Department of Creative Informatics
Graduate School of Information Science and Technology
THE UNIVERSITY OF TOKYO

Master's Thesis

MENACE as a Bayesian Observer: A Technical Analysis through the Free Energy Principle

ベイズ観測者としての MENACE：
自由エネルギー原理による技術的分析

Krzysztof Woś
クシシュトフ・ヴォシュ

Supervisor: Professor Hideki Nakayama

February 2026

Abstract

Active Inference under the Free Energy Principle (FEP) provides a unified objective for perception, learning, and action. Yet, despite its explanatory scope, end-to-end computational agents that are clearly driven by an explicit expected free energy objective remain difficult to build and to validate in fully enumerable domains. This thesis uses Donald Michie's MENACE—an interpretable matchbox-and-bead learner for Tic-Tac-Toe—as a concrete, fully analysable bridge between a working learning mechanism and an Active Inference interpretation.

We map MENACE's 287 matchboxes and bead updates to a Dirichlet–categorical model: bead counts act as Dirichlet pseudo-counts and random bead draws implement posterior predictive probability matching. Under explicit modelling commitments, MENACE corresponds to an instrumental special case of expected free energy minimisation ($\lambda = 0$), while Active Inference variants introduce epistemic value via the mutual information $I(o;\theta)$. We then test the correspondence empirically against Active Inference variants and tabular reinforcement learning baselines under controlled curricula and evaluation against optimal play.

The results quantify the exploration–exploitation trade-off in this setting: MENACE's exploration pressure decays endogenously as Dirichlet concentration increases, whereas a fixed epistemic weight λ fixes only a trade-off coefficient (the epistemic term itself typically diminishes with posterior concentration). With broader state-action coverage and a larger budget, tabular RL can also approach minimax performance, highlighting that the key differences are sample efficiency, opponent-distribution sensitivity, and interpretability rather than asymptotic capability. By grounding Active Inference in a mechanically implementable system, the thesis provides a concrete test case for methodological debates about the FEP and clarifies which aspects of MENACE are explained by Active Inference and which require additional assumptions.

概要

自由エネルギー原理（FEP）に基づく能動的推論は、知覚、学習、行動に対する統一的な目的関数を提供する。しかし、その説明範囲の広さにもかかわらず、明示的な期待自由エネルギー目標によって明確に駆動されるエンドツーエンドの計算エージェントは、完全に列挙可能なドメインにおいても構築および検証が困難である。本論文では、Donald Michie が考案した MENACE（三目並べのための解釈可能なマッチボックスとビーズによる学習器）を、動作する学習メカニズムと能動的推論の解釈をつなぐ具体的で完全に分析可能な橋渡しとして使用する。

MENACE の 287 個のマッチボックスとビーズ更新をディリクレ・カテゴリカルモデルに対応付ける：ビーズの数はディリクレの擬似カウントとして機能し、ランダムなビーズの抽出は事後予測確率マッチングを実装する。明示的なモデリングの下で、MENACE は期待自由エネルギー最小化の道具的な特殊ケース（$\lambda = 0$）に対応し、能動的推論の変種は相互情報量 $I(o;\theta)$ を通じて認識論的価値を導入する。次に、制御されたカリキュラムと最適プレイに対する評価の下で、能動的推論の変種およびテーブル型強化学習ベースラインに対して経験的に対応関係を検証する。

結果は、この設定における探索と活用のトレードオフを定量化する：MENACE の探索圧力はディリクレ濃度が増加するにつれて内生的に減衰するのに対し、固定された認識論的重み λ はトレードオフ係数のみを固定する（認識論的項自体は通常、事後濃度とともに減少する）。より広い状態行動カバレッジとより大きな予算があれば、テーブル型 RL もミニマックス性能に近づくことができ、主要な違いはサンプル効率、対戦相手分布感度、および解釈可能性であり、漸近的能力ではないことを強調する。能動的推論を機械的に実装可能なシステムに基盤づけることにより、本論文は FEP に関する方法論的議論の具体的なテストケースを提供し、MENACE のどの側面が能動的推論によって説明され、どの側面が追加の仮定を必要とするかを明確にする。

Contents

List of Figures

List of Tables

Part I

Theoretical Foundations

Chapter 1

Introduction

This chapter frames the problem motivating the thesis, states the research questions, and explains why MENACE is a suitable bridge between historical reinforcement learning mechanisms and contemporary Active Inference. The aim is to set expectations, delimit scope, and establish the terminology and structure followed in the remaining chapters.

