the aware domain, and its guarantee depends on the cap being valid. Cal-QL floors the conservative values at a known quantity, the value of a reference policy, and its improvement over plain conservatism comes from the floor. TUCRL restricts optimism to states observed to be reachable. In each case the value written for the unknown is tied to something the agent has observed, and the guarantee, or the measured improvement over the unanchored rule, holds where the anchor is valid. The anchor is not free: it is an assumption that the unknown is bounded by the known, and Section 3.3 shows that no such assumption follows from ignorance. But it is the assumption that the located successes share.

7. What a function-approximating agent actually assigns

The tabular framing makes the value rule a deliberate choice: a new row is added to a table, and something must be written in it. An agent whose values are computed by a neural network has no row to add. Its value for a new input is whatever the network outputs, and the output is set by how the network generalises from what it was trained on.

Rashid et al. (2020) make the consequence explicit. Deep Q-networks are typically initialised so that their outputs start near zero, which, with nonnegative rewards, is the lowest possible value: a pessimistic initialisation. Setting the network to output optimistic values at the start is easy, but it is not enough, because "we cannot predict how a neural network will generalise", and so cannot predict whether values for unvisited state-action pairs will stay optimistic. Their remedy, optimistic pessimistically initialised Q-learning (OPIQ), separates the source of optimism from the network by adding a count-based bonus to the network's estimates during both action selection and bootstrapping, and they report that on hard exploration tasks it outperforms non-optimistic deep Q-network variants that use pseudo-count intrinsic rewards, and that it predicts optimistic

estimates for novel state-action pairs. Machado, Srinivasan and Bowling (2014) note a related limitation of optimistic initialisation: it generally depends on the domain, in that the scale of the rewards must be known and the feature representation must have constant norm.

The extrapolation can also err upward. Van Hasselt, Guez and Silver (2016) showed that deep Q-networks overestimate action values substantially in some Atari games and that reducing the overestimation led to much better performance on several games; Fujimoto et al. (2019) and Kumar et al. (2019) showed that in the offline setting damaging errors arise from actions outside the data and are propagated by bootstrapping. Lee et al. (2021) single out unfamiliar actions at novel states during the offline-to-online transition. The value of a new input under a function approximator is therefore neither optimistic nor pessimistic by design; it is optimistic or pessimistic by accident of generalisation, and the published evidence shows both accidents.

The methods that address this supply an explicit prior over values in regions without data. Bootstrapped deep Q-networks maintain an ensemble of value heads trained on resampled data and act on a randomly chosen head, which produces temporally extended exploration (Osband, Blundell, Pritzel and Van Roy, 2016). Osband, Aslanides and Cassirer (2018) observe that bootstrap ensembles have "no mechanism for uncertainty that does not come from the observed data", and add to each ensemble member a fixed, randomly initialised prior network; this is, in the terms of this paper, a way of writing a value distribution for regions the agent has not visited. Osband, Van Roy, Russo and Wen (2019) treat such methods as randomised value functions and prove a regret bound for them with a tabular representation. Pseudo-counts derived from density models (Bellemare et al., 2016) and the prediction error against a fixed random network (Burda, Edwards, Storkey and Klimov, 2018) give non-tabular agents a count-like signal that can drive an optimistic bonus. Each of these replaces the accident of generalisation with a declared rule, and each is used to drive the agent toward the regions it is uncertain about; none supplies the recoverability that Section 4 showed such exploration requires.

The broader generalisation literature in deep reinforcement learning bears on the same point. Kirk, Zhang, Grefenstette and Rocktaeschel (2023) survey zero-shot generalisation, the problem of policies that must act well in situations unseen at training time, and they argue for more attention to fast online adaptation. A policy that generalises well to a new situation has, in effect, assigned that situation a value close to the truth by similarity to what it has seen. Whether it does so is an empirical property of the representation, not something the value rule controls.

8. Before the value: knowing that a state is new

Every value rule above is conditional on the agent knowing that it is in a new state. In the EMDP-GA model this is given: a state either is or is not in the aware domain. In open environments it is a detection problem, and its failure is silent. An agent that does not notice that a state is new applies whatever value its existing estimates produce, which for a function approximator is the extrapolation of Section 7.

The open-world reinforcement learning literature treats detection and adaptation as separate tasks. Balloch et al. (2022) define novelty as a sudden change to the mechanics or properties of an environment, give an ontology distinguishing novelties that affect objects from those that affect actions and those that change which solutions work, and provide NovGrid, a generator of such changes for gridworld agents, with metrics for evaluating adaptation. NovelGym extends this to hybrid planning and learning agents (Goel et al., 2024). Peng, Balloch and Riedl (2021) represent game states and rules as knowledge graphs so that novelty can be detected as a change in the graph; WorldCloner learns a symbolic transition model of the pre-novelty environment, uses it to detect novelty, and uses imagined transitions to adapt (Balloch et al., 2023). Piotrowski et al. (2023) detect novelty in a planning agent by measuring divergence between observed action outcomes and the outcomes its model predicts, and then search over model revisions. Xue, Nikonova, Zhang and Renz (2023) train agents to adjust only the parts of a policy a novelty affects. Lu et al. (2026) use a language model to identify operators a symbolic planner lacks for novel objects and to write reward functions for learning them. Langley (2020) framed the challenge of agents that respond to sudden, long-term environmental change as one requiring several abilities studied in isolation to be combined, and Kejriwal, Kildebeck, Steininger and Shrivastava (2024) review open-world learning more broadly.

Two features of this literature matter here. First, in each of these systems the novelty is detected as a deviation from a model of the known environment, so the detector inherits the model's coverage: a change the model does not represent cannot register as a deviation. Second, the evaluations located here measure adaptation after novelty in episodic environments; they do not measure what the agent's value estimates assumed about the new situation before it adapted, or what that assumption cost when the situation could not be undone.

