RAS Techniques and Instruments

RASTI **5,** 1–8 (2026) https://doi.org/10.1093/rasti/rzag005
Advance Access publication 2026 February 04

Solving the 2-level atom non-LTE problem using soft actor-critic reinforcement learning

Brandon Panos [1]★ and Ivan Milić [2]

[1]*Institute for Data Science, University of Applied Sciences and Arts Northwestern Switzerland (FHNW), Bahnhofstrasse 6, CH-5210 Windisch, Switzerland*
[2]*Institute for Solar Physics (KIS), Georges-Köhler-Allee 401a, D-79110 Freiburg, Germany*

Accepted 2026 January 6. Received 2025 December 19; in original form 2025 July 2

ABSTRACT

We present a novel reinforcement learning (RL) approach for solving the classical 2-level atom non-LTE radiative transfer problem by framing it as a control task in which an RL agent learns a depth-dependent source function $S(\tau)$ that self-consistently satisfies the equation of statistical equilibrium (SE). The agent's policy is optimized entirely via reward-based interactions with a radiative transfer engine, without explicit knowledge of the ground truth. This method bypasses the need for constructing approximate lambda operators (Λ^*) common in accelerated iterative schemes. Additionally, it requires no extensive precomputed labelled data sets to extract a supervisory signal, and avoids backpropagating gradients through the complex RT solver itself. Finally, we show through experiment that a simple feedforward neural network trained greedily cannot solve for SE, possibly due to the moving target nature of the problem. Our Λ^*- Free method offers potential advantages for complex scenarios (e.g. atmospheres with enhanced velocity fields, multidimensional geometries, or complex microphysics) where Λ^* construction or solver differentiability is challenging. Additionally, the agent can be incentivized to find more efficient policies by manipulating the discount factor, leading to a reprioritization of immediate rewards. If demonstrated to generalize past its training data, this RL framework could serve as an alternative or accelerated formalism to achieve SE. To the best of our knowledge, this study represents the first application of reinforcement learning in solar physics that directly solves for a fundamental physical constraint.

Key words: algorithms – numerical methods – simulations – radiative transfer - machine learning - reinforcement learning.

1 INTRODUCTION

Radiative transfer (RT) calculations are critical for interpreting light from diverse astronomical sources (e.g. S. Chandrasekhar 1950; R. G. Athay 1972), including the early universe (D. Korber et al. 2023), supernovae (X. Chen et al. 2022), galaxy formations (S. S. Sethuram et al. 2023), stars, and our Sun (H. Uitenbroek 2001). Researchers infer physical variables (e.g. magnetic field strength, temperature, density, abundance, velocity) by analysing spectra from specific atomic transitions sensitive to these local conditions. The modern inference process typically involves iteratively running forward RT simulations and adjusting parameters until the calculated spectrum matches the observation. The resulting parameters are then hypothesized to represent the object's physical state. This method crucially assumes that the simulation accurately captures all necessary physics. However, rigorously modelling the radiation-matter interaction, especially under non-local thermodynamic equilibrium (non-LTE) conditions relevant for many spectral lines, is often computationally intractable. The non-LTE conditions denote any departure of the system from the state of LTE and directly imply a strong, non-linear coupling between the state of gas and radiation. Understanding this coupling is also extremely important in the context of radiative (magneto) hydrodynamical simulations, where the radiation also serves as a crucial means of energy transport (e.g. B. V. Gudiksen et al. 2011; D. Przybylski et al. 2022). As a result, researchers frequently adopt unwarranted simplifying assumptions such as plane-parallel geometry (effective 1D atmospheres), complete frequency redistribution (CRD), stationary backgrounds, low-order numerical methods, and coarse spatial grids, which all call into question the reliability of the inferred physical parameters.

The problem of RT involves solving a massive coupled set of linear equations, which in the non-LTE regime is done iteratively using techniques such as the Lambda Iteration (LI) or Accelerated Lambda Iteration (ALI) (G. B. Rybicki & D. G. Hummer 1991, 1992, 1994), Gauss-Seidel (J. Trujillo Bueno & P. Fabiani Bendicho 1995), implicit lambda (I. Milić & O. Atanacković 2014), multi-grid (O. Steiner 1991; J. Štěpán & J. Trujillo Bueno 2013), bi-conjugate gradient methods (F. Paletou & E. Anterrieu 2009), and matrix-free approaches (P. Benedusi et al. 2023). Despite algorithmic progress, accurate and complete simulations still remain computationally prohibitive.

★ E-mail: brandon.panos@fhnw.ch

© The Author(s) 2026.
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/),which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/rasti/article/doi/10.1093/rasti/rzag005/8460525 by Fachhochschule Nordwestschweiz FHNW user on 10 February 2026

For this reason, researchers have started using machine learning-based methods to accelerate RT simulations, including learning a mapping between LTE and non-LTE level populations in 3D (B. A. Chappell & T. M. D. Pereira 2022), predicting the departure coefficients of atomic level populations (A. Vicente Arévalo, A. Asensio Ramos & S. Esteban Pozuelo 2022), and replacing the computational core of simulations with physics-informed neural networks (PINNs) (M. Raissi, P. Perdikaris & G. E. Karniadakis 2019). Direct emulation of the simulation engine has resulted in impressive offline performance gains for numerical weather predictions (R. Lagerquist et al. 2021), the forward modelling of galaxy spectral energy distributions (S. S. Sethuram et al. 2023), and accurate predictions of the complete 4D hydrogen fraction evolution of the epoch of reionization (D. Korber et al. 2023). Additionally, neural fields and PINNs have been applied to the inverse problem in solar physics, specifically for spectropolarimetric inversions to infer atmospheric parameters like the magnetic field (C. J. Díaz Baso et al. 2025; R. Jarolim et al. 2025). Despite these advancements, PINNs are still brittle under generalization, do not scale well, and struggle to function as RT emulators in real-time online tasks (B. Mu et al. 2023).