1.1 Motivation and central issue

Multiple computational implementations of Active Inference exist for discrete settings, including pymdp [1], ActiveInference.jl [2], and various scalable approximations. However, end-to-end agents whose planning and learning derive from a single expected free energy objective—without importing RL-style value functions or ad hoc exploration bonuses—remain comparatively rare in mainstream benchmarking. Most practical systems mix reinforcement-learning-style value updates with heuristic exploration, making it difficult to determine which behaviours genuinely arise from EFE minimisation.

Tic-Tac-Toe offers a tractable counterexample: every state can be enumerated, enabling exact accounting of preferences, beliefs, and exploration pressure. This thesis uses MENACE as a compact, fully analysable case study to clarify what is already "Active Inference-like" in a simple learning system and where explicit epistemic objectives change the exploration–exploitation trade-off.

Donald Michie's Machine Educable Noughts And Crosses Engine (MENACE) [3] is particularly suited to this purpose. Built from matchboxes and beads, it sidesteps perception by enumerating states and focuses solely on reinforcement—an architectural choice that makes it an ideal Rosetta Stone for connecting a physical learner to the Free Energy Principle (FEP) and Active Inference. We adopt an objective, falsifiable stance: the goal is to expose where the correspondence holds, which modelling commitments are required, and where the analogy breaks down.

1.2 Research questions and scope

- How can MENACE's matchboxes, beads, and update rules be mapped to the random variables and update equations of Active Inference under the FEP?
- Which aspects of exploration–exploitation in MENACE arise implicitly from posterior uncertainty, and which would require explicit epistemic objectives?
- How do instrumental ($\lambda = 0$) and epistemic ($\lambda > 0$) Active Inference variants compare to MENACE and tabular reinforcement learning baselines in a fully enumerable game?
- What limits apply to this correspondence (e.g., opponent modelling choices, preference specification, evaluation budget)?

The scope is deliberately narrow: we study Tic-Tac-Toe, finite-horizon play, and discrete-state Active Inference formulations. No claims are made about neuroscientific plausibility or performance beyond the stated domain.

1.3 Contributions

- A definition-before-use mapping between MENACE, reinforcement learning constructs, and Active Inference variables, aligned to a consistent notation (formalised in Chapter 2).
- A precise statement of the minimal generative model under which MENACE realises an instrumental special case of EFE minimisation ($\lambda = 0$), clarifying which assumptions are required.
- Clarification of historical and implementation details (state filters, bead schedules, reinforcement semantics) that resolves common ambiguities about reproducing MENACE.
- An experimental comparison of MENACE, instrumental/epistemic Active Inference variants, and tabular reinforcement learning baselines under controlled curricula, with ablations on state filters and restocking.
- Public artifacts—tables, figures, and code—that make the correspondence reproducible and auditable.

1.4 Thesis outline

Part I introduces the mathematical preliminaries, the expected free energy decomposition, and MENACE's historical context. Part II constructs the MENACE–Active Inference correspondence explicitly. Part III specifies the experimental design and reports empirical results. Part IV draws conclusions, states limitations, and outlines concrete directions for future work.

1.5 Historical framing

The problem of understanding how intelligent systems learn from experience has occupied researchers since the inception of both psychology and computer science. A particularly illuminating approach emerged in 1961 when Michie constructed MENACE. The device demonstrated that a purely mechanical system could learn to play Tic-Tac-Toe at near-optimal levels through reinforcement of successful strategies. Situated within the broader trial-and-error lineage from animal conditioning to modern reinforcement learning [4, 5], MENACE challenged the prevailing belief that a program could never outperform its programmer [6].