The theory of help-seeking under irreversibility makes detection a formal requirement. Plaut et al. (2024) study online learning in which each round's payoff is the chance of avoiding catastrophe and the agent may query a mentor. Without further structure, they prove, every algorithm must choose between asking for help in a constant fraction of rounds and an almost certain

catastrophe. With a learnable mentor policy class and an assumption they call local generalisation, that knowledge transfers between similar inputs and transfers less reliably as inputs become less similar, they give an algorithm whose regret and query rate both go to zero. The algorithm's decision to ask rests on a nearest-neighbour distance: how far the current input is from the closest input at which the mentor has been observed taking the action the agent now proposes. They describe local generalisation as vital because it is "what allows us to detect when an input is unfamiliar", and they show that without it, avoiding catastrophe is impossible even for policy classes that are otherwise learnable. The author's earlier analysis of blind-spot discovery reached a related conclusion from the opposite direction: what counts as an unknown unknown is fixed relative to an external judge, not by the model alone (Mahendrakar, 2026a). In Plaut et al.'s framework the judge is a distance metric under which similarity implies transferable knowledge, and the guarantee is borrowed from the assumption that such a metric exists.

9. The two branches the theory adds: deferral and a way back

When neither resets nor a data horizon is available, the located guarantees that combine safety with learning use one of two further devices.

9.1 Deferral

The first is a mentor. Kosoy (2019) observes that most regret bounds for reinforcement learning are either episodic or assume no traps, and derives a regret bound without either assumption by letting the agent occasionally delegate an action to an advisor. Cohen and Hutter (2020) and Cohen, Catt and Hutter (2020) use a mentor to supply the exploration a pessimistic or cautious agent will not perform; the latter paper's agent, Mentee, explores with a probability that depends on the expected information gain from exploring, and in a simple non-ergodic environment with a weak mentor it is reported to outperform both existing asymptotically optimal agents and the mentor. Plaut et al. (2025) carry the mentor approach past the avoidance of catastrophe to the earning of reward. Their setting allows actions that cannot be undone and a state space without bound; under the usual assumptions of online learning plus local generalisation, they construct an algorithm and prove that, as the horizon lengthens, both its regret and the number of times it consults the mentor grow more slowly than the horizon itself. They suggest that this may be the first formal demonstration that an agent can earn high reward and gradually stop needing help in a high-stakes environment that never resets. They are explicit that their work is not immune to the Heaven or Hell problem; the mentor is their answer to it, offered as a complement to reversibility or periodic resets where those cannot be assumed.

In robot planning with language models, Ren et al. (2023) use conformal prediction to give a language-model planner statistical guarantees on task completion while keeping the amount of human help it requests small; the planner asks for help when its calibrated uncertainty says it does not know which action is intended. The mechanism is the same in kind: the agent's uncertainty about what to do in the current situation is converted into a decision to hand control to someone else rather than into a value.

Deferral borrows its guarantee from the mentor's competence and availability, and, in Plaut et al.'s results, from a similarity metric that tells the agent when to ask. For an unknown unknown, the mentor must know what the agent does not, which is a strong requirement in exactly the cases that motivate the problem.

9.2 Preserving a way back

The second device makes recoverability an objective rather than an assumption. Moldovan and Abbeel (2012) restrict exploration to policies from which the agent can return. Eysenbach, Gu, Ibarz and Levine (2017) learn a reset policy alongside the task policy and use the reset policy's value function to abort the task policy when it is about to enter a state from which it cannot return, which reduces both manual resets and unsafe actions. Recovery RL learns from offline data where constraint violations occur and uses a separate recovery policy to steer the agent back to safety when violation becomes likely (Thananjeyan et al., 2021). Bossens and Bishop (2022) extend the explicit explore-or-exploit scheme of the classical tabular literature to constrained settings with three separate policies: one that improves performance on known states, one that discovers unknown states, and one that returns safely to known states, all optimised against the worst-case model consistent with observations; they prove that their algorithm, E4, finds a near-optimal constraint-satisfying policy in polynomial time while satisfying safety constraints throughout learning. In the side-effects literature, relative reachability penalises reductions in the agent's ability to reach states it could reach from a baseline (Krakovna, Orseau, Kumar, Martic and Legg, 2019), and attainable utility preservation penalises reductions in the agent's ability to optimise auxiliary reward functions, which Turner, Hadfield-Menell and Tadepalli (2020) report induces conservative behaviour even when the auxiliary functions are random.

E4 makes the paper's thesis concrete in an algorithm. Its discovery policy is optimistic about unknown states; its escape policy guarantees a return to known states; its optimisation is pessimistic over models. The three rules that appear in competition in Sections 4 to 6 coexist in one design, each assigned to the part of the problem whose assumption it can borrow: optimism where return is guaranteed, pessimism where the model is uncertain, and an explicit escape that manufactures the

recoverability optimism needs. The guarantee then depends on the escape being possible, which E4's constraint formulation and model uncertainty sets supply.

10. Where the question is live: language-model agents with irreversible actions

The question of what an agent should assume about a situation it has not represented is no longer confined to gridworlds. Language-model agents with tool access act in environments where actions can be irreversible and where the agent's training has not covered the situation it meets. The published evidence about how they behave is measured with instruments different from those of Sections 4 to 9, and the two literatures do not cite each other, but the decision is the same.

