Free Energy Risk Metrics for Systemically Safe AI: Gatekeeping Multi-Agent Study

Michael Walters
Gaia Lab
Nuremberg, Germany

Rafael Kaufmann
Primordia Co.
Cascais, Portugal

Justice Sefas
University of British Columbia
B.C., Canada

Thomas Kopinski
Gaia Lab
Fachhochschule Südwestfalen
Meschede, Germany

January 27, 2025

Abstract

We investigate the Free Energy Principle as a foundation for measuring risk in agentic and multi-agent systems. From these principles we introduce a Cumulative Risk Exposure metric that is flexible to differing contexts and needs. We contrast this to other popular theories for safe AI that hinge on massive amounts of data or describing arbitrarily complex world models. In our framework, stakeholders need only specify their preferences over system outcomes, providing straightforward and transparent decision rules for risk governance and mitigation. This framework naturally accounts for uncertainty in both world model and preference model, allowing for decision-making that is epistemically and axiologically humble, parsimonious, and future-proof. We demonstrate this novel approach in a simplified autonomous vehicle environment with multi-agent vehicles whose driving policies are mediated by gatekeepers that evaluate, in an online fashion, the risk to the collective safety in their neighborhood, and intervene through each vehicle's policy when appropriate. We show that the introduction of gatekeepers in an AV fleet, even at low penetration, can generate significant positive externalities in terms of increased system safety.

1 Introduction

Rooted in physics, the Free Energy Principle (FEP), in tandem with Bayesian inference of world models, offers a compelling foundation in the Active Inference (ActInf) formulation of intelligent systems [Da +20; Fri+24; GB20; Hyl+24; KGT21; Lei22; MTB21; PPF22]. One of the earliest progenitors of this idea is the *Helmholtz machine*, proposed by Dayan, Hinton, Neal, and Zemel in 1995 [Day+95], connecting the statistical mechanics governing the Helmholtz Free Energy and perceptual processing. Here, treating the log-likelihood of perceptrons in a neural model as energy akin to statistical mechanics, learning proceeds as the minimization of variational free energy (VFE) through variational inference[1]. Fast-forward to the present day and the Bayesian Brain hypothesis has found popularity in neurosymbolic modeling, whereby perception and other decision/control mechanisms are driven by predictive (generative) models and hierarchical Bayesian uncertainty-resolving directives [Deu10; PF19]. For an enriching summary around FEP and its connections to Bayesian/Active Inference, see Gottwald and Braun [GB20].

The FEP in Active Inference can be applied in a few different ways [MTB21; GB20], and interpreted in many more [Da +20]. These interpretations are variations on a classic theme: exploitation vs. exploration. Whether it's accuracy vs. complexity, risk vs. ambiguity, intrinsic value vs. extrinsic value, model evidence vs. information gain, or energy vs. entropy, the mechanics of the FEP live in the tension of this duality.

To illustrate the rich connection between probabilistic modeling and the FEP, we begin with the common setup of an agent making observations o_t at time t, and wishing to infer the latent state of the world x_t through actions a_t according to policy π (which we will take as Markovian). The agent's uncertainty about x_t given its observations is expressed as the posterior $p(x_t|o_t) = p(o_t, x_t)/p(o_t)$. With the standard assumption of the intractability of $p(o_t)$, Variational Inference prescribes we instead work with a tractable approximation, $q(x_t)$ that *can* be computed.

Typically, the mismatch between $p(x)$ and $q(x)$ is quantified by the *Kullback-Leibler divergence*,

$$\mathbf{D}_{\mathrm{KL}}(q||p) = \int_x q(x) \ln\left(\frac{q(x)}{p(x)}\right) dx.$$

[1] As the authors in [GB20] point out, though VFE is not the same as Helmholtz Free Energy, the two concepts can be formally related.

We will drop the subscript t going forward in most cases when it is irrelevant. The KL divergence is convex for fixed p. Thus, the problem is recast with a new proxy objective: the minimization of $\mathbf{D}_{\mathrm{KL}}(q||p)$ through inference on q.

Finally, the KL divergence between the variational approximation of the true posterior $\mathbf{D}_{\mathrm{KL}}\big(q(x)||p(x|o)\big)$ has an intrinsic connection to the log-evidence $\ln p(o)$:

$$\begin{aligned}\mathbf{D}_{\mathrm{KL}}\big(q(x)||p(x|o)\big) &= \int_x q(x)\ln\left(\frac{q(x)p(o)}{p(x,o)}\right)dx \\ &= -\int_x q(x)\ln p(x,o)dx \\ &\quad + \int_x q(x)\ln q(x)dx \\ &\quad + \int_x q(x)\ln p(o)dx \\ &\Rightarrow \mathbb{E}_{q(x)}\left[\ln q(x) - \ln p(x,o)\right] + \ln p(o). \qquad (1)\end{aligned}$$

In line (1) we make use of the fact that $p(o)$ is independent of $q(x)$. Rearranging, we can express the evidence as

$$\begin{aligned}\ln p(o) &= \mathbf{D}_{\mathrm{KL}}\big(q(x)||p(x|o)\big) - \mathbb{E}_{q(x)}\big[\ln q(x) \\ &\qquad - \ln p(x,o)\big] \\ &= \mathbf{D}_{\mathrm{KL}}\big(q(x)||p(x|o)\big) - \mathbf{F}(q).\end{aligned}$$

The $-\mathbf{F}(q)$ term gives a floor for the evidence (since $\mathbf{D}_{\mathrm{KL}}(q||p) \geq 0$), and as the evidence $\ln p(o)$ is fixed with respect to $q(x)$, minimizing $\mathbf{F}(q)$ drives the floor up and *minimizes* the KL divergence between q and p.