Reinforcement learning (RL) represents a powerful branch of machine learning that remains underutilized within the sciences despite possessing qualities seemingly highly advantageous for simulation tasks (R. S. Sutton & A. G. Barto 2018). Many computational physics problems involve iterative schemes to reach a solution, which, when employing supervised function approximation or emulation, can result in a fast accumulation of errors. Unlike most approximation paradigms, RL optimizes policies for sequential decision-making by maximizing a cumulative long-term reward signal. This framework inherently balances exploration (sampling novel strategies) and exploitation (leveraging effective known strategies), enabling the potential discovery of non-intuitive or more efficient solutions. Additionally, the design of the reward function provides flexibility that can promote the exploration of fundamentally new solutions in the case of a reward sparse framework (usually requiring more episodes and therefore compute), while dense, structured rewards can incorporate domain knowledge which forces the policy into the realm of imitation or expert learning (M. Zare et al. 2024). RL has achieved unprecedented success within the gaming industry thanks to its ability to automate the algorithmic discovery process, resulting in superhuman performance in complex strategic games such as Chess (D. Silver et al. 2017), Go (D. Silver et al. 2016), and Dota 2 (C. Berner et al. 2019). By formulating scientific problems within an RL framework ('gamification'), researchers have achieved substantial breakthroughs, including predicting protein structures with atomic-level accuracy (J. Jumper et al. 2021), developing novel turbulence modelling strategies via multi-agent RL (G. Novati, H. L. Laroussilhe & P. Koumoutsakos 2021), and discovering faster algorithms for fundamental basic computations such as matrix multiplication (A. Fawzi et al. 2022). These examples illustrate the potential of RL in the sciences, beyond traditional game environments.

In this work, we propose a novel solution to the RT problem that dispenses with manual heuristics, such as the pre-defined construction of an approximate lambda operator in ALI schemes, and avoids constructing labelled data sets for supervision. Instead, our approach relies exclusively on a carefully designed reward function that allows an RL agent to develop an optimal update strategy (a policy), that can efficiently drive a simple non-LTE atmosphere into its steady state of statistical equilibrium (SE), purely through reward-based interaction with a physics engine. Whether this framework generalizes to other atmospheric profiles and more realistic scenarios remains a task for future studies.

2 TWO-LEVEL ATOM NON-LTE PROBLEM

For a proof of concept, we selected a simple academic, non-LTE problem: a single species of 2-level atom under the assumption of complete frequency redistribution (CRD) in a 1D plane-parallel atmosphere. The depth grid is defined on $\log_{10}(\tau_L)$ ranging from -4 to 5 consisting of 91 depth points, where τ_L is the line-integrated optical depth, as in classic two-level non-LTE line formation problem (see e.g. I. Hubeny & D. Mihalas 2014), and the continuum opacity is set to zero. For simplicity, the atmosphere is assumed to be isothermal, represented by a constant Planck function, $B(\tau_L) = B = 1$. A Gaussian (Doppler) profile $\phi(\nu)$ is used to model the absorption line and is taken to be constant with depth. Frequency and angle integrations employ Gaussian quadrature schemes. For the angular integration we use a 3-point Gauss–Legendre quadrature on [0,1], while the frequency integration is performed on a uniform 21-point grid over $[-4, 4]$ Doppler widths using trapezoidal weights normalized to unity.

The radiative transfer equation (RTE) in this case reads:

$$\mu \frac{\mathrm{d}I(\tau_L, \mu, \nu)}{\mathrm{d}\tau_L} = \phi(\nu)\left[I(\tau_L, \mu, \nu) - S(\tau_L)\right]. \tag{1}$$

RTE was solved numerically along discrete rays (μ =const) using a second-order short characteristics method (G. L. Olson & P. Kunasz 1987) to compute the specific intensity $I(\tau_L, \mu, \nu)$. Finally, to mimic a solar atmosphere, we set the boundary conditions such that the incoming radiation ($\mu < 0$) at the top of the atmosphere was given by $I_{\rm in} = 0$ and the outgoing radiation ($\mu > 0$) at the bottom of the atmosphere was provided by the Planck function $I_{\rm out} = B$. The core of the non-LTE problem lies in satisfying the SE equation (referred to as 'kinetic equilibrium' by I. Hubeny & D. Mihalas 2014), which for a 2-level atom is given by:

$$S(\tau_L) = (1 - \epsilon)\bar{J}(\tau_L) + \epsilon B(\tau_L), \tag{2}$$

where $\bar{J}$ is the so-called scattering integral, i.e. absorption profile-averaged mean intensity:

$$\bar{J}(\tau_L) = \frac{1}{2}\int \phi(\nu) \int_{-1}^{1} I(\tau_L, \mu, \nu)\mathrm{d}\mu \, \mathrm{d}\nu. \tag{3}$$

SE refers to the steady-state condition in which, at each atmospheric depth, the atomic level populations do not evolve in time because all radiative and collisional excitation and de-excitation rates exactly balance, for each level. In our case, SE reduces to a condition in which the source function is self-consistent with the radiation field it produces, i.e. the source function equals the one implied by the radiative transfer solution. The value of the photon destruction probability ϵ in equation (2) dictates whether we are in a high-scattering or strongly absorbing region of the atmosphere. By 'scattering', we refer to photons that are re-emitted via radiative de-excitation, following radiative excitation, contributing to the mean intensity $\bar{J}$. By 'absorption' we mean radiative excitation, followed by de-excitation due to inelastic collisions with thermal electrons, resulting in the destruction of the original photon. This process is coupled to the probability of producing a photon via collisional excitation, which produces positive emissivity and therefore increases the source function.

Downloaded from https://academic.oup.com/rasti/article/doi/10.1093/rasti/rzag005/8460525 by Fachhochschule Nordwestschweiz FHNW user on 10 February 2026

When $0 \leq \epsilon \ll 1$, the source function decouples from the Planck function and is determined by the radiation field and non-local contributions. However, when $\epsilon \to 1$, the source function is entirely locally determined, i.e. we recover the assumption of LTE. We set $\epsilon = 10^{-4}$ to be constant with depth. The setup is similar to very well-known problems used to benchmark the operator perturbation techniques (see e.g. D. Hummer & G. Rybicki 1971; D. Mihalas 1978; G. L. Olson & P. Kunasz 1987). Note that the problem is linear in the source function and, after suitable discretization, can be solved directly. However, it serves as a pedagogical introduction to multilevel problems which can be highly non-linear (D. Mihalas 1978; G. B. Rybicki & D. G. Hummer 1991).