Michie's own accounts were explicit about what MENACE did and did not mechanise. In "Trial and Error" he split learning into "classification of the stimulus" and "reinforcement of the response," acknowledging that the former was "quite extraordinarily complicated" but that reinforcement "is much more tractable" once the discrete situations can be enumerated. MENACE sidesteps the classification challenge by enumerating every board pattern and focuses solely on reinforcement—an architectural choice that makes the system ideal for a Dirichlet–categorical analysis.

In parallel, theoretical neuroscience has advanced a complementary account of adaptive behaviour. The Free Energy Principle (FEP), as formulated by Friston and colleagues, posits that adaptive systems act to minimise variational free energy, a quantity that bounds surprise or prediction error [7, 8]. The FEP has attracted attention for its promise to unify perception, action, and learning; it has also generated debate about scope, falsifiability, and explanatory ambition [9, 10, 11]. By placing MENACE within this framework, we can evaluate concrete, falsifiable claims about how much of MENACE's behaviour is explained by an instrumental EFE objective and how much would require explicit epistemic drives or richer generative models.

1.6 Michie's motivating question

> In simple games for which individual storage of all past board positions is feasible, is any optimal learning algorithm known? ... The difficulty lies in costing the acquisition of information for future use at the expense of present expected gain. A means of expressing the value of the former in terms of the latter would lead directly to the required algorithm. [6]

Because Tic-Tac-Toe admits exhaustive enumeration, the Free Energy Principle provides exactly such a cost accounting: risk (the instrumental preference term, scored as cross-entropy or KL divergence depending on decomposition) and epistemic value (Dirichlet–categorical mutual information) share common units, allowing us to quantify how much information MENACE gathers by chance and how much an explicit epistemic drive would add.

1.7 Trial-and-error lineage

MENACE emerged from a lineage of trial-and-error machines linking behavioural psychology, cybernetics, and contemporary reinforcement learning. Thorndike's puzzle boxes framed learning as the gradual strengthening of stimulus–response associations, and early cybernetic projects—such as Grey Walter's tortoises—explored mechanical reinforcement with minimal internal state [5]. Michie's motivation followed this pattern: by enumerating every board position, MENACE could skip the perceptual classification problem and focus solely on reinforcement. This historical through-line underscores why Tic-Tac-Toe, with its finite decision space, is an ideal laboratory for connecting MENACE to the Bayesian formalism of the Free Energy Principle.

1.8 Summary

This chapter posed the problem of interpreting MENACE through Active Inference, stated the research questions, and outlined the contributions and structure of the thesis. The next chapter introduces the mathematical preliminaries and the shared notation that the remainder of the document follows.

Chapter 2

Mathematical Preliminaries

This chapter establishes the notation, information-theoretic quantities, and conjugacy results used throughout the thesis. The focus is on Dirichlet–categorical structure because it underpins both MENACE's bead mechanics and the expected free energy decomposition used in later chapters. Definitions appear before use, with short notes on how each concept is applied in the analysis.

2.1 Notation and Conventions

We adopt the conventions listed in Table 2.1. Random variables use uppercase letters, their realisations lowercase; expectations explicitly name the distribution (e.g., $\mathbb{E}_q[\cdot]$). We use $p(\cdot)$ for generative-model and preference distributions, and $q(\cdot)$ for predictive or posterior beliefs. Dirichlet concentration parameters are denoted α, categorical parameters θ, and policies π; the preference vector is C.

Table 2.1. Notation and symbols used throughout the thesis.