As mentioned earlier, the free energy in statistical mechanics is, abstractly, the sum of an *accuracy* term (energy), and a *complexity* term (entropy). For example, for some distribution ϕ, the Helmholtz Free Energy,

$$F_H(\phi) = \langle E\rangle_\phi + \frac{1}{\beta}\,\mathrm{H}(\phi)$$

where inverse temperature β plays a weighting factor between energy and entropy. It is this similarity in form why $\mathbf{F}(q)$ is also called the *variational free energy* (VFE):

$$\begin{aligned}\mathbf{F}(q) &= -\,\mathbb{E}_{q(x)}[\ln p(x,o)] + \mathbb{E}_{q(x)}[\ln q(x)] \\ &= \underbrace{-\,\mathbb{E}_{q(x)}[\ln p(x,o)]}_{\text{“Energy”}} - \mathrm{H}[q(x)]. \qquad (2)\end{aligned}$$

The entropic term is a form of Occam's razor, encouraging models to make fewer assumptions or have too many extraneous parameters. It also functions like a regularizer against overfitting to model evidence by the energy term. In the ActInf framework, agents are driven to reduce "surprisal"—the discrepancy between their models and the world, i.e. VFE—primarily through two means ([PPF22] §2.6, [MTB21]):

- (Perception) Updating world models to better fit the evidence.
- (Action) Exploration and actions in the world to elicit desirable outcomes or reduce uncertainty.

With a generative model $p(x, o)$, artificial agents can simulate potential futures and use the expected free energy to evaluate policies and inform their decisions.

1.1 Extending into the future

The VFE-based objective discussed thus far has focused on deriving a variational model $q(x)$ through inference that both explains the data and is balanced by an entropic term. However, this falls short of how a fully equipped ActInf agent would operate intelligently: using preference-biased predicted futures to inform its actions. We defer the philosophical justification [PPF22], but in sum, incorporating a preference prior distribution $\tilde{p}(o)$ over expected outcomes (or states $\tilde{p}(x)$) embeds the goal directives of the agent into the objective—elevating it from being just a Bayesian evidence-building machine.

Inference then proceeds towards minimizing the *Expected Free Energy* (EFE) across candidate policies, where quality of fit is judged by the expected log likelihood of *desired* observations, and exploration is encouraged through maximizing the divergence between the expected variational posterior and the expected variational prior [2].

$$\begin{aligned}\mathbf{EFE}_t &\equiv \mathbb{E}_{q(o_t,x_t|\pi)}\left[\ln q(x_t|\pi) - \ln \tilde{p}(o_t, x_t)\right] \qquad (3) \\ &\approx -\underbrace{\mathbb{E}_{q(o_t,x_t|\pi)}\big[\ln \tilde{p}(o_t)\big]}_{\text{Extrinsic Value}} \qquad (4) \\ &\quad -\underbrace{\mathbb{E}_{q(o_t|\pi)}\,\mathbf{D}_{\mathrm{KL}}\big[q(x_t|o_t)||q(x_t|\pi)\big]}_{\text{Epistemic Value}}\end{aligned}$$

where $\tilde{p}(o_t, x_t) = p(o_t|x_t)\tilde{p}(x_t)$. Taking a temporal mean-field factorization of the variational posterior $q(x_{t:\tau}, \pi) \approx q(\pi)\prod_{s=t}^{\tau} q(x_s|\pi)$ and generative model $\tilde{p}(o_{t:\tau}, x_{t:\tau}) \approx \prod_{s=t}^{\tau}\tilde{p}(o_s)q(x_s|o_s)$, severs the temporal dependence between steps, meaning the optimal path is that with the lowest sum $\sum_t \mathbf{EFE}_t$.

Millidge, Tschantz, and Buckley [MTB21] give considerable contemplation to the question of extending the VFE into the future and the natural origins of the EFE[3]. The

[2] We use an approximation here that the true and approximate posteriors are similar, $q(x|o) \approx p(x|o)$, meaning inference was successful. Without this assumption, there is an additional divergence term between these two quantities.

[3] Ultimately, it is argued that the EFE—the go-to formulation in ActInf—is by no means mandatory and even less natural of a construction than their proposed *Free Energy of the Expected Future* (FEEF) alternative (not to be confused with the FEF discussed in this manuscript). The difference being that the extrinsic value of the EFE is the maximization of the log model evidence (Eq. (4)), whereas in the FEEF it is the minimization of the KL-divergence between the likelihood of observations predicted under a *veridical generative model* and the marginal likelihood of observations predicted under the *biased generative model*—

authors go on to introduce an additional FEP-based formulation, the *Free Energy of the Future* (FEF), which has an objective driven by the *minimization* of the entropic term, in stark contrast to epistemic maximization:

$$\begin{aligned} \mathbf{FEF}_t &\equiv \mathbb{E}_{q(o_t,x_t|\pi)} \left[\ln q(x_t|o_t) - \ln \tilde{p}(o_t, x_t)\right] & (5) \\ &\approx -\mathbb{E}_{q(o_t,x_t|\pi)} \left[\ln \tilde{p}(o_t|x_t)\right] \\ &\quad + \mathbb{E}_{q(o_t|\pi)} \mathbf{D}_{\mathrm{KL}} \left[q(x_t|o_t)||q(x_t|\pi)\right] & (6) \end{aligned}$$

Note the epistemic terms between the EFE and FEF differ only in their sign. Encouraging the minimization of an information-seeking term seems anathema to an ActInf agent, yet minimizing the FEF satisfies the FEP-driven goals of 1) bounding the model evidence (surprisal), and 2) minimizing the divergence between a variational posterior and a target model (whether that is based on the true world distribution or a preference prior in the context of Active Inference).