A classical, simple solution of these problems is to iterate equations (3) and (2) until convergence. This is known as 'Lambda iteration'. For small values of ϵ, the procedure is extremely slow and converges linearly, which makes defining a stopping criterion difficult (a thorough discussion is presented in the textbook of D. Mihalas 1978). A widely accepted approach is to accelerate this procedure by the operator perturbation techniques, which are, in a way, equivalent to preconditioning (e.g. D. Mihalas 1978; G. L. Olson & P. Kunasz 1987; F. Paletou & E. Anterrieu 2009). See also H. Socas-Navarro & J. Trujillo Bueno (1997) for an explicit comparison between pre-conditioning and linearization approaches. The core of all these methods is the process of solving RTE for all necessary directions and frequencies, and then updating the source function (equation 2). Regardless of the method employed, we refer to this process as one Λ iteration.

3 REINFORCEMENT LEARNING FORMULATION

We reformulate the RT problem within an RL framework, which models the interaction between an agent and an environment. In this framework, the agent selects actions to influence the environment's state transitions, aiming to learn a policy $\pi(a|s)$ that maximizes cumulative future rewards. For the RT problem, the environment represents the physical system, with its state characterized by the current depth-dependent source function $S(\tau_L)$. The agent is implemented using the soft actor-critic (SAC) algorithm. Unlike traditional methods that modify $S(\tau_L)$ at discrete points, our agent's action is the selection of a low-dimensional parameter vector $\mathbf{a} \in [-1, 1]^4$. This vector is scaled to define four physical parameters, $\mathbf{p} =$ (floor, amplitude, centre, width), which collectively determine the source function across the entire depth grid via a smooth, generally monotonic sigmoid profile:

$$S(\tau_L; \mathbf{p}) = \text{floor} + \text{amplitude} \times \text{sigmoid}\left(\frac{\log_{10}(\tau_L) - \text{centre}}{\text{width}}\right). \tag{4}$$

The agent, as seen in Fig. 1, observes the resulting source function $S(\tau_L; \mathbf{p})$ generated by its chosen parameters. The core learning mechanism is the reward function, designed to drive the agent towards the SE solution. At each step t, given the agent's proposed parameters $\mathbf{p}_t$ yielding $S_t = S(\tau_L; \mathbf{p}_t)$, the environment calculates the corresponding mean intensity $\bar{J}(S_t)$ and determines the source function implied by SE:

$$S_{\text{implied},t} = (1 - \epsilon)\bar{J}(S_t) + \epsilon B. \tag{5}$$

One such step is equivalent to one Λ iteration in conventional approaches. The reward R_t is the negative mean squared error

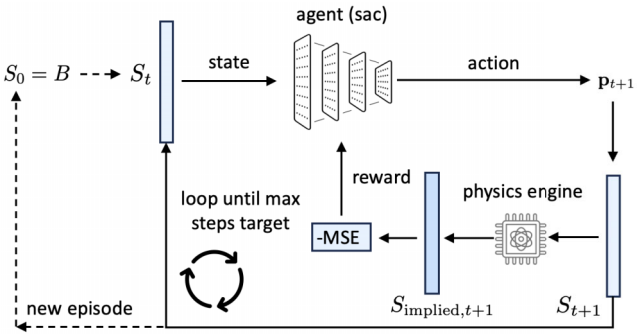

Figure 1. Diagram of the training loop: In a clockwise fashion; the source function is initiated to the Planck function B, which is then sent to the agent as a state. The agent's policy decides on an action that generates four parameters $\mathbf{p}$ that are used to construct a smooth, well-behaved source function across the entire depth scale. The agent's predicted solution is passed to the physics engine, which generates a conditioned 'implied' source function that would satisfy SE. The residual of the agent and implied source is used as a reward signal to instruct the agent's policy. The inner loop iterates until either a max step criterion is reached or the agent obtains the target, defining a single episode. Once the inner loop terminates, the source function is once again initiated to the Planck function, and the agent can try and refine its policy.

(MSE) between the parametrized and implied source functions:

$$R_t = -\frac{1}{N_D}\sum_{i=1}^{N_D}[S_t(\tau_{c,i}) - S_{\text{implied},t}(\tau_{c,i})]^2. \tag{6}$$

Maximizing this reward incentivizes the agent to find parameters $\mathbf{p}$ such that $S(\tau_L; \mathbf{p})$ globally satisfies the SE condition. An episode involves the agent iteratively proposing parameters and receiving rewards, terminating either after a maximum number of iterations or when the MSE between the agent's current S_t and a pre-computed, converged ALI solution falls below a threshold. The ALI solution serves solely as a termination criterion and does not inform the reward signal during training. For problems without reference solutions, the stopping criterion could instead be defined through residual thresholds. Note that all rewards are ≤ 0, meaning the agent tries to obtain a net reward that approaches zero over the maximum 50 permitted time-steps of exploration (inner loop counter). The problem of finding SE is what is termed a moving target problem within the domain of RL (V. Mnih et al. 2015), because the implied S used to calculate the residual depends on the actors actions. Moving target problems are notoriously difficult for direct gradient methods to solve, because the loss landscape continuously reshapes itself around the latest guess, resulting in noisy gradients (T. P. Lillicrap et al. 2016). A useful analogy can be found in dynamic traffic routing, where a navigation system might identify an empty road as the optimal route, but once many drivers take the recommendation, the road becomes congested, shifting the optimal route elsewhere.