Symbol	Meaning (domain)
$o \in \{\text{win}, \text{draw}, \text{loss}\}$	Terminal outcome from the agent's perspective
$s \in S$, $a \in A(s)$	Board state and legal action in that state
$p(o \mid C)$	Preference distribution over terminal outcomes
$q(o \mid \pi)$	Predicted outcome distribution under policy π
$\alpha, \alpha_{s,a}$	Dirichlet concentration (global; per-state/action bead counts)
θ	Categorical parameter vector ($\sum_i \theta_i = 1$)
π	Policy (action distribution or sequence)
$H[\cdot]$	Entropy of a distribution
$I(\cdot;\cdot)$	Mutual information between random variables
β_{amb}	Outcome-ambiguity weight on $H[q(o \mid \pi)]$ (set to 0 in Part III)
λ	Epistemic-weight coefficient in expected free energy
λ_{policy}	Policy softmax temperature (CLI: `--policy-lambda`)

Conventions: (i) $q(\cdot)$ vs. $p(\cdot)$ as above; (ii) Dirichlet concentration totals are $\alpha_0 = \sum_i \alpha_i$; (iii) preferences C are normalised such that $p(o \mid C)$ is a categorical distribution; (iv) when indexing state-specific parameters, we use double subscripts (e.g., $\alpha_{s,a}$); (v) all Tic-Tac-Toe counts refer to canonical, symmetry-reduced states unless stated otherwise; (vi) X moves first; MENACE is X; (vii) outcome vectors are ordered (win, draw, loss) throughout; (viii) logarithms are natural (information measured in nats).

2.2 Information Measures

■Entropy. Entropy quantifies the expected surprisal of a discrete distribution. For categorical $p(x)$,

$$H[p] = -\sum_x p(x) \ln p(x). \tag{2.1}$$

How we use it. Entropy measures predictive uncertainty (e.g., over terminal outcomes) and appears in both the expected free energy decomposition and the Dirichlet–categorical uncertainty analysis.

■Mutual information. Mutual information measures how much knowing one variable reduces uncertainty about another:

$$I(X;Y) = H[X] - H[X \mid Y]. \tag{2.2}$$

How we use it. The epistemic term $I(o;\theta)$ quantifies expected information gain about Dirichlet parameters from observing outcomes, providing the explicit epistemic-value term in the expected free energy objective.

2.3 Dirichlet Distributions and Categorical Conjugacy

In the study of discrete probability distributions, the Dirichlet distribution occupies a position of central importance analogous to that of the Gaussian distribution in continuous settings. Just as the Gaussian serves as the conjugate prior for the mean of another Gaussian, the Dirichlet distribution serves as the conjugate prior for the parameters of a categorical distribution. This conjugacy relationship is what makes Bayesian inference tractable in discrete settings and, as we shall demonstrate, is implicitly exploited by MENACE's learning mechanism.

To develop this theory properly, we must first establish notation and fundamental concepts. Consider a discrete random variable that can take one of k distinct values. The probability of observing each value is governed by a categorical distribution with parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, where θ_i represents the probability of outcome i. These parameters must satisfy the constraints $\theta_i \geq 0$ for all i and $\sum_{i=1}^{k} \theta_i = 1$, defining what is known as the $(k-1)$-dimensional probability simplex Δ^{k-1}.

The challenge in Bayesian inference is to maintain and update beliefs about these unknown probabilities θ as we observe data. The Dirichlet distribution provides an elegant solution to this challenge, serving as a probability distribution over probability distributions—a concept that may seem abstract but proves remarkably natural in practice.

Definition 2.1 (Dirichlet Distribution). *A random vector $\theta = (\theta_1, \ldots, \theta_k)$ follows a Dirichlet distribution with parameters $\alpha = (\alpha_1, \ldots, \alpha_k)$, denoted $\theta \sim \mathrm{Dir}(\alpha)$, if its probability density function is:*

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{2.3}$$

where $\theta_i \geq 0$, $\sum_i \theta_i = 1$, $\alpha_i > 0$, and $B(\alpha)$ is the multivariate beta function:

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)} \tag{2.4}$$

The gamma function $\Gamma(z)$ *is defined as:*

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \tag{2.5}$$

for $z > 0$*, with the property that* $\Gamma(n) = (n-1)!$ *for positive integers* n*. The normalising constant* $B(\alpha)$ *ensures that the density integrates to unity over the probability simplex; closed-form evaluation via gamma functions makes Bayesian inference tractable.*