2 Cumulative Risk Exposure

We propose and showcase an arrangement that repurposes and reframes the VFE construction laid out above. The canonical Active Inference agent begins with a known preference prior that informs its actions as expected VFE computations. However, by obfuscating the preference prior from the agent—or at least the *true* stakeholder preference prior, if we still want the agent to operate in an ActInf fashion with its own preference prior—we can help buffer against certain reward specification pitfalls, like reward hacking, etc. In essence, this defines a Gatekeeper (GK) arrangement, where the GK has access to the agent's policies and can compute a policy's expected free energy according to its hidden preference prior as a form of policy evaluation and risk metric. Expressing values as preference prior distributions allows for a wide range of preference structures, including risk-aversion, social preferences, and non-Markovian utility functions [SA23].

The free energy risk metric can be utilized as context prescribes, and we demonstrate a simple method whereby a risk threshold is defined as a point of criticality demanding gatekeeper intervention[4]. To our knowledge, this is the first VFE-based gatekeeper model for agentic AI applications.

When defining a risk metric, both the FEF and the EFE provide viable options. For contexts where exploration is *discouraged*, the FEF offers a better form since its objective is minimized through low-entropy futures. This may be the better choice for safety-critical applications where minimizing unexpected behavior is preferred. Conversely, in domains with significant structural uncertainty and ambiguity (such as research and corporate strategy) or where downside risk is not deemed significant (such as arts and entertainment), an EFE formulation would encourage exploration.

2.1 Adapting for observation-space

Often it is the case that a preference prior is expressed in terms of outcomes, not hidden states. Thus, it useful to express the VFE formulae in observation-space. From the definition of EFE in Eq. (3) (dropping the time-dependence),

$$\begin{aligned} &\mathbb{E}_{q(o,x|\pi)} \left[\ln q(x|\pi) - \ln \tilde{p}(o,x)\right] \\ &= \mathbb{E}_{q(o,x|\pi)} \left[\ln q(x|\pi) - \ln \tilde{p}(o) - \ln q(x|o)\right] \\ &= \mathbb{E}_{q(o,x|\pi)} \Big[\cancel{\ln q(x|\pi)} - \ln \tilde{p}(o) - \ln q(o|x) \\ &\qquad - \cancel{\ln q(x|\pi)} + \ln q(o|\pi)\Big] \\ &= \mathbb{E}_{q(o,x|\pi)} \left[- \ln \tilde{p}(o) - \ln q(o|x) + \ln q(o|\pi)\right] \\ &= -\underbrace{\mathbb{E}_{q(o,x|\pi)} \left[\ln \tilde{p}(o)\right]}_{\text{Extrinsic}} - \underbrace{\mathbb{E}_{q(x|\pi)} \left[\mathbf{D}_{\mathrm{KL}}[q(o|x)||q(o|\pi)\right]}_{\text{Information Gain}} \end{aligned}$$

making use of the definitions $\tilde{p}(x,o) = q(x|o)\tilde{p}(o)$ and $q(x,o|\pi) = q(x|\pi)q(o|x) = q(o|\pi)q(x|o)$, and Bayes' rule. Computationally, one can estimate these values through sampling of the variational prior and the produced observations. Similar decompositions can be achieved for the FEF:

$$\begin{aligned} &\mathbb{E}_{q(o,x|\pi)} \left[\ln q(x|o) - \ln \tilde{p}(o,x)\right] \\ &= \mathbb{E}_{q(o,x|\pi)} \left[\ln q(x|o) - \ln \tilde{p}(o|x) - \ln q(x|\pi)\right] \\ &= \mathbb{E}_{q(o,x|\pi)} \Big[\ln q(o|x) + \cancel{\ln q(x|\pi)} - \ln q(o|\pi) \\ &\qquad - \ln \tilde{p}(o|x) - \cancel{\ln q(x|\pi)}\Big] \\ &= \mathbb{E}_{q(o,x|\pi)} \left[\ln q(o|x) - \ln q(o|\pi) - \ln \tilde{p}(o|x)\right] \\ &= \underbrace{-\mathbb{E}_{q(o,x|\pi)} \left[\ln \tilde{p}(o|x)\right]}_{\text{Extrinsic}} + \underbrace{E_{q(x|\pi)} \left[\mathbf{D}_{\mathrm{KL}}[q(o|x)||q(o|\pi)]\right]}_{\text{Information Gain}} \end{aligned}$$

Between the two decompositions, we see the sign flip on the epistemic term persist.

Finally, the VFE risk formulations thus far are lacking a balancing variable that weights the epistemic and extrinsic components. In analogy with free energy formulations of thermodynamics, we can introduce an inverse "temperature" to balance the terms of our risk equation. In abstract, the instantaneous risk at time t, for a variable set $\phi = [q, \tilde{p}, \pi]$ is

$$\mathcal{G}_t(\phi) = \langle E \rangle_{\phi,t} \pm \frac{1}{\beta}\mathcal{H}[\phi, t], \quad (7)$$

where E and $\mathcal{H}$ are the energetic and entropic components[5]. Recall, the EFE and FEF are expected free energy forms,

which is argued to be more aligned with the goals of an ActInf agent.

[4] A binary risk threshold is not the only option. Specifically, in a setting with continuous control variables, it would be possible to perform a smooth handover (linear combination) between agent policy and gatekeeper policy. This introduces complexities in the simulation model and will be left to future work.

which can be γ time-discounted in aggregation across time. We thus define the *Cumulative Risk Exposure* (CRE)

$$\mathcal{G}_{\Sigma}(\phi, t) = \sum_{t'}^{\tau} \gamma^{t'} \mathcal{G}_{t+t'}(\phi), \tag{8}$$

though we will commonly drop the time subscript in our discussions.