Fig. 2 illustrates the flexibility of the parametrized source function by visualizing its solution space. The background heatmap represents the probability density, computed from 50 000 uniformly sampled parameter sets and displayed on a logarithmic scale. This reachable space is compared against the initial guess (dashed black line) and the desired target solution obtained from ALI (dashed blue line). Ten randomly selected examples of the parametrized function (grey solid lines) illustrate some of the characteristic shapes generated within the defined parameter boundaries, showing the parametrization's ability (or

Downloaded from https://academic.oup.com/rasti/article/doi/10.1093/rasti/rzag005/8460525 by Fachhochschule Nordwestschweiz FHNW user on 10 February 2026

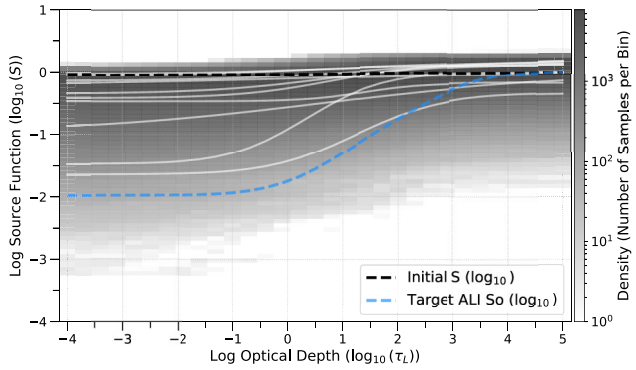

Figure 2. Reachable solution space of the parametrized source function. The density heatmap (log scale) shows the frequency of $\log_{10}(S)$ values versus $\log_{10}(\tau_L)$ based on 50 000 random parameter samples. Overlays show the initial guess (black), target ALI solution (blue dashed), and example random profiles (light grey).

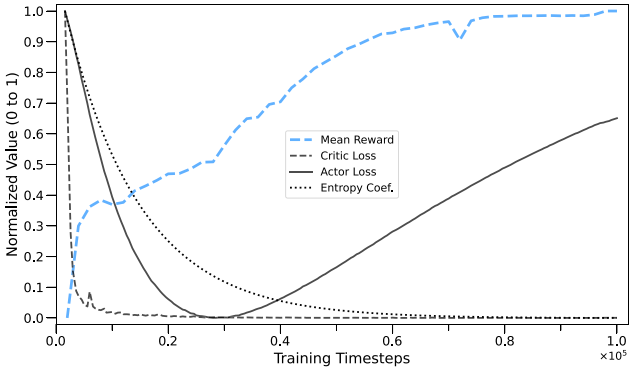

Figure 3. Training performance of the SAC agent. The figure shows the mean normalized reward, critic loss, actor loss, and entropy coefficient α, as a function of training steps. The steadily increasing reward demonstrates successful learning, while the critic loss decreases, indicating convergence of the value function estimate. The decreasing entropy coefficient signifies a shift from exploration towards policy exploitation.

limitations) in approximating the target profile across different optical depths. Due to its limited flexibility, there will inevitably be a small residual between the target and the best solution, as verified by a Levenberg–Marquardt least-squares fit. This implies that our agent is exploring the solution space to effectively find the infimum of the residual.

To assess the accuracy achievable within the analytic profile of equation (4), we directly fit this parametrization to the converged ALI source function using a non-linear least-squares procedure. The best fit yields a mean squared error of 4.7×10^{-5} and a relative L_2 error $E_{\rm rel} = \|S_{\rm fit} - S_{\rm ALI}\|_2 / \|S_{\rm ALI}\|_2 = 1.56 \times 10^{-2}$, representing the best attainable accuracy within this parametrized manifold. The trained SAC policy achieves a residual error indistinguishable from this lower bound, indicating that the agent effectively exploits the full expressive power of the chosen parametrization while just falling short of exactly reproducing the true non-LTE solution.

4 THE SOFT ACTOR-CRITIC ALGORITHM

The agent depicted in Fig. 1 represents a state-of-the-art, off-policy, stochastic, model-free RL algorithm called soft actor-critic (SAC), which uniquely addresses problems with continuous action spaces (T. Haarnoja et al. 2018b). It is composed of an ensemble of backpropagation function approximators working in tandem to extract an efficient policy. SAC is based on the maximum entropy actor-critic framework, which is both sample efficient (due to an experience replay buffer) and insensitive to hyperparameters. The actor maps the observed state (the S-profile) to a probability distribution over actions (the parameters $\mathbf{p}$), while the critic (value function) estimates the expected future cumulative reward (Q-value) for taking a given action in a given state. In practice, two main and two target critics are used to stabilize training. SAC optimizes a policy π to maximize a trade-off between the expected cumulative reward and the policy's entropy H:

$$J(\pi) = \sum_{t=0}^{T} E_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[R(\mathbf{s}_t, \mathbf{a}_t) + \alpha H(\pi(\cdot|\mathbf{s}_t)) \right], \tag{7}$$

where $\mathbf{s}_t$ is the state (S-profile), $\mathbf{a}_t$ is the action (parameters), ρ_π is the state-action distribution induced by policy π, R is the reward from equation (6), and α is a temperature parameter balancing reward maximization and entropy maximization. Effectively, equation (7) instructs the agent to find the optimal policy while behaving as randomly as possible. If the entropy is maximized then each of the four variables will be sampled from a flat distribution, implying zero strategic decision making and consequently a low net reward. On the other hand, maximizing the reward directly could result in suboptimal behaviour due to insufficient exploration. Finding a balance between these two tensions (which the algorithm regulates internally) encourages a healthy exploration and prevents the policy from collapsing to a deterministic, potentially suboptimal solution, which as we will see, might be an important attribute that allows SAC to solve this particular moving target problem.

