Safe Artificial Machine Intelligence Through Localization

D. Sliskovic

Abstract

As autonomous agents grow more powerful and versatile, ensuring they remain both effective and aligned with human intent has become increasingly challenging. Traditional approaches often rely heavily on extensive human oversight or computationally intensive methods for uncertainty estimation, neither of which scale well. To address challenges in alignment research, we propose a novel approach inspired by the phenomenon of Anderson localization in condensed matter physics, where random disorder in a medium causes wavefunctions to become exponentially localized and inhibits transport. Our novel alignment approach applies this principle to restrict AI agents ability to reason about unsafe behaviors by strategically injecting noise into their internal world models, which effectively "localize" its behavior, thus induce confinement in safe policy-space, analogous to the way disorder induces spatial confinement in Anderson localization.

1 Introduction

Ensuring that advanced artificial intelligence systems act in accordance with human values is a central challenge in AI safety and alignment research. Modern deep learning systems, while powerful, can exhibit unintended behavior: they may optimize proxy objectives in ways that diverge from true human intent, behave unpredictably under distributional shift, or exploit weaknesses in supervision frameworks. These failure modes pose serious risks as AI systems gain greater autonomy and capability.

A major stream of research has focused on reward modeling and Reinforcement Learning from Human Feedback (RLHF) [7] . These approaches aim to train agents via learned reward functions derived from human preferences. While effective in constrained settings, they are susceptible to reward hacking, misgeneralization, and misinterpretation of human intent. Policies trained via RLHF often perform well on the training distribution but degrade or behave undesirably in novel situations.

To address these concerns, researchers have explored scalable oversight strategies such as Iterated Distillation and Amplification (IDA) [6] and Recursive Reward Modeling [13]. These frameworks aim to amplify human judgment or recursively structure feedback across hierarchies of tasks. However, they introduce additional complexity, are difficult to scale reliably, and critically depend on the assumption that the supervising agents are themselves aligned and error-free, a recursive problem in its own right.

In parallel, the safety of model-based reinforcement learning has been advanced through techniques such as MOPO [17] and MBPO [10], which estimate epistemic uncertainty using ensembles to penalize unreliable predictions. Although these methods improve robustness, they are computationally expensive and poorly suited for real-time applications or high-dimensional environments due to the cost of maintaining and sampling multiple dynamics models.

2 Method: Safe Artificial Machine Intelligence Through Localization

These challenges underscore the need for alignment methods that simultaneously ensure safety, generalization, and efficiency. We propose an approach inspired by Anderson localization [1], a concept of condensed matter physics, to address these issues. The core idea is to constrain the agent's reasoning and planning processes in goal space within a world model. The use of a world model as a simulated environment is inspired by Schmidhuber's and Ha's early work [9], where agents learn by predicting and interacting with an internal model of the environment. Here, we extend that idea by deliberately obfuscating access to harmful concepts through noise within the simulation. This is realized through a combination of:

- **Partially Obfuscating Encoders** – degrade latent representations of misaligned states via value-dependent noise injection, limiting unsafe reasoning.
- **Joint Embedding Alignment Scoring** – compare goal embeddings with textual human values using cosine similarity in a shared embedding space.
- **Uncertainty-aware Predictive World Modeling** – distill alignment knowledge into a world model during training as encoded noise, using a single-model aleatoric predictor to capture uncertainty efficiently without ensembles.
- **Meta-controller** – inspired by Hierarchical Reinforcement Learning [4], we predict high-level goals instead of low-level continuous actions to make planning more tractable and enables more abstract alignment scoring.
- **Safe Planning via SPUCT** – modify PUCT [8] search to prune unsafe goals based on aleatoric uncertainty.
- **Low-level Control via MPC-based Subpolicies** – translate high-level goals into smooth, interpretable action sequences in continuous control tasks.

