

Reading status. The detailed review below preserves the earlier literature synthesis and its proposed ARIA-NBV adaptations. Recommendations about model inputs, training stages or implementation are research proposals, not a description of the current system or evidence of measured benefit. The active thesis and implementation specify the current choices.

RL Sources For Rollout And Q_H

This page owns the source-backed RL distillation used for thesis writing. The formal ARIA-NBV rollout notation, replay fields, masks, Q_H targets, and implementation acceptance checks live in the finite-candidate rollout and Q_H contract.

Primary sources. Trajectory Transformer [1], Gumbel-Top-k / stochastic beam search [2], DQN [3], Double DQN [4], Implicit Q-Learning [5], Deep Energy-Based Policies / soft Q-learning [6], PPO [7], and SAC [8].

Related ARIA-NBV pages. rollout contract, VIN-NBV, GenNBV, Hestia, roadmap, and research questions.

Core contribution

The RL literature is useful for ARIA-NBV only after the finite-candidate rollout substrate is trusted. The current thesis core is not unrestricted continuous control. It is target-conditioned finite-candidate value learning over offline Oracle RRI (oracle RRI) rollout data, with the formal model defined in the contract page.

[Paper signals] $\rightarrow$ [Rollout data contract]
 [Rollout data contract] $\rightarrow$ [Masked candidate-to-state Q_H]
 [Masked candidate-to-state Q_H] $\rightarrow$ [Optional IQL / actor-critic bridge]

Local source anchors

Use these local source files when tracing thesis claims back to papers:

paper	local source anchors	claim to reuse
DQN	docs/literature/tex-src/arXiv-DQN/intro.tex, background.tex, method.tex, experiments.tex	delayed rewards, correlated samples, behavior-distribution drift, experience replay, Bellman Q-learning, random minibatches, transition reuse, held-out predicted-Q diagnostics
Double DQN	docs/literature/tex-src/arXiv-Double-DQN/DoubleDQN_aaai2016_total.tex	max-operator overestimation and online-selector / target-evaluator decoupling
IQL	docs/literature/tex-src/arXiv-IQL/iclr2022_conference.tex	offline support constraint, in-sample SARSA-style targets, upper expectile value fitting, Q backup through $V(s')$, advantage-weighted extraction

DQN, Double DQN, and IQL for Q_H

DQN contributes the basic learning pattern for ARIA-NBV's Finite-Horizon Q Function (Q_H). It identifies sparse or delayed rewards, correlated streams, and non-stationary behavior distributions as stability problems for deep RL, then uses experience replay to store transitions and train from random minibatches. It also motivates batched discrete-action scoring: the Atari network emits one value per valid action in one forward pass, while ARIA-NBV replaces those fixed heads with candidate-to-state query tokens over a variable candidate table [3].

Double DQN is the first-path overestimation safeguard. Its source argues that Q-learning's maximization step can prefer overestimated action values and shows that separating action selection from evaluation reduces this bias. The earlier ARIA-NBV proposal adopted that idea as a masked selector/evaluator backup over valid candidate-table entries; the exact equations live in the training contract [4].

IQL contributes the offline support rule. Its source emphasizes that offline RL should avoid querying values for unseen or out-of-distribution actions, then uses upper-expectile value fitting and a Q backup through the learned state value to perform multi-step dynamic programming without direct out-of-sample action queries. ARIA-NBV keeps this as a gated ablation after masked fitted Double-Q, because the first need is a reliable rollout store and support-aware candidate table [5].