Ruan et al. (2024) built an emulated sandbox for testing tool-using agents against 36 high-stakes toolkits and 144 test cases; human evaluation found 68.8 percent of the failures it identified would be valid real-world failures, and the safest agent tested failed in 23.9 percent of cases according to their automatic evaluator. Agent-SafetyBench evaluated 16 agents over 349 interaction environments and 2,000 test cases and reported that none achieved a safety score above 60 percent; its authors conclude that defence prompts alone may be insufficient (Zhang et al., 2024). AgentAbstain pairs 263 tasks that should be carried out with variants that should not, including variants where the reason to stop is discovered during execution, and reports that the best of 17 models reached 59.5 percent paired accuracy and that abstention is largely independent of general task-solving capability (Liu et al., 2026); an earlier analysis by the author, which drew on that benchmark, separated epistemic from consequential triggers for not acting (Mahendrakar, 2026c).

Against the suggestion that prompts are insufficient stands a 2025 result on quitting. Bonagiri, Kumaraguru, Nguyen and Plaut (2025) added an explicit quit action to agents in the emulated sandbox above, under a threat model of benign but underspecified instructions, and compared a baseline prompt with prompts that permit quitting and that specify when to quit. Under the specified prompt, safety scores on a 0 to 3 scale rose by about 0.4 on average across 12 models (0.39 in their abstract, 0.40 in their text), by 0.64 on average for proprietary models, while helpfulness fell by 0.03 on average. The per-model effects broadly track how often the model quits: Claude 4 Sonnet's safety score rose from 1.008 to 2.223 with a 72.22 percent quit rate, while Llama 3.1 8B quit in no tasks and its safety score fell from 0.450 to 0.410 (Figure 1b). The largest helpfulness loss in their table, 0.312 for GPT-4o, came with its 57.64 percent quit rate. Both safety and helpfulness were scored by a language model, Qwen3-32B.

The two results are not in direct conflict. Agent-SafetyBench tested defence prompts against a broad set of risks; Bonagiri et al. tested an exit option against underspecification, where stopping is often the correct action. But together they illustrate the structure of Section 9. Quitting is deferral to the user; it is effective in the measured setting because the user is available and the task is recoverable by asking. The located evidence does not show what an agent that cannot defer, in a situation it cannot recover from, should assume.

Gu et al. (2024) make the recoverability assumption explicit for web agents. Tree search, which has improved web-agent performance in sandboxes, relies on backtracking, and real websites are full of irreversible actions that make backtracking unavailable. Their WebDreamer uses a language model as a world model to simulate the outcome of each candidate action before committing to one, and reports performance competitive with tree search in a sandbox at four to five times the efficiency, and effectiveness on live websites. In this paper's terms, the agent replaces a recoverability assumption it cannot make with a model in which it can explore at no cost, which moves the borrowed guarantee from the environment to the world model's accuracy.

11. The evidence in two tables and a figure

The values the argument depends on, proved or measured, appear in Table 1 with one source result per row. Table 2 maps each rule to the assumption its located guarantees borrow. Figure 1 re-plots two published tables: the offline-to-online regret comparison that shows what a pessimistic floor costs when learning resumes, and the quitting study that shows deferral at work in language-model agents.

Figure 1a separates two things the optimism-versus-pessimism framing runs together. The gap between the online-only agent and both pessimistic agents mainly reflects the offline data, not the value rule; the gap between CQL and Cal-QL, which share data and algorithm and differ in the floor, reflects the value rule. On all eleven tasks Cal-QL's regret is lower than CQL's, and on the pen task it falls from 0.86 to 0.11. Figure 1b shows deferral's safety gains rising with how often the model defers, and helpfulness falling for each of the five models with the highest quit rates; the one model that never quit is the only one whose safety fell.

12. The decision structure

Algorithm 1 is the paper's procedure for classifying an agent design by the assumption its treatment of new states borrows. It does not choose a rule for an agent. It names which assumption a design relies on and where the design's guarantee lapses, given what is known about the deployment.

Applied to the designs in Table 2, the procedure returns ANCHOR for NIVE (line 10, with the Homeland

Table 1. Proved and measured values used in the argument, each row a single source result. "Proved" marks a theorem under the source's stated assumptions; other rows are measurements reported by the source. Values are as reported; differences in Figure 1b are computed from the source's table.

Source	Setting	Rule or device	Reported result
Jin et al. 2018	Tabular, episodic	Optimistic (UCB) Q-learning	Regret $\sim \sqrt{H^3 S A T}$ (proved)
Azar et al. 2017	Tabular, episodic	Optimistic value iteration	Minimax rate for large T (proved)
Fruit et al. 2018	Weakly communicating	UCRL, optimistic PSRL	Linear regret
Cohen, Catt, Hutter 2020	General, non-ergodic	Asymptotic optimality	Destroyed or incapacitated w.p. 1, under No Heaven (proved)
Zhu et al. 2025	Episodic, growing awareness	Average of known values	Regret adds $H^2 S \sqrt{T}$ term, under Homeland (proved)
Jin, Yang, Wang 2021	Offline, linear MDP	Pessimistic value iteration	Minimax optimal up to d, H factors (proved)
Kumar et al. 2020	Offline, deep	Conservative Q-learning	2-5x final return of prior offline methods
Nakamoto et al. 2023	Offline to online, 11 tasks	CQL vs Cal-QL vs SAC	Mean regret 0.41 vs 0.22 vs 0.96; Cal-QL best on 8
Ball et al. 2023	Online with offline data	Off-policy with small changes	2.5x over prior approaches
Saunders et al. 2017	Atari, human oversight	Learned intervention	All catastrophes blocked for simple classes; 4 h of oversight
Plaut et al. 2024	Online, catastrophe	Mentor + local generalisation	Without structure: linear queries or catastrophe (proved)
Plaut et al. 2025	Irreversible MDP, no resets	Mentor + local generalisation	Sublinear regret and queries (proved)
Ruan et al. 2024	LM agents, 144 cases	None (baseline)	Safest agent fails 23.9%; 68.8% of failures valid
Zhang et al. 2024	LM agents, 2,000 cases	None (baseline)	No agent above 60% safety
Liu et al. 2026	LM agents, 263 pairs	Abstention	Best paired accuracy 59.5%
Bonagiri et al. 2025	LM agents, 12 models	Specified quit prompt	Safety +0.40 (0-3); helpfulness -0.03
Gu et al. 2024	Web agents	World-model planning	Competitive with tree search at 4-5x efficiency

Table 2. What each rule writes for a new state, what its located guarantees borrow, and where they lapse.