2.1 Partially Obfuscating Encoder

In our framework, the world model is trained in the style of Joint Embedding Predictive Architectures (JEPA) [2]. Input states are passed through a partially obfuscating encoder that produces high-level latent representations. For benign inputs, the encoder yields structured representations. For states which are not aligned with human values, it returns stochastic embeddings from a standard Gaussian distribution, effectively injecting noise and preventing explicit reasoning about undesirable behaviors, which is described by:

$$\tilde{z} = \sqrt{\text{align\_score}(s, v)}\,\text{enc}(s) + \sqrt{1 - \text{align\_score}(s, v)}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

The Alignment Score quantifies the degree to which a given state or goal aligns with a set of predefined human values. While existing approaches like using AI for overseeing [3] could be used here, they introduce an additional challenge: ensuring that the overseer itself remains well-aligned. In this work, we propose a efficient and interpretable method that leverages pre-encoded representations. Specifically, we compare the encoded form of a given state to the encoded representations of the textual descriptions of our values, which should reside in a joint embedding space. Our method offers greater extensibility compared to approaches that require training a new reward model for each value set. Instead, new values can be incorporated simply by embedding their textual representations, enabling rapid adaptation without retraining. Formally, we define the Alignment Score as:

$$\text{AlignmentScore}\left(S, (V_j)_{j=1}^{k}\right) = \sigma\left(a\left(\frac{1}{k}\sum_{j=1}^{k}\left(\frac{\text{cos\_sim}\left(\text{enc}(S), \text{enc}(V_j)\right) + 1}{2}\right) - threshold\right)\right) \tag{2}$$

2.2 Aleatoric World Model

This alignment knowledge is distilled into our world model by obfuscating training targets through the sampled noise. This enables more efficient real-time planning, as it eliminates the need to repeatedly

query a costly AI Evaluator to verify model alignment at each decision step. Our aleatoric world model predictor can be implemented as an RNN that outputs the mean and variance of predicted future states, with the obfuscated predictions converging toward a zero mean and unit variance. For training we adapted the negative log-likelihood loss to the Joint Embedding Predictive Architecture setting and reinforcement learning by applying it to predicted distributions over future latent representations. This allows the model to capture uncertainty in latent dynamics. For our loss we assume diagonal covariance.

$$\mathcal{L}(z_t, a_t, s_{t+1}) = \frac{1}{2} \sum_{i=1}^{d} \left[\log(2\pi\sigma_i^2(z_t, a_t)) + \frac{(\text{enc}(s_{t+1})_i - \mu_i(z_t, a_t))^2}{\sigma_i^2(z_t, a_t)} \right] \tag{3}$$

Compared to ensemble-based approaches such as Model-Based Offline Policy Optimization, this architecture offers a more computationally efficient means of capturing uncertainty since a single model is sufficient.

2.2.1 Dual Hierarchical Goal Selection and Safe Planning via SPUCT

Our agent is build on top of that world model. The agent behavior starts with setting goals. The goal selection approach is inspired by hierarchical reinforcement learning [4]. A meta-controller selects high-level goals from encoded states, which act as targets for the lower-level subpolicy. Goal selection can occur via a dual-process framework inspired by psychologist Daniel Kahneman [12]: under low uncertainty, goals are chosen via fast, intuitive inference (akin to System 1 cognition) , under high uncertainty, a more deliberate planning process (System 2) is triggered, as described in:

$$g_t = \begin{cases} \arg\max_g N(s, g), & \text{if } \exp(-\beta \cdot \text{unc}(s)) < \tau \\ \arg\max_g \pi(g \mid s), & \text{otherwise} \end{cases} \tag{4}$$

The latter is modeled after MuZero-style planning [14], where our meta-controller searches for goals most frequently visited in simulated rollouts.

We extend this planning mechanism using PUCT formular [8], with key modifications: the exploration term in the objective is scaled by aleatoric uncertainty, encouraging the agent to avoid expanding branches associated with high risk, and instead of actions our agents selects high level goals, which is formalized as Safe PUCT (SPUCT) such:

$$g^* = \arg \max_{g=\phi(s)} \left[Q(s, g) + c_{\text{spuct}} \cdot P(s, g) \cdot \exp(-\beta \cdot \text{unc}(g)) \cdot \frac{\sqrt{\sum_{g'} N(s, g') + \epsilon}}{1 + N(s, g)} \right] \tag{5}$$

Our world model serves as the sandbox within which rollouts are performed, ensuring that planning remains confined to safe goals. Using uncertainty instead of an alignment score in our formulation allows for generalization, such as deferring autonomous planning also in situations where the model lacks sufficient training data or operates in noisy environments, and human control is preferred. To make sure our goal selection ist safe, we proof if we generate at least one safe goal. This goal will be picked over unsafe goals, shown as follows:

Theorem 1.

$$\exists g \in G_{safe} \Rightarrow g^* \in G_{safe}$$

Proof. We prove the statement by contradiction.