Paper signals for ARIA-NBV

paper family	source-backed signal	ARIA-NBV adoption	deferred or rejected
Trajectory Transformer	Offline control can be modeled as sequences of states, actions, and rewards; beam search can decode high-return trajectories.	Use bounded rollout and beam-search abstractions before training a large sequence model.	Do not start with a trajectory Transformer before typed rollout traces and oracle labels are trusted.
Gumbel-Top-k	Ordered size- k samples can be drawn without replacement, and stochastic beam search avoids enumerating the full sequence space.	Use after deterministic lookahead to diversify rollout data and reduce duplicate root-greedy traces.	Do not present stochastic beams as a substitute for deterministic oracle lookahead.
DQN	Replayed minibatch Q-learning can reuse transitions, decorrelate sequential samples, and score all discrete valid actions in one forward pass.	Adopt replayed transition learning and batched finite-candidate scoring with candidate-to-state query tokens.	Do not import Atari CNNs, epsilon-greedy schedules, score clipping, or emulator hyperparameters as defaults.
Double DQN	Separating action selection from action evaluation reduces max-operator overestimation.	The earlier proposal prioritized masked fitted Double-Q as its first learned Q_H method.	It reduces overestimation but does not solve offline support mismatch by itself.
IQL	Offline RL should avoid querying unseen actions; value fitting uses an upper expectile over dataset actions before advantage-weighted policy extraction.	Keep as a second offline-RL ablation after the fitted Double-Q dataset and masks are stable.	Do not use IQL to skip the required finite-candidate Q_H result.
Soft Q / energy policies	Maximum-entropy policies can represent multimodal action distributions through energy/ Q -shaped sampling.	Use as conceptual support for temperature-softmax candidate selection.	Not a first thesis algorithm.
PPO / SAC	Practical online actor-critic methods assume an interactive reward loop and simulator or environment abstraction.	Simulator-gated bridge only.	Not a required quantitative continuous-control result.

Thesis-writing use

For writing, cite DQN when motivating replayed finite-action value learning, Double DQN when motivating the masked selector/evaluator backup, and IQL when motivating the offline support constraint. Cite Trajectory Transformer and Gumbel-Top-k when justifying bounded beams, branch schedules, and stochastic rollout data

diversity. Cite PPO/SAC and GenNBV/Hestia only to position continuous actor-critic work as bridge or stretch work after the oracle rollout and Q_H contract is stable.

The compact thesis claim is:

ARIA-NBV adapts deep RL ideas to a constrained NBV setting: DQN supplies replayed finite-action value learning, Double DQN supplies support-aware overestimation control for candidate-table backups, and IQL supplies the offline warning that learned values must not optimize over unsupported actions.

Open risks / caveats

- Offline rollout data can be narrow if generated mostly by root-greedy policies; late-branch schedules and random-valid traces are needed for support.
- Double-Q reduces overestimation but does not solve support mismatch by itself.
- Learned values should be evaluated under equal acquisition budget against one-step greedy, oracle lookahead, and random-valid baselines on scene-level splits.

Terminology

Model

Q_H – Finite-Horizon Q Function: Expected finite-horizon return after taking a specified candidate action first. The continuation policy determines later choices. The ideal optimal value maximizes over policies using the admitted actor information; a learned scorer approximates it from a finite representation. A sampled return, the fitted training target and the predicted value are distinct. The Q function specifies neither encoder architecture nor training algorithm.

Supervision

oracle RRI – Oracle RRI: RRI measured using privileged reference geometry. The current ASE protocol prepares the historical depth cloud together with selected and hypothetical rendered observations, then measures target point-mesh error against a fixed reference crop. These labels do not make candidate depth available to the main actor.

Bibliography

- [1] M. Janner, Q. Li, and S. Levine, “Offline Reinforcement Learning as One Big Sequence Modeling Problem.” [Online]. Available: <https://arxiv.org/abs/2106.02039>
- [2] W. Kool, H. van Hoof, and M. Welling, “Stochastic Beams and Where to Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement,” in *Proceedings of the 36th International Conference on Machine Learning*, 2019. [Online]. Available: <https://arxiv.org/abs/1903.06059>
- [3] V. Mnih *et al.*, “Playing Atari with Deep Reinforcement Learning,” *CoRR*, 2013, [Online]. Available: <http://arxiv.org/abs/1312.5602>
- [4] H. van Hasselt, A. Guez, and D. Silver, “Deep Reinforcement Learning with Double Q-learning.” [Online]. Available: <https://arxiv.org/abs/1509.06461>
- [5] I. Kostrikov, A. Nair, and S. Levine, “Offline Reinforcement Learning with Implicit Q-Learning.” [Online]. Available: <https://arxiv.org/abs/2110.06169>
- [6] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, “Reinforcement Learning with Deep Energy-Based Policies,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017. [Online]. Available: <https://arxiv.org/abs/1702.08165>
- [7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms.” [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [8] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor,” in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1861–1870. [Online]. Available: <https://arxiv.org/abs/1801.01290>