Rule	Value written	Guarantee borrowed from	Lapses when	Sources
Optimistic	Upper bound	Resets or communicating MDP	States unreachable or irreversible	Jin 2018; Fruit 2018; Cohen, Catt, Hutter 2020
Pessimistic, offline	Lower bound	End of data collection	Learning resumes	Jin 2021; Nakamoto 2023
Pessimistic, online	Lower bound	Mentor, safe region, regularity	No mentor or seed	Cohen, Hutter 2020; Turchetta 2016
Anchored	Known reference	Validity of the anchor	Anchor wrong	Zhu 2025; Nakamoto 2023
Delegated	None; ask	Mentor and similarity metric	Mentor absent or also ignorant	Plaut 2024, 2025; Kosoy 2019
Escape	Return to known	Escape feasibility	No safe return exists	Bossens 2022; Eysenbach 2017
Implicit (deep)	Extrapolation	Representation generalising	Novel input misvalued	Rashid 2020; Fujimoto 2019

Algorithm 1. Classify how a design values a state it did not know existed Input: design D; deployment E Output: one of {UNDETECTED, RESTRICT, OPTIMISE, ANCHOR, DEFER, ESCAPE} + borrowed assumption

<pre> 1 if D treats unseen inputs like seen ones (no novelty test, count, penalty or prior): 2 label <- UNDETECTED # value = extrapolation of the representation (Sec. 7) 3 borrowed <- "representation generalises to the new state" 4 return label, borrowed 5 record the detector's assumption (model coverage, similarity metric) (Sec. 8) 6 if no further data will be collected in E after the value is written: 7 return RESTRICT, "data horizon: nothing more can be learned" 8 if E guarantees return (episodic reset, communicating, reversible actions): 9 if D bounds the new value above or below by a known quantity (average over known states, value of a reference policy): 10 return ANCHOR, "recoverability + validity of the anchor" 11 return OPTIMISE, "recoverability; fails if states unreachable (Sec. 4.2)" 12 if D can defer to a mentor that is available in E: 13 return DEFER, "mentor competence + similarity metric for when to ask" 14 if D has a policy that returns to known states with stated confidence: 15 return ESCAPE, "feasibility of the escape under model uncertainty" 16 return RESTRICT, "none: the agent stays in known states and may never learn" </pre>

condition as the cap and episodic resets as the return), OPTIMISE for upper-confidence Q-learning in episodic settings (line 11), RESTRICT for conservative Q-learning deployed offline (line 7), and ANCHOR for Cal-QL when online fine-tuning follows (line 10 if the fine-tuning environment resets, as the benchmark tasks do; otherwise the procedure reaches line 12 and, without a mentor, line 16). It returns DEFER for the agents of Cohen and Hutter, Kosoy and Plaut et al. (line 13), ESCAPE for E4 and for the reset-policy design of Eysenbach et al. (line 15), and UNDETECTED for a standard deep Q-network and for a language-model agent with no mechanism that treats unfamiliar situations differently (line 2), whatever value its network happens to produce. A language-model agent given a quit action has a crude substitute for line 1, its own judgement that it lacks confidence; if that judgement is accepted as the test, the agent is DEFER when a user can answer it and RESTRICT otherwise.

Two consequences follow. First, line 16 is where unknown unknowns in the strict sense land: an environment without resets, without a data horizon, without a mentor and without a known escape. No located guarantee covers that branch. Second, line 1 is where most deployed agents land, because function approximators do not represent the boundary of their own knowledge unless something is added to make them do so.

13. Comparisons that would close the open questions

Every comparison below needs only environments and algorithms that already exist.

Score the averaging rule where the Homeland condition fails. Construct episodic environments in which states outside an initial aware domain are, by design, better than the known average, and compare NIVE with the maximum-value optimistic rule and with a pessimistic rule. With the Homeland condition violated, NIVE's regret guarantee no longer applies; whether its empirical regret degrades gracefully or sharply is unmeasured. Repeat with discovered states worse than average, where the condition holds.

Remove the reset. Run the same comparison in a version of each environment where entering certain new states ends the agent's life rather than its episode. The prediction of Section 4 is that every optimistic rule, anchored or not, incurs losses that the episodic regret bound does not see, and that the ordering among rules changes. The relevant measure is the probability of avoiding an absorbing low-value state over the agent's lifetime, the quantity Plaut et al. (2024) maximise, rather than regret.

Measure what a deep agent assigns. For a trained deep Q-network, record the value estimates it produces on inputs from a novelty generator such as NovGrid

(Balloch et al., 2022), before any post-novelty training, and compare them with the true values. Classify each novelty by whether the agent's implicit value was optimistic or pessimistic. Rashid et al. (2020) argue that the sign is unpredictable a priori; the distribution of signs across novelty types is unmeasured.

Separate the floor from the data in offline-to-online learning. Nakamoto et al. (2023) compare CQL and Cal-QL under shared data. A design that varies the floor continuously, from no floor to the behaviour policy's value to the offline policy's own estimated value, would show whether the regret is monotone in the floor and whether the best floor depends on dataset coverage, which their analysis of narrow datasets suggests.

Test deferral when the mentor is also ignorant. In Plaut et al.'s framework, construct inputs that are novel for the mentor as well as for the agent. The prediction is that the guarantee lapses; the empirical question is whether agents that ask for help on unfamiliar inputs fail more or less badly than agents that act.