The SAC agent was trained using Stable Baselines3 (A. Hill et al. 2018) for a specified number of total time-steps. The training progress was monitored, and the best-performing model (based on evaluation episodes) was saved. The learning dynamics of the actor-critic system can be seen in Fig. 3. Here, the steadily increasing mean reward indicates successful learning, as the SAC algorithm extracts an efficient policy that finds a set of parameters $\{\mathbf{p}\}$ which populate and cluster around the target SE solution, resulting in high relative accumulative rewards. The critic networks (two main and two target) learn to approximate the Q-function. The decrease in critic loss (representing the error between the neural network's predicted return and the actual returns estimated via the Bellman update) signifies that the critic's predictions of future cumulative rewards are becoming more accurate based on the observed transitions and rewards. The solid black line represents the loss associated with the actor network (the policy), minimizing this loss is equivalent to maximizing the expected Q-value estimated by the critic. Its initial decrease corresponds to the policy rapidly improving, while the subsequent increase is not uncommon and does not necessarily indicate poor performance, as evidenced by the steady increase in mean reward. It is possible for the policy loss to both increase while improving the estimate of the source function. This effect is consistent with the mechanics of the SAC algorithm (T. Haarnoja et al. 2018a, b), and simply reflects the changing scale of the actor's objective function as the improved policy guides the trajectory through increasingly rewarding regions of the state space. Finally, the black dotted line shows the entropy coefficient,

Downloaded from https://academic.oup.com/rasti/article/doi/10.1093/rasti/rzag005/8460525 by Fachhochschule Nordwestschweiz FHNW user on 10 February 2026

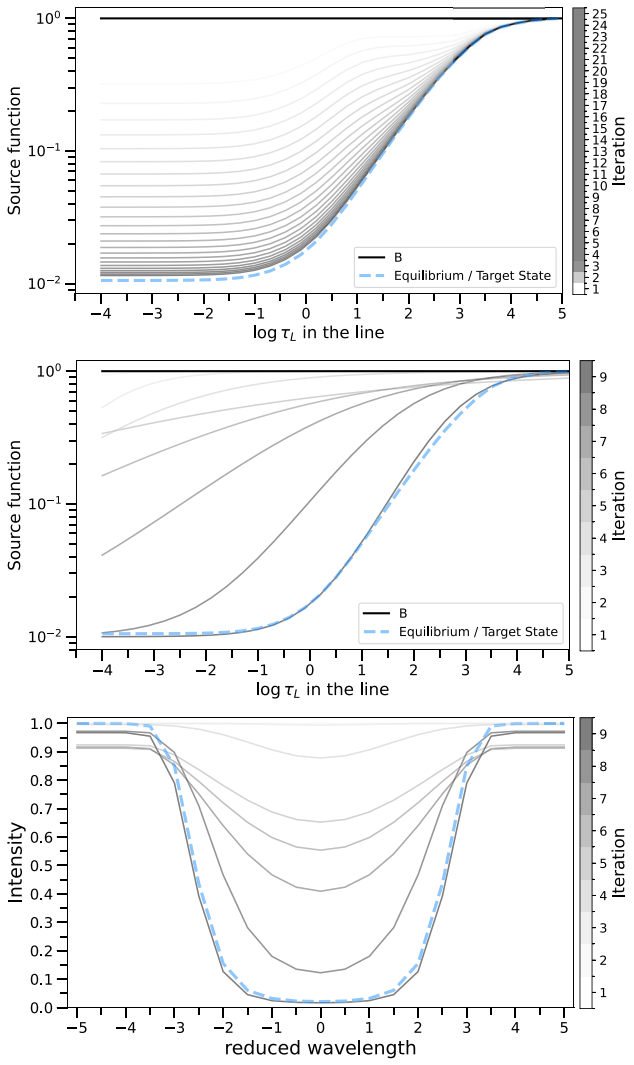

Figure 4. Upper panel: ALI method converges to SE (dashed blue line) after multiple iterations. Middle panel: The SAC policy drives the simple non-LTE simulation into SE with fewer iterations. Decreasing the discount factor promotes policies that converge faster. Lower panel: Evolution of line-of-sight ($\mu = 1$) observed intensity as a function of the agents policy.

α, which weights the importance of the policy's entropy in the objective function. The trend shows that the agent starts with an aggressive exploration (high entropy) strategy, and then over the course of training, slowly focuses more on exploiting the known good actions (lower entropy). In the context of our parametrization, high entropy implies a broad Gaussian distribution over the source function parameters **p** (wide exploration), while low entropy indicates the agent has become confident in a specific spectral profile (narrowing the distribution).

5 RESULTS – COMPARING SAC POLICY PERFORMANCE

We evaluate the performance of the trained SAC agent by comparing its convergence behaviour to the ALI scheme. The upper panel of Fig. 4 depicts the standard ALI convergence, starting from an initial guess where the source function equals the Planck function ($S = B$, solid black line). ALI iteratively refines the source function profile across the optical depth scale. The greyscale lines represent successive iterations, gradually approaching the target SE state (dashed blue line).

The middle panel depicts the same process, but in this case, the source function is driven by the trained SAC agent's optimal policy, corresponding to a single evaluation episode within the inner loop of Fig. 1. The agent begins with an initial state ($S = B$) and, at each step, selects an action (a set of four parameters **p**) based on its learned policy $\pi^*(a|s)$. This action generates a parametrized source function according to equation (4), represented by the greyscale lines. The plot demonstrates that the SAC policy rapidly drives the system towards the target SE solution (dashed blue line), achieving the same tolerance as the ALI scheme with fewer steps, representing a significant reduction in the number of solver interactions.

Finally, the lower panel shows the corresponding evolution of the emergent line-of-sight intensity profile, $I(\mu = 1)$, as a function of wavelength (normalized to Doppler width). As the SAC policy adjusts the source function parameters step-by-step (middle panel), the resulting spectrum evolves from the initial flat continuum ($I = B = 1$) towards the final absorption line expected at convergence (dashed blue line), providing a visual confirmation of the method's legitimacy.

Furthermore, the agent can be incentivized to find more optimal policies by reducing the discount factor $\gamma < 0.99$, which serves as a pre-factor for the trajectory's reward chain $\gamma^t r_t$, leading to a prioritization of immediate returns. It is unclear how the choice of discount factor will impact the agent's ability to generalize and find optimal solutions. We hypothesize that the agent's optimal solution would be compromised with lower γ as exploration will be dampened.

It is important to note that this comparison focuses on the number of iterations required after training. The RL agent requires a separate, potentially computationally intensive, training phase. The performance only has meaningful acceleration potential if the given policy generalizes to other unseen atmospheric configurations.