The parameters α_i of the Dirichlet distribution have a remarkably intuitive interpretation: they can be thought of as pseudo-counts representing prior observations of each category. This interpretation becomes precise when we consider the expected value of the distribution: $\mathbb{E}_{\mathrm{Dir}(\alpha)}[\theta_i] = \alpha_i/\alpha_0$, where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$ is the total pseudo-count. Thus, a Dirichlet distribution with parameters $(3, 2, 5)$ can be interpreted as encoding the belief arising from having previously observed 3 instances of category 1, 2 instances of category 2, and 5 instances of category 3.

This count interpretation is not merely a convenient metaphor—it is fundamental to understanding how Bayesian updating works in discrete settings. When we observe new data, the posterior distribution remains Dirichlet with updated parameters that simply add the observed counts to the prior counts. This remarkable property is formalised in the following theorem:

How we use it. Bead counts in MENACE act as Dirichlet concentrations: pseudo-counts summarise past evidence and determine posterior-predictive action probabilities.

Theorem 2.2 (Conjugacy)**.** *If* $\theta \sim \mathrm{Dir}(\alpha)$ *and we observe* n *categorical outcomes with* n_i *occurrences of category* i*, then the posterior distribution is:*

$$\theta | data \sim \mathrm{Dir}(\alpha + n) \tag{2.6}$$

where $n = (n_1, \ldots, n_k)$*.*

Proof. By Bayes' theorem:

$$p(\theta|\text{data}) \propto p(\text{data}|\theta) p(\theta) \tag{2.7}$$

For categorical likelihood:

$$p(\text{data}|\theta) = \prod_{i=1}^{k} \theta_i^{n_i} \tag{2.8}$$

Therefore:

$$p(\theta|\text{data}) \propto \prod_{i=1}^{k} \theta_i^{n_i} \cdot \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} = \prod_{i=1}^{k} \theta_i^{\alpha_i + n_i - 1} \tag{2.9}$$

This is the kernel of $\mathrm{Dir}(\alpha + n)$, completing the proof. □

How we use it. Conjugacy justifies the additive bead updates used throughout the thesis and enables the closed-form mutual information expressions referenced in later chapters.

2.4 Expected Values and Uncertainty

The conjugacy property established above provides computational tractability, but to understand how the Dirichlet distribution represents uncertainty about categorical probabilities, we must examine its moments and information-theoretic properties. These quantities will prove crucial in understanding how MENACE balances exploration and exploitation through its representation of uncertainty.

For a Dirichlet distribution with parameters α, the expected value of each component is given by a simple ratio:

$$\mathbb{E}_{\mathrm{Dir}(\alpha)}[\theta_i] = \frac{\alpha_i}{\alpha_0} \tag{2.10}$$

where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$ is termed the concentration parameter or precision. This formula confirms our interpretation of the α_i as pseudo-counts: the expected probability of category i is simply the proportion of counts (real or pseudo) allocated to that category.

The variance of each component reveals how the concentration parameter controls uncertainty:

$$\mathrm{Var}[\theta_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{2.11}$$

This expression has several noteworthy properties. First, the variance decreases as $O(1/\alpha_0)$ as we accumulate more observations, reflecting increasing confidence in our estimates. Second, for fixed α_0, the variance is maximised when $\alpha_i = \alpha_0/2$, corresponding to maximum uncertainty about a binary outcome. Third, the factor $(\alpha_0 + 1)$ in the denominator arises from the additional uncertainty inherent in the Dirichlet distribution compared to a fixed categorical distribution. Note that under posterior-predictive probability matching (used by MENACE), selection probability depends on posterior mean proportions α_i/α_0, not on variance directly; what shrinks with experience is the effective learning rate (each update becomes a smaller fractional change to the proportions).