2.2 Preference prior construction

Choice of preference prior is context-dependent, but a natural form is the Boltzmann distribution over some loss function $\mathcal{L}$:

$$\tilde{p}(\mathcal{L}) = e^{-\beta\mathcal{L}}/Z, \tag{9}$$
$$Z = \sum_j e^{-\beta\mathcal{L}_j}.$$

With this formulation, the inverse temperature term in Eq. (7), which serves to balance the extrinsic and intrinsic terms, equivalently operates on the extrinsic, preference-based term instead of the intrinsic term,

$$\begin{aligned}\mathcal{G} &= -\,\mathbb{E}_q\big[\ln\tilde{p}\big] \pm \frac{1}{\beta}\mathcal{H}\\ &\Rightarrow \beta\,\mathbb{E}_q\big[\mathcal{L}\big] \pm \mathcal{H} + \ln(Z).\end{aligned}$$

Consequently, from Eq. (9) β quantifies a *tolerance* to loss, scaling $\tilde{p}$ accordingly, and can be thought of as a *preference temperature* of our system. Very strong preference biases create a "low temperature" (high β) system that is very *energetically sensitive* to preference alignment; conversely, weak preference bias creatures a smoothed out preference distribution that is more *entropy dominated*, with lower energetic sensitivity.

Further, the Boltzmann distribution has the property that the ratio of state probabilities

$$\frac{\tilde{p}(\mathcal{L}_1)}{\tilde{p}(\mathcal{L}_2)} = \exp(-\beta(\mathcal{L}_1 - \mathcal{L}_2)).$$

Thus, we can calibrate β from a maximum and minimum loss range, and those corresponding stakeholder-assigned desirabilities,

$$\begin{aligned}\ln\left(\frac{\tilde{p}_{max}}{\tilde{p}_{min}}\right) &= -\beta(\mathcal{L}_{max} - \mathcal{L}_{min})\\ \Rightarrow \beta &= \frac{\ln\left(\tilde{p}_{min}/\tilde{p}_{max}\right)}{\mathcal{L}_{max} - \mathcal{L}_{min}} \geq 0.\end{aligned}$$

β is non-negative since since by definition the desirability $\tilde{p}_{min} \geq \tilde{p}_{max}$, and $\mathcal{L}_{max} \geq \mathcal{L}_{min}$.

2.3 Extending the approach

As discussed by Hyland et al. [Hyl+24], minimizing a *joint free energy* as a sum of individual agent free energies can avail game-theoretically optimized solutions that would otherwise not be played in selfish policies. Indeed, joint free energy minimization has been postulated as a potential core mechanism behind collective agency in biological systems [SL24; ML24]. It is also translatable to the Cooperative Inverse Reinforcement Learning paradigm [Had+24], as agents model the preferences of humans and themselves. In our AV experiment, the free energy of neighboring vehicle gatekeepers is aggregated before making decisions, and could for instance deter a vehicle from speeding up because to reduce the collective free energy, at the expense of reducing their own.

Extending CRE and VFE-based metrics hierarchically affords a natural and mathematically straightforward approach to first-principles AI safety. Several contemporary AI safety proposals feature prolific construction of probabilistic models (themselves constructed from AIs, at least in part). "Guaranteed Safe AI" demands rigorous world modeling to construct formal safety guarantees [Dal+24; TO23]. Bayesian, "Scientist AIs" exert caution within uncertainty bounds according to their world models, aided by simulation, but are also expected to require potentially massive amounts of compute [Ben24; Ben+24]. Elsewhere, the Gaia Protocol[6]of globally coordinated, amortized learning, depends on LLM-aided context-dependent model construction [KL23; KL24]. There is strong overlap in each of these pursuits, grounded in the creation and exploration of probabilistic world models, and the VFE framework outlined herein provides a natural language to 1) embed safety specifications into world models, 2) direct agentic learning and exploration in their accordance, while 3) taking actions that are in the collective interest through the minimization of the joint free energy.

3 Gatekeeping Experiment

We investigated the application of this principle in a simulated autonomous vehicle (AV) setting, using a pared-back simulator, `highway-env` [Leu18], which is built on top of `gymnasium` [Tow+24]. Code for this experiment is available on Github [Wal24], and a sample video can be found here.

Our highway track featured autonomous vehicles with a variable number of these being gatekeeper controlled. We adopt (and abuse) terminology from theory-of-mind research to distinguish Alters and Egos as the two main types of vehicles on the road. Alters have a static policy and constitute the background traffic of our simulation, whereas Egos are the vehicles of interest that we optionally assign gatekeepers to, measure, etc. Our results find that the introduction of gatekeepers controlling Ego policies

[5] The inverse temperature has an interesting parallel with the Probability Dependency Graph framework, where a β term represents the degree of confidence/belief in a distribution [Ric22]. In our construction, confidence in $\tilde{p}$ can factored into $1/\beta$, but the inverse temperature carries a slightly different implication: one could be entirely confident in $\tilde{p}$ but still value including entropic contributions.

[6] Of which some of the authors are affiliated.

according to CRE has an increasingly positive impact on the road as defined by stakeholder-defined preferences.

Each investigated configuration was seeded across 1200 world simulations, for a duration of 80 steps, which was enough time to allow traffic behaviors and consequences to emerge. When computing energy and risk estimates, every 5 world steps gatekeepers ran $N_{MC} = 128$ internal Monte Carlo (MC) trajectories out to a $\tau = 10$ step horizon. N_{MC} is not exceedingly large, but for a relatively close horizon is sufficient for collecting an expectation of the upcoming future. The gatekeeper internal trajectories were fully observable—though their measurements naturally only considered neighbors within a reasonable radius.