Although this work is focused on a proof of concept, we provide some computational details. All computations were performed on a MacBook Pro (Apple M2 Pro chip, 32GB RAM). Training the SAC agent (100 000 time-steps) took approximately 4 min. Once trained, the policy acts as an extremely fast approximate solver. A single forward pass of the policy network takes ~ 1 ms, and with an average of eight steps to reach equilibrium, the total time to solution is ~ 0.01 s. In comparison, the pedagogical Python-based ALI scheme used in this environment requires 1 s per iteration and ~ 25 s to fully converge. While the ALI benchmark reflects an unoptimized implementation, the orders-of-magnitude difference highlights the potential of the RL approach to bypass expensive formal solution loops during online inference (keeping in mind the high upfront pre-training cost).

5.1 Policy stability and comparison with direct optimization

Extracting RL policies for an optimization task often requires a massive number of agent-environment interactions and is therefore computationally expensive. It then becomes important to demonstrate superior performance over simpler, more direct methods. To test the benefits of a sophisticated, long time horizon, delayed-reward planning algorithm over an essentially

Downloaded from https://academic.oup.com/rasti/article/doi/10.1093/rasti/rzag005/8460525 by Fachhochschule Nordwestschweiz FHNW user on 10 February 2026

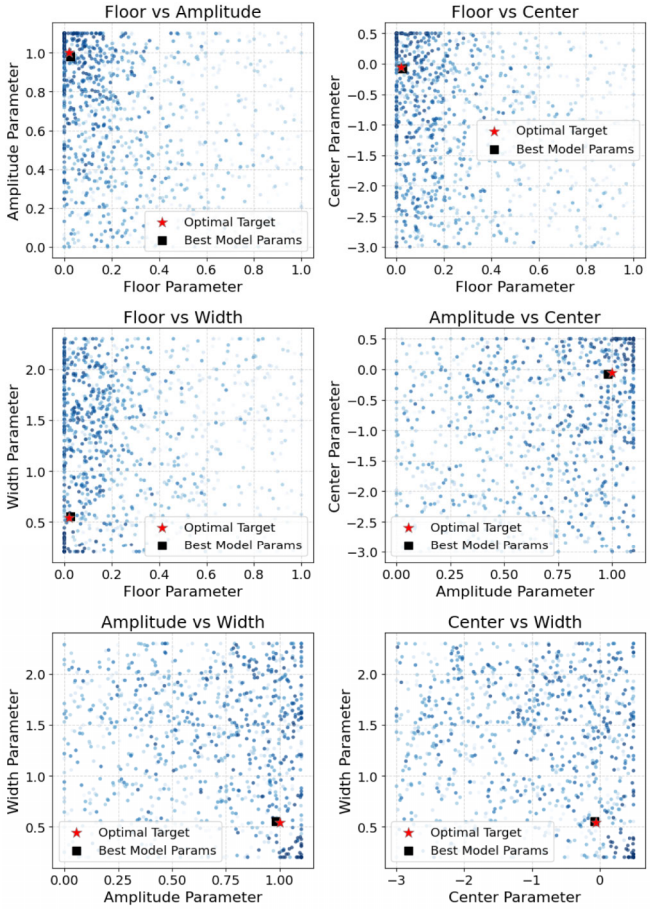

Figure 5. Evolution of parameters during SAC training: The six phase planes show how the source function parameters change over time, with darker blue dots indicating later agent actions. The target solution of SE in the parameter space is indicated by a red star, while the policy's optimal solution is indicated by a black square. The plot shows how the parameters migrate during training towards the target setting.

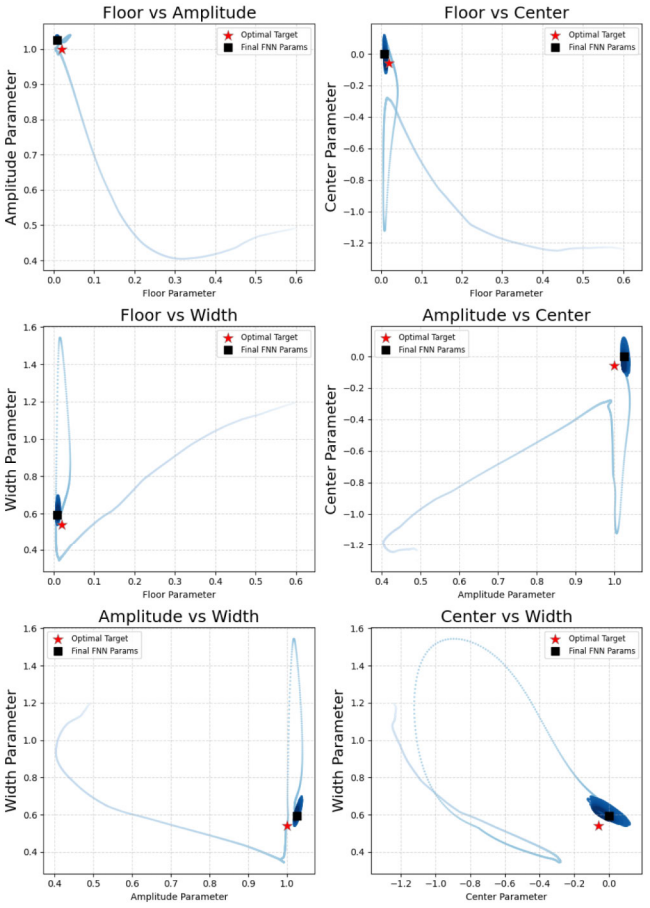

Figure 6. Evolution of parameters using direct optimization: Same as Fig. 5. In this case, the simple FNN directly seeks the solution; however, due to the moving target nature of the problem, the model gets trapped in a suboptimal region of the parameter space, a behaviour that is consistent regardless of model complexity or training time.

greedy policy, we replaced the SAC agent in Fig, 1 with a fully connected feed forward network (FNN), while keeping the remaining logic of the loop consistent. The FNN was trained using the same residual-based feedback loop. The performance and stability of these two approaches differ significantly, as illustrated by comparing the evolution of the source function parameters during training.