To quantify the overall (predictive) uncertainty about the next outcome, we compute the entropy of the posterior-predictive categorical distribution:

$$H[\mathrm{Cat}(\mathbb{E}[\theta])] = -\sum_{i=1}^{k} \frac{\alpha_i}{\alpha_0} \ln \frac{\alpha_i}{\alpha_0} \tag{2.12}$$

This quantity is the predictive entropy H_{pred}—the uncertainty over the next observation after marginalising out parameter uncertainty. It is useful to separate this predictive term into aleatoric and epistemic contributions. The expected (aleatoric) categorical entropy under the Dirichlet posterior is

$$H_{\text{ale}} = \mathbb{E}_{\theta\sim\mathrm{Dir}(\alpha)}\left[-\sum_i \theta_i \ln \theta_i\right] = -\sum_i \frac{\alpha_i}{\alpha_0}\big(\psi(\alpha_i{+}1) - \psi(\alpha_0{+}1)\big), \tag{2.13}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function (logarithmic derivative of the gamma function). The epistemic component is captured by the mutual information between future outcomes and the parameters:

$$I(o;\theta) \;=\; H_{\text{pred}} - H_{\text{ale}} \;=\; \sum_i \frac{\alpha_i}{\alpha_0}\left[\psi(\alpha_i{+}1) - \psi(\alpha_0{+}1) - \ln\frac{\alpha_i}{\alpha_0}\right] \;\geq\; 0. \tag{2.14}$$

Note that H_{pred} depends only on the *proportions* α_i/α_0: scaling $\alpha \mapsto c\alpha$ leaves H_{pred} unchanged. What decreases with concentration is the *epistemic* component $I(o;\theta)$. In the limit $\alpha_0 \to \infty$ (holding α_i/α_0 fixed), parameter uncertainty collapses, $I(o;\theta) \to 0$, and the expected categorical entropy H_{ale} approaches the predictive entropy H_{pred}.

The relationship between the Dirichlet parameters and uncertainty has profound implications for sequential decision-making. An agent maintaining Dirichlet beliefs naturally exhibits exploration behaviour: when counts are small, posterior uncertainty is large, which increases the probability of trying under-sampled actions under posterior sampling schemes. This uncertainty-driven exploration can be implemented through two related mechanisms: Thompson sampling [12, 13], where one samples parameters $\theta \sim \text{Dir}(\alpha)$ then selects $a = \arg\max_i \theta_i$, or posterior-predictive probability matching, where actions are selected with probability α_i/α_0. Both approaches avoid hand-designed exploration schedules by tying exploration pressure to posterior uncertainty. As we shall see, MENACE implements the latter approach through its bead-drawing mechanism.

2.5 Summary

This chapter set the notation, information measures, and Dirichlet–categorical results that ground the rest of the thesis. The next chapter applies these tools to the Free Energy Principle, defining variational and expected free energy in the discrete setting we use for MENACE.

Chapter 3

The Free Energy Principle for Discrete State Spaces

This chapter introduces the Free Energy Principle (FEP) and Active Inference in the discrete setting used throughout the thesis. We define variational free energy F for inference, expected free energy G for policy evaluation, and make explicit the assumptions (categorical distributions, fully observed board states, mean-field policy/state factorisation) needed to align the mathematics with MENACE. The exposition is computational and domain-specific: claims are restricted to the generative models and preferences we specify rather than broad statements about neuroscience or cognition.

Having established the mathematical foundations of Dirichlet–categorical inference, we now turn to the Free Energy Principle itself. This principle, developed by Karl Friston and colleagues, provides a unified mathematical framework for understanding perception, action, and learning in adaptive systems. While the principle applies broadly to continuous and discrete systems alike, our focus here is on its formulation for discrete state spaces, as this is the setting most directly relevant to MENACE.

The Free Energy Principle rests on a deceptively simple premise: adaptive systems act to minimise surprise, where surprise is defined information-theoretically as the negative log probability of observations. Since surprise itself cannot be computed directly (it would require knowing the true generative process of observations), systems instead minimise an upper bound on surprise known as variational free energy. This quantity, borrowed from statistical physics and variational Bayesian methods, serves as a tractable objective that, when minimised, ensures both accurate perception and adaptive action.