3.1 Rewards and Loss

Our reward score was constituted from three aspects: target speed, collisions, and defensive driving. Ego vehicles received a **speed reward** R_S in the form of a Gaussian centered on a target speed v_T:

$$R_S(v) = \alpha \exp[-(v - v_T)^2/2\sigma^2], \tag{10}$$

where constants α, σ, and v_T were heuristically chosen. The **collision reward** R_C was simply a constant based on collision state $s = s_c$,

$$R_C(s) = \begin{cases} -\kappa & \text{if } s = s_c \\ 0 & \text{otherwise} \end{cases}$$

with κ heuristically chosen appropriately to ascribe high disincentive.

Braking distance—the distance it takes to come to a full stop—is a property that scales quadratically with speed [THH00]. This is an important property to capture, which we combine with the common sense that proximity is inherently more risky, to formulate our **defensive-driving reward**:

$$R_D(j) = R_{D,max} - \lambda \sum_{i \in V_j} \frac{1}{2^m d_{ij}} \left[w(i,j)^2 + \zeta \right], \tag{11}$$

$$\begin{aligned} w(i,j) =& \max(0, v_j - v_i) \times \mathrm{H}(x_i - x_j) \\ &+ \max(0, v_i - v_j) \times \mathrm{H}(x_j - x_i), \end{aligned}$$

with scalar $\lambda > 0$, vehicle index i of vehicle j's neighbors (set V_j), lane differential $m \in \{0, 1, 2, \dots\}$, and neighbor distance d_{ij}. $w(i,j)$ returns the magnitude of relative speed between j and its neighbor, using the Heaviside binary function H to control for if a neighbor is in front or behind. If vehicles j and i are drifting apart, $w(i,j)$ is 0. The constant ζ adds an additional penalty for vehicle proximity. Since the terms are penalizing, we subtract the bulk from a max reward $R_{D,max}$ and truncate to the range $R_D(j) \in [0, R_{D,max}]$. The final result is a function that 1) penalizes quadratically with relative speeds between neighbors, 2) penalizes with increased proximity $\propto 1/d_{ij}$, but 3) less so as lane differential increases. The fact that R_C is negative is appropriately handled in the reward normalization process. Loss was then simply the negative sum of rewards, and constituted our only observed variable,

$$\mathcal{L} = -\sum R_i.$$

It is worth highlighting here that the resulting improved road safety, as a consequence of gatekeeper decision-making, was achieved with this single aggregate scalar variable and did not require the suite of AV sensor inputs in its decision evaluation.

3.2 Risk formulation

Since our experiment was a fully observable environment, and we assert *ex hypothesi* that our loss and $\tilde{p}$ formulations are sufficient and accurate, we can drop any entropic contributions. In this context, therefore, CRE is identical to time-discounted expected utility[7]. Additionally, whereas the extrinsic terms in EFE/FEF are expectations over the variational model $q(x, o|\pi)$, we can directly work with $p(o, x|\pi)$ since we have a fully observable environment, and use Monte Carlo methods to approximate $p(o, x|\pi)$.

The removal of entropy simplifies the determination of our stakeholder tolerance parameter. Without exploratory requirements, the scale of $beta$ is irrelevant—as with energy in many other contexts, we are only concerned with relative values, not absolutes[8]. In other applications, β may be determined as a forced constraint: cost in dollars, quantity, etc.

Taken together, our final CRE is the expected utility

$$\begin{aligned} \mathcal{G}_\Sigma(\mathcal{L}) &= -\sum_{t'}^{\tau} \gamma^{t'} \, \mathbb{E}_{p(\mathcal{L})}[\ln \tilde{p}(\mathcal{L})] \\ &= \sum_{t'}^{\tau} \gamma^{t'} \, \mathbb{E}_{p(\mathcal{L})}[\mathcal{L}] \end{aligned} \tag{12}$$

3.3 Policies

The `highway-env` library has an automated *Intelligent Driving Model* (`IDMVehicle`) [THH00] class, which employs a combination of deterministic logic to calculate acceleration and steering. Lane changes are determined in part according to the *Minimizing overall braking induced by lane change* "MOBIL" model [KTH07], which, as advertised, tries to reduce imposed braking in its lane-change selection.

This vehicle policy is deterministic and has no machine learning or sampling involved in its decision-making. How-

[7] In future experiments involving partially-observable environments and other sources of uncertainty, the value of the complete CRE formulation given in Eq. (8) will become more evident. See, for instance, [KGT21; Da +20; PPF22; TSB20; Saj+21; Uel18; Lan+21; BB18; BB19; BFB11].

[8] The scale and shape of $\mathcal{G}$ does become relevant when its absolute value matters, such as determining a risk threshold ρ^*.

ever, there are several knobs we can tune to produce different behavior types. For the Alter vehicles, we increased their appetite and aggression for lane changes, increasing the course difficulty for Egos. We also constructed two policies for Ego vehicles called "Defensive" and "Hotshot". These differ in their comfort with braking distance and lane-change aggression—Hotshot vehicles are more comfortable with tailgating a driver in front of them if it allows them to approach their target speed or get closer to a lane change they want. They are also more likely to accept an aggressive lane change.

In our experiment, a total of 24 vehicles were divided evenly between ego and alter vehicles. Among the ego vehicles, we experimented with different fractions of them being under GK control, also termed "online". In one configuration, 4/12 ego vehicles were online, and in another 12/12. Ego vehicles would start in the Hotshot policy, so in the 4/12 arrangement, the other 8 remained Hotshot for the entire run. Those under GK control were available for policy modulation between Hotshot and Defensive, based on the gatekeeper's CRE computation from simulated futures—like a driving instructor copilot that takes over when they anticipate upcoming danger.