Fig. 5 visualizes the trajectory of the four source function parameters (floor, amplitude, centre, width) explored by the SAC agent during training, displayed across the six pairwise parameter phase planes. Each blue dot represents the parameter set generated by an agent action, with the colour darkening over time to indicate later steps in the training process.

The agent initially explores a wide region of the parameter space (lighter dots) before progressively clustering around the target SE solution represented by the red star in each panel. This behavioural tendency from exploration into exploitation is a signature characteristic of SAC and is consistent with the decreasing entropy curve in Fig. 3. The plot clearly shows the agent effectively navigating the parameter space, with the parameter samples migrating and clustering around the policy's optimum (black squares), which closely approximates the true target (red star). In each case the target and model solution are on top of one another or overlapping. The minor difference between the policy's optimum and target is likely due to the aforementioned limitations of the sigmoid function, since its lack of flexibility results in at best the discovery of an infimum. This fixed best case error is also likely responsible for pushing the model's parameter search to its boundaries, resulting in a darkening of the plot edges, as SAC searches for a better configuration that does not exist.

In contrast, Fig. 6 illustrates the parameter evolution when using the direct FNN optimization approach. Unlike the stochastic nature of SAC, the FNN greedy optimization scheme performed a controlled, rapid, and direct parameter trajectory toward the optimal settings, in far fewer learning cycles, however, the model never obtains the target, but oscillates in a neighbourhood close to the true solution. Note that the optimal setting is not degenerate, and that all six parameters have to therefore be aligned with the stars for an optimal solution.

We hypothesize that the moving target nature of the problem, where the target for the model depends on its output, results in a continuous undulation and restructuring of the loss landscape, something that proves difficult for standard backpropagation FNN to learn, regardless of the model's complexity or training time.

Downloaded from https://academic.oup.com/rasti/article/doi/10.1093/rasti/rzag005/8460525 by Fachhochschule Nordwestschweiz FHNW user on 10 February 2026

SAC on the other hand decouples policy improvement from value estimation, where the critic bootstraps value estimates via the Bellman equation and smooths them with soft targets using Polyak averaging, generating a low-variance learning signal, even within a moving target scenario. The actor can then separately optimize against these stable targets. Additionally, the entropy-driven search of SAC allows it to escape local optima that trap the FNN.

6 CONCLUSION AND OUTLOOK

In this work, we have presented a novel reinforcement learning solution using SAC for the classical 2-level atom non-LTE radiative transfer problem. By framing SE as a control task, an agent learns an optimal policy to parametrize the source function $S(\tau_L)$ based solely on reward signals derived from the SE residual, obtained via interaction with a formal solver. This method does not require direct knowledge of the ground truth, nor does it require a labelled data set or the backpropagation of gradients through the RT solver itself. Our results demonstrate that the agent successfully learns a policy satisfying the SE constraint for a simplified solar atmospheric model. Once trained, this policy drives the system to equilibrium in significantly fewer iterations than a standard ALI scheme for the tested configuration, suggesting a potential for computational acceleration. An important result is SAC's ability to handle the moving-target nature of the SE problem, where classical gradient-based optimization fails. This demonstrates the advantage of entropy regularized RL in problems where the fixed point depends on the agent's own actions. The SAC method converged reliably where a direct feedforward network optimization failed due to instabilities in the gradient signal. To the best of our knowledge, this represents the first application of RL to directly enforce SE in solar physics or radiative transfer. Our method offers an 'Approximate-Lambda-Operator-Free' path to equilibrium, while the formal solve to obtain the mean intensity $\bar{J}$ (the result of the true Lambda operator) is necessary to structure the reward function, the technique avoids the explicit construction and inversion of the approximate operators (Λ^*) central to ALI. This feature is potentially advantageous for complex scenarios where formulating effective operators is challenging.

While promising, this work serves as a proof of concept and several natural extensions exist. Although the trained optimal SAC policy is highly efficient, the acceleration gains are only meaningful if the policy can be shown to generalize to different environments, without the need for expensive pre-training. Furthermore, the system should be submerged into a more realistic complex environment such as that provided by the Lightweaver code C. M. Osborne & I. Milić (2021). The effect the discount factor has on the policy as well as the density of the reward signal provided to the actor should be numerically tested. We hypothesize that reward sparse configurations, where the inner loop's reward is only dispensed to the agent at modular intervals in Fig. 1, would result in more optimal convergence at the price of increased compute but also decreased generalization. On the other hand, a reward-rich scenario could promote generalization, as the agent's policy is more tightly constrained and tethered to the behaviour of the underlying simulation. In the case where the system is out of equilibrium, a general imitation learning approach can be deployed (M. Zare et al. 2024).

Although in this work we adopt a simple analytic parametrization of the source function (equation 4), more realistic radiative-transfer problems such as multilevel atoms or PRD, where S becomes frequency and angle-dependent, will require more flexible state and action representations. A natural generalization is for the agent to operate on the departure coefficients (or other population-based quantities) rather than directly on S, which avoids prescribing an analytic form and remains physically meaningful even for complex atoms. Another promising direction is to learn a low-dimensional representation of the source function (e.g. via PCA or an autoencoder), allowing the agent to act in this compact space. These approaches would enable the method to scale to more complex radiative-transfer settings without relying on hand-crafted analytic parametrizations. A related extension is to use the SAC-derived source function as an initial guess for a short sequence of Λ-iterations. In this hybrid strategy, the RL agent rapidly drives the system into a neighbourhood of the SE fixed point within the restricted analytic profile, and a few classical iterations then remove the small parametric bias.

Finally, we plan to extend the RL framework to related optimization problems in astrophysics that are easily 'gamified', such as spectropolarimetric inversions with discrete node adjustments. It is our belief that reinforcement learning provides an exciting and flexible new paradigm for tackling complex, physics-constrained problems within solar physics, enabling agents to learn directly from simulations and uncover efficient, non-intuitive solutions.