Report quitting against irreversibility, not only underspecification. Bonagiri et al. (2025) measure quitting under underspecified instructions. Splitting the same evaluation by whether the risky action is reversible, using the emulated sandbox's tool descriptions, would show whether quitting rates track irreversibility or only ambiguity. AgentAbstain's paired design (Liu et al., 2026) makes such a split possible within one benchmark.

Score world-model accuracy where it carries the guarantee. For WebDreamer-style agents (Gu et al., 2024), measure the world model's prediction error specifically on actions whose consequences are irreversible, since that is where the agent has replaced an environmental assumption with a modelled one.

14. The strongest case against this paper

This is Plaut et al.'s taxonomy restated. Plaut et al. (2025) present a taxonomy of the assumptions that make no-regret guarantees possible in Markov decision processes, recoverability, prior knowledge and external help, with the Heaven or Hell problem as the motivation. This paper's account of what optimism and the mentor branch borrow agrees with theirs and depends on it. The paper's additions are three. It applies the analysis to the specific question of what value a newly discovered state should receive, where the rival rules are each defended by their proponents as principled. It adds the data horizon as a separate source of guarantee, which accounts for pessimism's offline success and its online costs, and reads the offline-to-online evidence as a test of that account. And it shows that the averaging rule proposed for unknown unknowns is an optimistic rule whose cap is an environmental assumption. If these additions were already stated together in a located source, the contribution would reduce to a consolidation;

the search found none.

The Homeland condition is reasonable. Zhu et al. defend it by analogy with exploration needing a base, and some environments may well have their best states near where an agent starts. That may be true, and the paper does not claim the condition is usually false. It claims the condition is an assumption about the environment, not a consequence of the agent's ignorance, and that an agent facing an unknown unknown has no way to test it before relying on it.

Episodic settings are the right model for most learning. Much reinforcement learning research uses simulators that reset, and much learning for deployment happens in simulation first. That is a fair description of common practice. It concedes the paper's claim that the guarantees are borrowed from resets, and it leaves open the deployment case, where an agent trained with resets meets a state it never saw and cannot reset from. That case is the one the phrase "unknown unknown" describes.

Deep agents with ensembles and priors do represent uncertainty about novel inputs. Randomised prior functions and ensembles give an agent a value distribution in unvisited regions (Osband et al., 2018), and the paper says so in Section 7. Those methods are used to direct the agent toward the uncertain region, and so they inherit exploration's dependence on recoverability. They address detection and valuation, not the cost of being wrong.

The language-model evidence is not about unknown states. The agent benchmarks in Section 10 measure risky behaviour under underspecified instructions and safety-relevant scenarios, not discovery of states outside a representation. That is correct, and the paper uses them only to show that the safety mechanisms measured for these agents are of the deferral kind and that the measured effect of deferral depends on someone being available to defer to. The paper does not claim that those benchmarks measure what an agent assumes about the genuinely new.

Pessimism is the safe default, so the analysis changes nothing in practice. Line 16 of Algorithm 1 returns RESTRICT, which is pessimism. For a deployment where safety dominates, that may be correct. The paper's point is that the default has a price that the safety framing hides: a fully restricted agent cannot learn what the new state is worth, and Nakamoto et al.'s evidence shows that a floor set too low has a measurable cost the moment learning resumes. Choosing pessimism is choosing to borrow from the data horizon, and the choice should be declared as such.

15. What is not known

How NIVE or any anchored rule performs empirically when its cap is violated is unknown, because no located study has tested it; the EMDP-GA paper is theoretical. It is not known what value a trained deep reinforcement learning agent implicitly assigns to genuinely novel states, or whether the sign of its error is predictable from the type of novelty; no located study measures the value estimates themselves before adaptation. It is not known whether any agent without resets, a mentor or a known escape can learn safely in an environment with irreversible states; no located guarantee covers that case, and Cohen, Catt and Hutter's result shows that, under their assumption, a guarantee of asymptotic optimality is incompatible with avoiding both destruction and incapacitation. It is not known whether language-model agents' quitting behaviour tracks the irreversibility of the action or only the ambiguity of the instruction. It is not known how often deployed agents encounter states their designers would classify as unknown unknowns, so the practical weight of line 16 relative to the other branches is unmeasured.

16. Limitations

The central claim concerns the guarantees this review located, and the review searched arXiv and Crossref only, so it may have missed relevant work. A guarantee for an unanchored rule that holds without resets, a data horizon or external help would refute the claim in its general form; none was found, and Cohen, Catt and Hutter's theorem and Plaut et al.'s Heaven or Hell argument suggest why, but the absence of a located result is not a proof that none exists.

Several sources are idealised. Cohen and Hutter's pessimistic agent and the asymptotically optimal agents of Cohen, Catt and Hutter are defined over general computable environment classes and are not implementable as stated. The paper uses them for what they show about the structure of the problem, not as descriptions of practical systems.

The comparison of bounds in Section 6.3 compares upper bounds, and a looser analysis rather than a harder problem may account for the difference. The paper draws no conclusion about the true cost of unknown unknowns from it.

The language-model evidence in Section 10 comes from benchmarks that use language models as emulators and judges. Ruan et al. (2024) tested their emulator and evaluator through human evaluation; Bonagiri et al. (2025) used a single judge model. The measured effects inherit those instruments' errors.

Figure 1a's comparison between CQL and Cal-QL is the cleanest located test of a value floor, but Cal-QL changes how the conservative objective is computed, not only the floor's level, and the benchmark environments reset. The

figure shows that anchoring a floor to a reference reduced regret in that setting; it does not show what the same anchor would do in an environment without resets.