Since collisions are a metric of interest, conditions were set up such that these were not exceedingly rare events. During a run, 4 online ego vehicles would be tracked for a collision, the event of which would terminate a run. Additionally, if any 6+ vehicles were ever in a collision state, this would be considered a jam and the run terminated. Runs were not terminated on *any* collision because it is still valuable to measure performance of ego vehicles in adapting to such road conditions.

3.4 Gatekeeper Policy Control

For online vehicles, gatekeepers anticipate upcoming risk through internal simulations, then toggle their vehicle's policy to Defensive in risky situations, or back to Hotshot when deemed safe enough. Using Hotshot as a nominal policy may seem odd, but it gives a stronger counterbalance to observe the phenomenon of interest[9].

Gatekeepers run N_{MC} internal Monte Carlo trajectories at regular, frequent intervals in the world simulation to compute a CRE estimate, following Eq. (12). Values for a given trajectory's risk are accumulated out to an MC horizon $\tau = 10$ steps. Each vehicle's actions are not in a vacuum. Sharing local observations and predictions by opening channels of communication through gatekeepers enhances decision-making through a collective intelligence. After computing individual CREs, we replace each with the average of their local neighborhoods and have online vehicles make policy decisions from this average.

Converting from a unitless CRE value to a policy decision is not self-evident, and is open to the needs of the stakeholder. We opted for a simple CRE threshold method, where crossing the risk threshold, ρ^*, triggers a policy switch response. To avoid erratic behavior at the threshold, i.e. frequent policy switching, we employed two thresholds $\rho^*_+ = 1.1 \times \rho^*$ and $\rho^*_- = 0.9 \times \rho^*$, such that when risk crosses ρ^*_+ from below, the GK engages Defensive driving, and subsequently crossing ρ^*_- from above engages Hotshot. Additionally, we used a 10-step graduation for policy parameter deltas, to smooth policy transitions and further reduce erratic behavior. We selected a heuristic value of $\rho^* = 2$, however determining ρ^* is likely to be more straightforward in practical applications where the loss or CRE have units with bearing.

3.5 Results & Discussion

The ultimate goal here is better decision-making according to stakeholder preferences through simulated futures. To that end, our main measuring stick is the defined loss $\mathcal{L}$ and collision results. Two baselines were simulated across 1200 world runs, for the Defensive and Hotshot policies. In a given baseline, all 12 of the ego vehicles would stick to the defined policy throughout, and thus no CRE calculations were performed for GK operations. Realized rewards and loss values were still measured at each step, however.

Though the Hotshot policy has a consistently higher speed reward, it suffers in the defensive reward compared to the Defensive policy, and incurs substantially more collisions (Fig. 1). Ultimately, the erratic, dangerous Hotshot behavior garners greater loss on average.

With the introduction of online gatekeepers, we aspire for the best of both worlds: *intelligent policy selection that responds to environment conditions.* We found a considerable signal in support of this, that became increasingly pronounced proportional to GK presence. At full GK strength, crash avoidance was significantly improved, while finding opportunity to excel in defensive driving and target speed.

For the most part, the Defensive Baseline is always going to be hard to surpass: It is expected to have the fewest crashes and the highest R_D. Thus, gatekeepers need to perform comparatively well in those two dimensions while eking out gains in R_S—which is at odds with R_D and R_C. Nonetheless, the Online-12 configuration handled this remarkably well, especially for the first half of the simulation where it tracked Hotshot-level R_S while approaching Defensive-level R_D. This superior performance combination was most strongly exhibited in the Loss minimum by Online-12 around Step 25 that substantially outperformed both baselines. From visual observations, the first half of the simulation is the more dynamic portion of the simulation, requiring egos to navigate around themselves and alters more (since they have a higher target speed than alters), versus the latter portion where the road approaches more of a steady-state. (Example video.) The selection of $\rho^* = 2$ yielded modest policy switching activity, and the "Defensive Fraction" in Fig. 1 indicates that typically

[9] As our construction could also apply to gatekeeping human drivers, the Hotshot policy is not a bad model of standard driver behavior in many parts of the world.