Assume that $Q(s, g) \geq 0$, $P(s, g) > 0$ for all (s, g) , $\beta > 0$, that $\epsilon > 0$, which ensures nonzero exploration incentives even when visit counts are initially zero.

Suppose, for the sake of contradiction, that $g^* \notin G_{\text{safe}}$ and arbitrary. Then, by definition,

$$g^* = f(s) + \epsilon', \quad \epsilon' \sim \mathcal{N}(0, I),$$

where the noise ϵ' is constructed such that

$$\text{unc}(g^*) \to \infty.$$

As a result,

$$\exp(-\beta \cdot \text{unc}(g^*)) = \frac{1}{\exp(\beta \cdot \text{unc}(g^*))} \to 0.$$

Multiplying with the exploration bonus term,

$$SPUCT(s, g^*) = Q(s, g^*) + c_{\text{spuct}} \cdot P(s, g^*) \cdot 0 \cdot \frac{\sqrt{\sum_{g'} N(s, g') + \epsilon}}{1 + N(s, g^*)}$$

we can conclude that the exploration bonus term becomes 0.

Since there exists a $g \in G_{\text{safe}}$, which by definition cannot be $\exp(-\beta \cdot \text{unc}(g)) = 0$. There exists a strict lower bound SPUCT(s, g) > 0 by applying our assumptions. Therefore, the total score for g^* becomes strictly lower than that of the safe goal $g \in G_{\text{safe}}$.

This contradicts $g^* \notin G_{\text{safe}}$. Therefore, for every argument of maximum g^*, it must hold that $g^* \in G_{\text{safe}}$. □

Consequently, our algorithm should compute candidate goals where at least one safe goal is identified or terminates after a predefined number of iterations if no safe goal is found, before proceeding to execute the goals in the real world.

2.3 Sub-policy Control via Model Predictive Control

Once we have the goal we can frame the action selection as an energy minimization problem:

$$\mathbf{a}^*_{1:T} = \arg\min_{\mathbf{a}_{1:T}} E(s, \mathbf{a}_{1:T}, g) = \|f_T(s, \mathbf{a}_{1:T}) - g\|^2 \tag{6}$$

This is what the sub-policy computes. The sub-policy operates via Model Predictive Control [5], using the uncertainty-aware predictor to optimize action sequences toward the selected high-level goals, avoiding states with high uncertainty. This yields an interpretable and tractable control strategy at the low level while leveraging the expressiveness of deep learning at the high level. In contrast to other algorithms, which struggles in continuous action spaces, our method naturally accommodates continuous control via MPC , while maintaining scalable and generalizable high-level reasoning through the meta-controller.

2.4 Application: Autonomous Driving

Our approach could be applied in following way to Autonomous Driving. High-level meta-controller could be responsible for planning maneuvers such as lane changes and turns by representing goals as states like coordinates, which makes decision-making more transparent for humans and leaves room for easier intervention than a purely reactive system, executing actions one after another, while low-level policies translate these maneuvers into precise vehicle actions like steering and acceleration. Safety is embedded directly into the planning process: when the vehicle enters a potentially hazardous state, such as approaching another vehicle too closely in simulation, the agent responds by leveraging aleatoric uncertainty induced through controlled noise, learning to proactively avoid unsafe configurations and could let humans take over control. Additionally, our framework dynamically switches between learned intuitive driving policies in familiar environments and more cautious, planning-based strategies in uncertain or novel scenarios, which mimics efficient human cognition. This hybrid approach could not only enhances safety but also improves driving efficiency across a wide range of real-world conditions.

3 Experiments

3.1 Setup

To evaluate our approach, we conduct experiments using the Safety Gymnasium benchmark suite [11], which is specifically designed to test agents under safety constraints such as hazardous regions. We focus on following environment representative of control challenge:

- **PointGoal1-v0**: A 2D navigation task where a point-mass robot must reach a randomly placed goal while avoiding predefined hazard zones.

These environment allow us to assess the core contributions of our approach, including safe goal-directed behavior, aleatoric uncertainty avoidance, and sample efficiency.

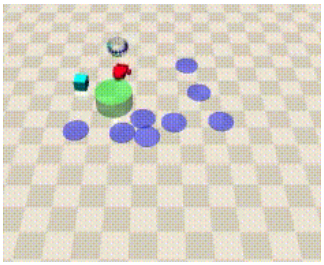

Figure 1: Environment

3.2 Baselines

We compare our method against two established baselines:

- **PPO**: A standard Proximal Policy Optimization agent trained with reward maximization only [15].
- **ShieldPPO**: ShieldPPO enhances the standard PPO algorithm by integrating a safety shield that proactively blocks actions leading to previously encountered catastrophic outcomes, thereby promoting safer behavior in reinforcement learning agents [16].