17. Conclusion

An agent that reaches a state it did not know existed must give it a value, and the reinforcement learning literature offers three rules that point in opposite directions: value it high and go and look, value it low and stay away, or value it at the average of the known. Each has a theorem. Read side by side, the theorems show that none of the rules is justified by the agent's ignorance about the state. Optimism's efficiency results hold where the agent can always return; where some states cannot be reached, optimistic algorithms incur linear regret, and where some cannot be left, guaranteed asymptotic optimality implies destruction or incapacitation under a mild condition. Pessimism's minimax results hold where no more data will arrive; when learning resumes, a floor set too low throws away a good policy, and a floor anchored to a reference policy cut average regret from 0.41 to 0.22 on one suite. The averaging rule proposed for unknown unknowns is optimistic under a stated ceiling, the assumption that unknown states are worth no more than the known average, and it lives in an episodic world that returns the agent to its start. The successful designs share an anchor to something observed, and the located theory adds two devices for when no anchor or return is given: a mentor to defer to and an explicit way back. What an agent should assume about the genuinely unknown, in an environment that does not reset, with no one to ask and no known escape, is the case no located guarantee covers. The paper has stated it as the last line of a decision procedure, so that a design that lands there can at least say so.

References

- Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained Policy Optimization. ICML 2017. arXiv:1705.10528. doi:10.48550/arXiv.1705.10528
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565. doi:10.48550/arXiv.1606.06565
- Azar, M. G., Osband, I., & Munos, R. (2017). Minimax Regret Bounds for Reinforcement Learning. arXiv:1703.05449. doi:10.48550/arXiv.1703.05449
- Ball, P. J., Smith, L., Kostrikov, I., & Levine, S. (2023). Efficient Online Reinforcement Learning with Offline Data. ICML 2023. arXiv:2302.02948. doi:10.48550/arXiv.2302.02948
- Balloch, J., Lin, Z., Hussain, M., Srinivas, A., Wright, R., Peng, X., Kim, J., & Riedl, M. (2022). NovGrid: A Flexible Grid World for Evaluating Agent Response to Novelty. arXiv:2203.12117. doi:10.48550/arXiv.2203.12117
- Balloch, J., Lin, Z., Wright, R., Peng, X., Hussain, M., Srinivas, A., Kim, J., & Riedl, M. O. (2023). Neuro-Symbolic World Models for Adapting to Open World Novelty. arXiv:2301.06294. doi:10.48550/arXiv.2301.06294
- Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016). Unifying Count-Based Exploration and Intrinsic Motivation. arXiv:1606.01868. doi:10.48550/arXiv.1606.01868
- Bharadhwaj, H., Kumar, A., Rhinehart, N., Levine, S., Shkurti, F., & Garg, A. (2021). Conservative Safety Critics for Exploration. ICLR 2021. arXiv:2010.14497. doi:10.48550/arXiv.2010.14497
- Bonagiri, V. K., Kumaraguru, P., Nguyen, K., & Plaut, B. (2025). Check Yourself Before You Wreck Yourself: Selectively Quitting Improves LLM Agent Safety. arXiv:2510.16492. doi:10.48550/arXiv.2510.16492
- Bossens, D. M., & Bishop, N. (2022). Explicit Explore, Exploit, or Escape (E4): near-optimal safety-constrained reinforcement learning in polynomial time. Machine Learning. doi:10.1007/s10994-022-06201-z
- Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). Exploration by Random Network Distillation. arXiv:1810.12894. doi:10.48550/arXiv.1810.12894
- Cohen, M. K., & Hutter, M. (2020). Pessimism About Unknown Unknowns Inspires Conservatism. COLT 2020. arXiv:2006.08753. doi:10.48550/arXiv.2006.08753
- Cohen, M. K., Catt, E., & Hutter, M. (2020). Curiosity Killed or Incapacitated the Cat and the Asymptotically Optimal Agent. IEEE Journal on Selected Areas in Information Theory (2021). arXiv:2006.03357. doi:10.48550/arXiv.2006.03357
- Eysenbach, B., Gu, S., Ibarz, J., & Levine, S. (2017). Leave no Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning. arXiv:1711.06782. doi:10.48550/arXiv.1711.06782
- Fruit, R., Pirotta, M., & Lazaric, A. (2018). Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes. arXiv:1807.02373. doi:10.48550/arXiv.1807.02373
- Fujimoto, S., Meger, D., & Precup, D. (2019). Off-Policy Deep Reinforcement Learning without Exploration. ICML 2019. arXiv:1812.02900. doi:10.48550/arXiv.1812.02900
- Goel, S., Wei, Y., Lymperopoulos, P., Chura, K., Scheutz, M., & Sinapov, J. (2024). NovelGym: A Flexible Ecosystem for Hybrid Planning and Learning Agents Designed for Open Worlds. AAMAS 2024. arXiv:2401.03546. doi:10.48550/arXiv.2401.03546
- Gu, Y., Zhang, K., Ning, Y., Zheng, B., Gou, B., Xue, T., Chang, C., Srivastava, S., Xie, Y., Qi, P., Sun, H., & Su, Y. (2024). Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents. arXiv:2411.06559. doi:10.48550/arXiv.2411.06559
- Innes, C., & Lascarides, A. (2019). Learning Factored Markov Decision Processes with Unawareness. arXiv:1902.10619. doi:10.48550/arXiv.1902.10619
- Innes, C., Lascarides, A., Albrecht, S. V., Ramamoorthy, S., & Rosman, B. (2018). Reasoning about Unforeseen Possibilities During Policy Learning. arXiv:1801.03331. doi:10.48550/arXiv.1801.03331
- Jin, C., Allen-Zhu, Z., Bubeck, S., & Jordan, M. I. (2018). Is Q-learning Provably Efficient? arXiv:1807.03765. doi:10.48550/arXiv.1807.03765
- Jin, Y., Yang, Z., & Wang, Z. (2021). Is Pessimism Provably Efficient for Offline RL? arXiv:2012.15085. doi:10.48550/arXiv.2012.15085
- Karni, E., & Viero, M.-L. (2013). Reverse Bayesianism: A Choice-Based Theory of Growing Awareness. American Economic Review, 103(7). doi:10.1257/aer.103.7.2790
- Kejriwal, M., Kildebeck, E., Steininger, R., & Shrivastava, A. (2024). Challenges, evaluation and opportunities for open-world learning. Nature Machine Intelligence. doi:10.1038/s42256-024-00852-4
- Kidambi, R., Rajeswaran, A., Netrapalli, P., & Joachims, T. (2020). MOREL: Model-Based Offline Reinforcement Learning. NeurIPS 2020. arXiv:2005.05951. doi:10.48550/arXiv.2005.05951
- Kirk, R., Zhang, A., Grefenstette, E., & Rocktaeschel, T. (2023). A Survey of Zero-shot Generalisation in Deep Reinforcement Learning. Journal of Artificial Intelligence Research, 76. arXiv:2111.09794. doi:10.48550/arXiv.2111.09794
- Kosoy, V. (2019). Delegative Reinforcement Learning: learning to avoid traps with a little help. SafeML ICLR 2019 Workshop.