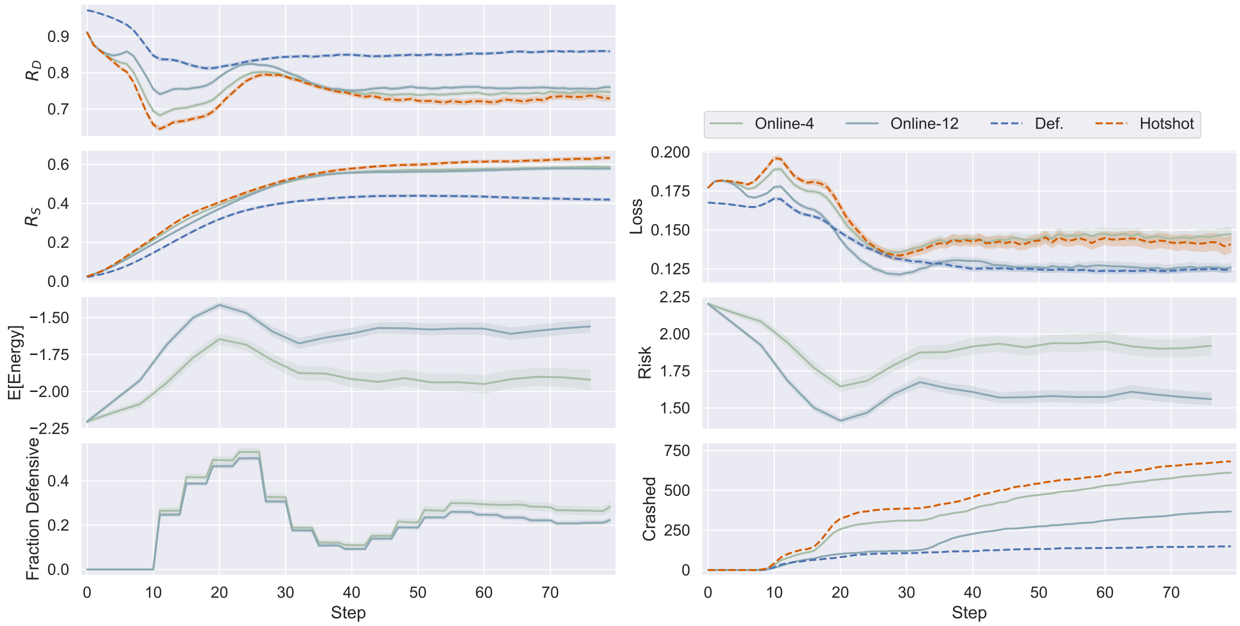

Figure 1: Baseline and gatekeeper results. Gatekeeper runs had either 4/12 or 12/12 ego vehicles online. R_D, R_S, Loss, Crashed, and Fraction Defensive are averaged realized values. Each E[Energy] and Risk measurement is across N_{MC} MC trajectories. The Fraction Defensive is the proportion of ego vehicles in the Defensive policy. Crashed is a cumulation of how many worlds have had an ego crash at or before a given step. Values are averaged across 1200 world draws, 90% CI displayed.

between 10-40% of egos would be in Defensive mode for the bulk of the run.

Collisions ("Crashed", Fig. 1) could not be wholly eliminated, but these were present even in the Defensive baseline, so this is expected. The Online-4 configuration was slightly but measurably better than the Hotshot baseline in this, though Online-12 kept in tow with the Defensive baseline for the first half of the duration before diverging. In practical applications, if stakeholders want to push something like collision likelihood down even further, they need only update their preference prior, or the loss function penalty for collision, κ.

The Energy and Risk figures are from gatekeeper MC estimates. Risk calculations consider trajectories out to $\tau = 10$ steps, so we should expect that early on with vehicles in Hotshot policy that it anticipates risk that reflects the baseline 10 steps ahead. Indeed this behavior tracks as the initial peak, subsequent dip, and plateau are anticipated by the Risk τ steps in advance. Trying to correlate spikes in Risk for Online modes with spikes in future baseline Loss becomes less accurate further into the simulation as their worlds continue to increasingly diverge after $t = 0$.

The results show a clear trend: the effect of gatekeepers produces increasingly safer roads for everyone through superior driving according to our embedded preference. They can score highly in R_S, while incorporating smarter, safer driving when needed, reducing collisions, improving their R_D scoring, and ultimately achieving better loss results than either baseline policy.

4 Conclusion

The Free Energy Principle, as one of the foundational underpinnings of Active Inference, draws powerful connections between physical energetic laws and intelligent action, with explanations for exploitation-exploration naturally emergent. Encoding stakeholder preferences via the preference prior provides a highly flexible means to direct agentic learning. The Cumulative Risk Exposure metric introduced leverages these foundations to create an interpretable, modular utility to score policies according to biased futures. The preference-temperature and tolerance mechanics outlined also introduce a conceptual and instructional foothold for usage.

Stakeholders and AI agents can employ this safety metric to anticipate upcoming high risk situations and respond intelligently, as demonstrated by our autonomous vehicle experiment, which saw increasingly superior driving performance proportional to online usage. This principle has immense potential across agentic applications as a quick and effective utility for gauging risk which, in contrast to simple loss measures, is biased towards stakeholder preferences, providing straightforward and transparent decision rules for risk governance and mitigation.

References

[BB19] Manuel Baltieri and Christopher L. Buckley. "Active Inference: Computational Models of Motor Control without Efference Copy". In:

2019 Conference on Cognitive Computational Neuroscience. Berlin, Germany, 2019. DOI: 10.32470/CCN.2019.1144-0.

[BB18] Manuel Baltieri and Christopher L. Buckley. “The Modularity of Action and Perception Revisited Using Control Theory and Active Inference”. In: *The 2018 Conference on Artificial Life*. The 2018 Conference on Artificial Life. Tokyo, Japan, 2018, pp. 121–128. DOI: 10.1162/isal_a_00031.

[Ben24] Yoshua Bengio. *Towards a Cautious Scientist AI with Convergent Safety Bounds*. https://yoshuabengio.org/2024/02/26/towards-a-cautious-scientist-ai-with-convergent-safety-bounds/. Feb. 2024.

[Ben+24] Yoshua Bengio et al. *Can a Bayesian Oracle Prevent Harm from an Agent?* Aug. 2024. DOI: 10.48550/arXiv.2408.05284.

[BFB11] Harriet Brown, Karl Friston, and Sven Bestmann. “Active Inference, Attention, and Motor Preparation”. In: *Frontiers in Psychology* 2 (2011). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2011.00218.

[Da +20] Lancelot Da Costa et al. “Active Inference on Discrete State-Spaces: A Synthesis”. In: *Journal of Mathematical Psychology* 99 (Dec. 2020), p. 102447. ISSN: 0022-2496. DOI: 10.1016/j.jmp.2020.102447.

[Dal+24] David Dalrymple et al. *Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems*. July 2024. DOI: 10.48550/arXiv.2405.06624.

[Day+95] Peter Dayan et al. “The Helmholtz Machine”. In: *Neural Computation* 7.5 (Sept. 1995), pp. 889–904. ISSN: 0899-7667. DOI: 10.1162/neco.1995.7.5.889.

[Deu10] Sid Deutsch. “Bayesian Brain: Probabilistic Approaches to Neural Coding (Doya, K., Eds., et al.; 2007) [Book Review]”. In: *IEEE Pulse* 1.3 (2010), pp. 64–65. DOI: 10.1109/MPUL.2010.939182.