ACKNOWLEDGEMENTS

The author thanks the University of Applied Sciences and Arts Northwestern Switzerland for providing the resources and support, particularly through conference opportunities, that greatly contributed to the development of this work. We would also like to thank Dr Reza Kakooee for his expert knowledge of reinforcement learning and for early fruitful discussions. The machine learning aspects of the project were developed using PYTORCH (A. Paszke et al. 2019) and the open-source STABLE BASELINES3 (A. Hill et al. 2018) reinforcement learning library, while the forward transfer model made use of a publicly available two-level Non-LTE code (I. Milic 2023).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY

The source code developed for this research, including the reinforcement learning environment and the implementation of the Soft Actor-Critic agent, is publicly available on GitHub at https://github.com/brandonlpanos/LightLogic.git

REFERENCES

Athay R. G., 1972, Radiation Transport in Spectral Lines. Springer, Dordrecht
Benedusi P., Riva S., Zulian P., Štěpán J., Belluzzi L., Krause R., 2023, J. Comput. Phys., 479, 112013
Berner C. et al., 2019, preprint (arXiv:abs/1912.06680)
Chandrasekhar S., 1950, Radiative Transfer. Oxford University Press, Oxford
Chappell B. A., Pereira T. M. D., 2022, record ascl: 2202.024
Chen X., Jeffery D. J., Zhong M., McClenny L., Braga-Neto U., Wang L., 2022, preprint (arXiv:2211.05219)

Downloaded from https://academic.oup.com/rasti/article/doi/10.1093/rasti/rzag005/8460525 by Fachhochschule Nordwestschweiz FHNW user on 10 February 2026

Díaz Baso C. J., Asensio Ramos A., de la Cruz Rodríguez J., da Silva Santos J. M., Rouppe van der Voort L., 2025, A&A, 693, A170
Fawzi A. et al., 2022, Nature, 610, 47
Gudiksen B. V., Carlsson M., Hansteen V. H., Hayek W., Leenaarts J., Martínez-Sykora J., 2011, A&A, 531, A154
Haarnoja T. et al., 2018a, preprint (arXiv:1812.05905)
Haarnoja T., Zhou A., Abbeel P., Levine S., 2018b, Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, Vol. 80, PMLR, ICML, p. 1861
Hill A. et al., 2018, Stable Baselines. Available at: https://github.com/hill-a/stable-baselines
Hubeny I., Mihalas D., 2014, Theory of Stellar Atmospheres: An Introduction to Astrophysical Non-equilibrium Quantitative Spectroscopic Analysis. Princeton Univ. Press, Princeton
Hummer D., Rybicki G., 1971, MNRAS, 152, 1
Jarolim R., Molnar M. E., Tremblay B., Centeno R., Rempel M., 2025, ApJL, 985, L7
Jumper J. et al., 2021, Nature, 596, 583
Korber D., Bianco M., Tolley E., Kneib J.-P., 2023, MNRAS, 521, 902
Lagerquist R., Turner D., Ebert-Uphoff I., Stewart J., Hagerty V., 2021, J. Atmos. Ocean. Technol., 38, 1673
Lillicrap T. P., Hunt J. J., Pritzel A., Heess N., Erez T., Tassa Y., Silver D., Wierstra D., 2016, 4th International Conference on Learning Representations (ICLR 2016)
Mihalas D., 1978, Stellar Atmospheres. W.H. Freeman, San Francisco
Milic I., 2023, 2lvl_nlte: Pedagogical non-LTE radiative transfer with a two-level atom. Available at: https://github.com/ivanzmilic/2lvl_nlte
Milić I., Atanacković O., 2014, Adv. Space Res., 54, 1297
Mnih V. et al., 2015, Nature, 518, 529
Mu B., Chen L., Yuan S., Qin B., 2023, Front. Earth Sci., 11, 1149566
Novati G., de Laroussilhe H. L., Koumoutsakos P., 2021, Nature Mach. Intell., 3, 87
Olson G. L., Kunasz P., 1987, J. Quant. Spectrosc. Radiat. Transfer, 38, 325
Osborne C. M., Milić I., 2021, ApJ, 917, 14
Paletou F., Anterrieu E., 2009, A&A, 507, 1815
Paszke A. et al., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché Buc F., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems 32. Vol. 32, Curran Associates, Inc., Red Hook, NY, p. 8024. Available at: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
Przybylski D., Cameron R., Solanki S., Rempel M., Leenaarts J., Anusha L., Witzke V., Shapiro A., 2022, A&A, 664, A91
Raissi M., Perdikaris P., Karniadakis G. E., 2019, J. Comput. Phys., 378, 686
Rybicki G. B., Hummer D. G., 1991, A&A, 245, 171
Rybicki G. B., Hummer D. G., 1992, A&A, 262, 209
Rybicki G. B., Hummer D. G., 1994, A&A, 290, 553
Sethuram S. S., Cochrane R. K., Hayward C. C., Acquaviva V., Villaescusa-Navarro F., Popping G., Wise J. H., 2023, MNRAS, 526, 4520
Silver D. et al., 2016, Nature, 529, 484
Silver D. et al., 2017, Nature, 550, 354
Socas-Navarro H., Trujillo Bueno J., 1997, ApJ, 490, 383
Steiner O., 1991, A&A, 242, 290
Štěpán J., Trujillo Bueno J., 2013, A&A, 557, A143
Sutton R. S., Barto A. G., 2018, Reinforcement Learning: An Introduction, 2 edn. MIT Press, Cambridge, MA. Available at: http://incompleteideas.net/book/the-book-2nd.html
Trujillo Bueno J., Fabiani Bendicho P., 1995, ApJ, 455, 646
Uitenbroek H., 2001, ApJ, 557, 389
Vicente Arévalo A., Asensio Ramos A., Esteban Pozuelo S., 2022, ApJ, 928, 101
Zare M., Kebria P. M., Khosravi A., Nahavandi S., 2024, IEEE Trans. Cybernet., 54, 7173

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/rasti/article/doi/10.1093/rasti/rzag005/8460525 by Fachhochschule Nordwestschweiz FHNW user on 10 February 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.