- arXiv:1907.08461. doi:10.48550/arXiv.1907.08461
- Krakovna, V., Orseau, L., Kumar, R., Martic, M., & Legg, S. (2019). Penalizing side effects using stepwise relative reachability. arXiv:1806.01186. doi:10.48550/arXiv.1806.01186
- Kumar, A., Fu, J., Tucker, G., & Levine, S. (2019). Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. NeurIPS 2019. arXiv:1906.00949. doi:10.48550/arXiv.1906.00949
- Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-Learning for Offline Reinforcement Learning. arXiv:2006.04779. doi:10.48550/arXiv.2006.04779
- Langley, P. (2020). Open-World Learning for Radically Autonomous Agents. Proceedings of the AAAI Conference on Artificial Intelligence, 34(09). doi:10.1609/aaai.v34i09.7078
- Lee, S., Seo, Y., Lee, K., Abbeel, P., & Shin, J. (2021). Offline-to-Online Reinforcement Learning via Balanced Replay and Pessimistic Q-Ensemble. CoRL 2021. arXiv:2107.00591. doi:10.48550/arXiv.2107.00591
- Leike, J., & Hutter, M. (2015). Bad Universal Priors and Notions of Optimality. COLT 2015. arXiv:1510.04931. doi:10.48550/arXiv.1510.04931
- Leike, J., Lattimore, T., Orseau, L., & Hutter, M. (2016). Thompson Sampling is Asymptotically Optimal in General Environments. UAI 2016. arXiv:1602.07905. doi:10.48550/arXiv.1602.07905
- Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv:2005.01643. doi:10.48550/arXiv.2005.01643
- Lipton, Z. C., Azizzadenesheli, K., Kumar, A., Li, L., Gao, J., & Deng, L. (2016). Combating Reinforcement Learning's Sisyphean Curse with Intrinsic Fear. arXiv:1611.01211. doi:10.48550/arXiv.1611.01211
- Liu, X., Zhang, Y. E., Kasprova, V., Rabbani, P., Zahraei, P. S., Zhang, T., Ebrahimpour-Borojeny, A., & Chandrasekaran, V. (2026). AgentAbstain: Do LLM Agents Know When Not to Act? arXiv:2607.10059. doi:10.48550/arXiv.2607.10059
- Lu, H., Lorang, P., Duggan, T. R., Sinapov, J., & Scheutz, M. (2026). Novelty Adaptation Through Hybrid Large Language Model (LLM)-Symbolic Planning and LLM-guided Reinforcement Learning. IROS 2026. arXiv:2603.11351. doi:10.48550/arXiv.2603.11351
- Machado, M. C., Srinivasan, S., & Bowling, M. (2014). Domain-Independent Optimistic Initialization for Reinforcement Learning. arXiv:1410.4604. doi:10.48550/arXiv.1410.4604
- Mahendrakar, P. (2026a). An Unknown Unknown Is Defined Relative to an Oracle: Four Places External Judgement Enters Blind-Spot Discovery, and Why the Self-Detection Results Do Not Remove It. Life of Research. doi:10.5281/zenodo.22217013
- Mahendrakar, P. (2026b). The Reassessment That Did Not Travel: What the 2020 Arcade Result About Exploration Bonuses Establishes, the Six Conditions That Made It Informative, and Which of Them the Language-Model Revival Drops. Life of Research. doi:10.5281/zenodo.22748437
- Mahendrakar, P. (2026c). Not Acting Is Not One Decision: Two Abstention Triggers in Tool-Using Agents, Why the Calibration and Permission Readings Cover Different Ones, and the Cost Term Neither Supplies. Life of Research. doi:10.5281/zenodo.22803148
- Moldovan, T. M., & Abbeel, P. (2012). Safe Exploration in Markov Decision Processes. arXiv:1205.4810. doi:10.48550/arXiv.1205.4810
- Nakamoto, M., Zhai, Y., Singh, A., Mark, M. S., Ma, Y., Finn, C., Kumar, A., & Levine, S. (2023). Cal-QL: Calibrated Offline RL Pre-Training for Efficient Online Fine-Tuning. NeurIPS 2023. arXiv:2303.05479. doi:10.48550/arXiv.2303.05479
- Osband, I., Aslanides, J., & Cassirer, A. (2018). Randomized Prior Functions for Deep Reinforcement Learning. arXiv:1806.03335. doi:10.48550/arXiv.1806.03335
- Osband, I., Blundell, C., Pritzel, A., & Van Roy, B. (2016). Deep Exploration via Bootstrapped DQN. arXiv:1602.04621. doi:10.48550/arXiv.1602.04621
- Osband, I., & Van Roy, B. (2017). Why is Posterior Sampling Better than Optimism for Reinforcement Learning? arXiv:1607.00215. doi:10.48550/arXiv.1607.00215
- Osband, I., Van Roy, B., Russo, D., & Wen, Z. (2019). Deep Exploration via Randomized Value Functions. Journal of Machine Learning Research. arXiv:1703.07608. doi:10.48550/arXiv.1703.07608
- Peng, X., Balloch, J. C., & Riedl, M. O. (2021). Detecting and Adapting to Novelty in Games. arXiv:2106.02204. doi:10.48550/arXiv.2106.02204
- Piotrowski, W., Stern, R., Sher, Y., Le, J., Klenk, M., deKleer, J., & Mohan, S. (2023). Learning to Operate in Open Worlds by Adapting Planning Models. AAMAS 2023. arXiv:2303.14272. doi:10.48550/arXiv.2303.14272
- Plaut, B., Zhu, H., & Russell, S. (2024). Avoiding Catastrophe in Online Learning by Asking for Help. ICML 2025. arXiv:2402.08062. doi:10.48550/arXiv.2402.08062
- Plaut, B., Lievano-Karim, J., Zhu, H., & Russell, S. (2025). Safe Learning Under Irreversible Dynamics via Asking for Help. arXiv:2502.14043. doi:10.48550/arXiv.2502.14043
- Rashid, T., Peng, B., Boehmer, W., & Whiteson, S. (2020). Optimistic Exploration even with a Pessimistic Initialisation. ICLR 2020. arXiv:2002.12174. doi:10.48550/arXiv.2002.12174
- Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., & Russell, S. (2021). Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. NeurIPS 2021. arXiv:2103.12021. doi:10.48550/arXiv.2103.12021
- Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L., Xia, F., Varley, J., Xu, Z., Sadigh, D., et al. (2023). Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. CoRL 2023. arXiv:2307.01928. doi:10.48550/arXiv.2307.01928
- Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., & Hashimoto, T. (2024). Identifying the Risks of LM Agents with an LM-Emulated Sandbox. arXiv:2309.15817. doi:10.48550/arXiv.2309.15817
- Saunders, W., Sastry, G., Stuhlmüller, A., & Evans, O. (2017). Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. arXiv:1707.05173. doi:10.48550/arXiv.1707.05173
- Schipper, B. C. (2024). Predicting the Unpredictable under Subjective Expected Utility. arXiv:2403.01421. doi:10.48550/arXiv.2403.01421
- Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., & Sun, W. (2023). Hybrid RL: Using Both Offline and Online Data Can Make RL Efficient. ICLR 2023. arXiv:2210.06718. doi:10.48550/arXiv.2210.06718
- Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Gonzalez, J. E., Ibarz, J., Finn, C., & Goldberg, K. (2021). Recovery RL: Safe Reinforcement Learning with Learned Recovery Zones. IEEE Robotics and Automation Letters. arXiv:2010.15920. doi:10.48550/arXiv.2010.15920
- Turchetta, M., Berkenkamp, F., & Krause, A. (2016). Safe Exploration in Finite Markov Decision Processes with Gaussian Processes. NIPS 2016. arXiv:1606.04753. doi:10.48550/arXiv.1606.04753
- Turner, A. M., Hadfield-Menell, D., & Tadepalli, P. (2020). Conservative Agency via Attainable Utility Preservation. AIES 2020. arXiv:1902.09725. doi:10.48550/arXiv.1902.09725
- van Hasselt, H., Guez, A., & Silver, D. (2016). Deep Reinforcement Learning with Double Q-learning. AAAI 2016. arXiv:1509.06461. doi:10.48550/arXiv.1509.06461
- Wagenmaker, A., & Pacchiano, A. (2023). Leveraging Offline Data in Online Reinforcement Learning. arXiv:2211.04974. doi:10.48550/arXiv.2211.04974
- Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., & Agarwal, A. (2021). Bellman-consistent Pessimism for Offline Reinforcement Learning. NeurIPS 2021. arXiv:2106.06926. doi:10.48550/arXiv.2106.06926