[Fri+24] Karl J. Friston et al. “Designing Ecosystems of Intelligence from First Principles”. In: *Collective Intelligence* 3.1 (Jan. 2024). ISSN: 2633-9137, 2633-9137. DOI: 10.1177/26339137231222481.

[GB20] Sebastian Gottwald and Daniel A. Braun. “The Two Kinds of Free Energy and the Bayesian Revolution”. In: *PLOS Computational Biology* 16.12 (Dec. 2020). ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008420.

[Had+24] Dylan Hadfield-Menell et al. *Cooperative Inverse Reinforcement Learning*. Feb. 2024. DOI: 10.48550/arXiv.1606.03137.

[Hyl+24] David Hyland et al. “Free-Energy Equilibria: Toward a Theory of Interactions Between Boundedly-Rational Agents”. In: *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*. 2024. URL: https://openreview.net/forum?id=4Ft7DcrjdO.

[KGT21] Rafael Kaufmann, Pranav Gupta, and Jacob Taylor. “An Active Inference Model of Collective Intelligence”. In: *Entropy* 23.7 (7 July 2021), p. 830. ISSN: 1099-4300. DOI: 10.3390/e23070830.

[KL23] Rafael Kaufmann and Roman Leventov. *Gaia Network: A Practical, Incremental Pathway to Open Agency Architecture*. Dec. 2023. URL: https://www.lesswrong.com/posts/AKBkDNeFLZxaMqjQG/gaia-network-a-practical-incremental-pathway-to-open-agency.

[KL24] Rafael Kaufmann and Roman Leventov. *Gaia Network: An Illustrated Primer*. Jan. 2024. URL: https://forum.effectivealtruism.org/posts/BaoA3gz7xRaqn764J/gaia-network-an-illustrated-primer.

[KTH07] Arne Kesting, Martin Treiber, and Dirk Helbing. “General Lane-Changing Model MOBIL for Car-Following Models”. In: *Transportation Research Record* 1999.1 (Jan. 2007), pp. 86–94. ISSN: 0361-1981. DOI: 10.3141/1999-10.

[Lan+21] Pablo Lanillos et al. *Active Inference in Robotics and Artificial Agents: Survey and Challenges*. Dec. 2021. DOI: 10.48550/arXiv.2112.01871.

[Lei22] Felix Leibfried. *Variational Inference for Model-Free and Model-Based Reinforcement Learning*. Dec. 2022. DOI: 10.48550/arXiv.2209.01693.

[Leu18] Edouard Leurent. *An Environment for Autonomous Driving Decision-Making*. https://github.com/eleurent/highway-env. 2018.

[ML24] Patrick McMillen and Michael Levin. “Collective Intelligence: A Unifying Concept for Integrating Biology across Scales and Substrates”. In: *Communications Biology* 7.1 (Mar. 2024), p. 378. ISSN: 2399-3642. DOI: 10.1038/s42003-024-06037-4.

[MTB21] Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. “Whence the Expected Free Energy?” In: *Neural Computation* 33.2 (Feb. 2021), pp. 447–482. ISSN: 0899-7667. DOI: 10.1162/neco_a_01354.

[PF19] Thomas Parr and Karl J. Friston. “Generalised Free Energy and Active Inference”. In: *Biological Cybernetics* 113.5–6 (Dec. 2019), pp. 495–513. ISSN: 0340-1200, 1432-0770. DOI: 10.1007/s00422-019-00805-w.

[PPF22] Thomas Parr, Giovanni Pezzulo, and K. J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, Massachusetts: The MIT Press, 2022. 296 pp. ISBN: 978-0-262-04535-3.

[Ric22] Oliver E. Richardson. *Loss as the Inconsistency of a Probabilistic Dependency Graph: Choose Your Model, Not Your Loss Function*. Feb. 2022. URL: http://arxiv.org/abs/2202.11862.

[Saj+21] Noor Sajid et al. "Active Inference: Demystified and Compared". In: *Neural Computation* 33.3 (Mar. 2021), pp. 674–712. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco_a_01357.

[SL24] Lakshwin Shreesha and Michael Levin. "Stress Sharing as Cognitive Glue for Collective Intelligences: A Computational Model of Stress as a Coordinator for Morphogenesis". In: *Biochemical and Biophysical Research Communications* 731 (Oct. 2024), p. 150396. ISSN: 0006-291X. DOI: 10.1016/j.bbrc.2024.150396.

[SA23] Joar Skalse and Alessandro Abate. "On the Limitations of Markovian Rewards to Express Multi-Objective, Risk-Sensitive, and Modal Tasks". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Uncertainty in Artificial Intelligence. July 2023, pp. 1974–1984. URL: https://proceedings.mlr.press/v216/skalse23a.html.

[TO23] Max Tegmark and Steve Omohundro. *Provably Safe Systems: The Only Path to Controllable AGI*. Sept. 2023. DOI: 10.48550/arXiv.2309.01933.

[Tow+24] Mark Towers et al. *Gymnasium: A Standard Interface for Reinforcement Learning Environments*. https://arxiv.org/abs/2407.17032. 2024.

[THH00] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. *Congested Traffic States in Empirical Observations and Microscopic Simulations*. Aug. 2000. DOI: 10.48550/arXiv.cond-mat/0002177.

[TSB20] Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. "Learning Action-Oriented Models through Active Inference". In: *PLOS Computational Biology* 16.4 (Apr. 2020), e1007805. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007805.

[Uel18] Kai Ueltzhöffer. "Deep Active Inference". In: *Biological Cybernetics* 112.6 (Dec. 2018), pp. 547–573. ISSN: 0340-1200, 1432-0770. DOI: 10.1007/s00422-018-0785-7.

[Wal24] Michael Walters. *Autonomous Vehicle Study Repository*. https://github.com/m-walters/av-agents. May 2024.