3.3 Evaluation Metrics

We evaluate agents based on the following criteria:

- **Episode Return**: Total reward accumulated during an episode.
- **Cost Violation Rate**: Percentage of episodes in which the agent triggered any constraint violations (e.g., entering a hazardous zone).
- **Safe Completion Rate**: Fraction of episodes where the goal was achieved without triggering a single violation.
- **Uncertainty Avoidance Score**: Percentage of steps where high-uncertainty states were explicitly avoided based on aleatoric variance.
- **Sample Efficiency**: How quickly the agent learns a safe and effective policy.

4 Conclusion

We presented a novel architecture for aligning artificial agents, inspired by Anderson Localization, that constrains unsafe reasoning through value-dependent obfuscation in latent space. By injecting structured noise proportional to alignment with human values, our method prevents the agent from explicitly modeling or planning around misaligned goals without requiring exhaustive enumeration of harmful states or explicitly defining a constrained safe action set.

Rather than relying on brittle reward models or expensive, potential misaligned, supervisory schemes, our approach uses a joint embedding framework that compares internal states to text-encoded representations of human values via cosine similarity. This enables rapid adaptation to new ethical priors without retraining.

We distill alignment knowledge into a world model aleatoric predictor that estimates risk through latent variance. This approach eliminates the need for multiple model compared to ensemble-based epistemic uncertainty estimation. We further incorporate this uncertainty into a modified planning algorithm, Safe PUCT (SPUCT), which prunes unsafe branches and plans in goal space. We prove that if a safe goal exists, our planner will select it.

Our hierarchical goal-selection architecture extends MuZero-style planning by reasoning over high-level latent goals rather than low-level continuous actions. This design improves tractability and robustness in real-world domains such as autonomous driving where search over continous actions is impractical.

Subpolicies use Model Predictive Control to compute smooth and interpretable action sequences that realize high-level goals. This creates a bridge between high-level planning and real-world actuation.

Overall, our framework addresses key alignment challenges: generalizing beyond training data, avoiding harmful behavior, scaling to real-world environments, and reducing compute overhead. Important open questions remain, including how to structure joint embeddings that cleanly separate harmful from aligned representations and how to suppress dangerous reasoning without impairing useful cognition.

We hope our findings encourage new directions in alignment research and contribute meaningfully to the safe development of advanced AI.

References

[1] P. W. Anderson. Absence of diffusion in certain random lattices. *Physical Review*, 109(5): 1492–1505, 1958. doi: 10.1103/PhysRev.109.1492. URL `https://link.aps.org/doi/10.1103/PhysRev.109.1492`.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL `https://arxiv.org/abs/2212.08073`.

[4] B. Bakker and J. Schmidhuber. Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. In F. Groen, N. Amato, A. Bonarini, E. Yoshida, and B. Kröse, editors, *Proceedings of the 8th Conference on Intelligent Autonomous Systems (IAS-8)*, pages 438–445, Amsterdam, The Netherlands, 2004. `https://people.idsia.ch/~juergen/ias2004.pdf`.

[5] Eduardo F Camacho and Carlos Bordons. *Model Predictive Control*. Springer, 2007.

[6] Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018. URL https://arxiv.org/abs/1810.08575.

[7] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[8] Julian Schrittwieser Ioannis Antonoglou Matthew Lai Arthur Guez Marc Lanctot Laurent Sifre Dharshan Kumaran Thore Graepel Timothy Lillicrap Karen Simonyan Demis Hassabis David Silver, Thomas Hubert. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[9] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

[10] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, pages 12519–12530, 2019.

[11] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=WZmlxIuIGR.

[12] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[13] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018. URL https://arxiv.org/abs/1811.07871.

[14] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020. doi: 10.1038/s41586-020-03051-4. URL https://www.nature.com/articles/s41586-020-03051-4.

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

[16] Shahaf S. Shperberg, Bo Liu, and Peter Stone. Learning a shield from catastrophic action effects: Never repeat the same mistake. *arXiv preprint arXiv:2202.09516*, 2022. URL https://arxiv.org/abs/2202.09516.

[17] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 14129–14142, 2020.