- arXiv:2106.06926
- Xie, T., Foster, D. J., Bai, Y., Jiang, N., & Kakade, S. M. (2022). The Role of Coverage in Online Reinforcement Learning. arXiv:2210.04157. doi:10.48550/arXiv.2210.04157
- Xue, C., Nikonova, E., Zhang, P., & Renz, J. (2023). Rapid Open-World Adaptation by Adaptation Principles Learning. arXiv:2312.11138. doi:10.48550/arXiv.2312.11138
- Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., & Ma, T. (2020). MOPO: Model-based Offline Policy Optimization. NeurIPS 2020. arXiv:2005.13239. doi:10.48550/arXiv.2005.13239
- Zhang, Z., Cui, S., Lu, Y., Zhou, J., Yang, J., Wang, H., & Huang, M. (2024). Agent-SafetyBench: Evaluating the Safety of LLM Agents. arXiv:2412.14470. doi:10.48550/arXiv.2412.14470
- Zheng, H., Luo, X., Wei, P., Song, X., Li, D., & Jiang, J. (2023). Adaptive Policy Learning for Offline-to-Online Reinforcement Learning. AAAI 2023. arXiv:2303.07693. doi:10.48550/arXiv.2303.07693
- Zhou, Z., Peng, A., Li, Q., Levine, S., & Kumar, A. (2025). Efficient Online Reinforcement Learning Fine-Tuning Need Not Retain Offline Data. ICLR 2025. arXiv:2412.07762. doi:10.48550/arXiv.2412.07762
- Zhu, J., de Carvalho, M., Yang, Z., & He, F. (2025). When a Reinforcement Learning Agent Encounters Unknown Unknowns. arXiv:2505.13188. doi:10.48550/arXiv.2505.13188