3.1 Variational Free Energy

Under the FEP, an agent maintains a generative model that specifies its beliefs about how observations are generated. For discrete systems engaged in sequential decision-making, this generative model takes the form $p(o, s, \pi)$, encoding beliefs about observations o, hidden states s, and policies π. Here, a policy represents a sequence of actions or a mapping from states to actions—in MENACE's case, this is the strategy for selecting moves in each board position. The agent also maintains an approximate posterior distribution $q(s, \pi)$ representing its current beliefs about states and policies given observations.

The variational free energy quantifies the divergence between these beliefs and the true posterior that would result from exact Bayesian inference [8]:

$$F = \mathbb{E}_{q(s,\pi)}[\ln q(s, \pi) - \ln p(o, s, \pi)] \tag{3.1}$$

This can be decomposed as:

$$F = D_{\mathrm{KL}}[q(s,\pi)\|p(s,\pi|o)] - \ln p(o) \tag{3.2}$$

This decomposition reveals the dual nature of free energy minimisation. The first term, the Kullback-Leibler (KL) divergence, measures how far the approximate posterior $q(s,\pi)$ is from the true posterior $p(s,\pi|o)$. The second term is the log model evidence or marginal likelihood. Since $\ln p(o)$ is constant with respect to q, minimising F is equivalent to minimising the KL divergence—that is, making the approximate posterior as close as possible to the true posterior.

Throughout, we assume a mean-field factorisation $q(s,\pi) = q(s \mid \pi)\,q(\pi)$ for discrete states and policies and a fully observed board (the observation model is identity), aligning the math with the Tic-Tac-Toe setting.

This formulation immediately suggests a principle for perception: by adjusting beliefs $q(s)$ to minimise free energy, an agent performs approximate Bayesian inference, updating its beliefs about hidden states to best explain observations. However, the Free Energy Principle goes beyond passive inference to encompass action selection, and this is where the framework becomes particularly powerful.

Throughout this thesis we reserve F for *variational* free energy (inference about the present) and G for *expected* free energy (planning over future outcomes under candidate policies).

3.2 Expected Free Energy

For policy selection, Active Inference extends the free energy framework by introducing expected free energy [14, 8]:

$$G(\pi) = \mathbb{E}_{q(o_{t:T}, s_{t:T}|\pi)}[\ln q(s_{t:T}|\pi) - \ln p(o_{t:T}, s_{t:T}|\pi)]. \tag{3.3}$$

This expectation can be rearranged in several illuminating ways. A form that is especially useful for the present discussion separates epistemic and instrumental motives:

$$G(\pi) = \underbrace{\mathbb{E}_{q(o|\pi)}[-\ln p(o \mid C)]}_{\text{risk / instrumental cost}} + \underbrace{\mathbb{E}_{q(s|\pi)}\big[H\big(p(o \mid s)\big)\big]}_{\text{likelihood ambiguity}} - \underbrace{\mathbb{E}_{q(o|\pi)}\big[D_{\mathrm{KL}}\big(q(s \mid o,\pi) \parallel q(s \mid \pi)\big)\big]}_{\text{epistemic value}}. \tag{3.4}$$

We follow the standard Active Inference convention: policies are selected by *minimising* $G(\pi)$, so epistemic value enters with a negative sign (information gain is something the agent seeks, not pays). The risk term scores how well a policy is expected to realise prior preferences encoded in C (negative expected log preference). The likelihood ambiguity term penalises policies that lead to uncertain observations under the agent's likelihood model. Many expositions emphasise the risk+ambiguity form of $G(\pi)$ [8]. The decomposition above makes explicit how an epistemic drive can be switched off by setting its weight to zero.

Remark 3.1 (Terminology Note: Likelihood Ambiguity vs Outcome Entropy)**.** *In the standard expected-free-energy decomposition above, the term labelled* likelihood ambiguity *is* $\mathbb{E}_{q(s|\pi)}\big[H(p(o \mid s))\big]$*, which penalises policies that lead to intrinsically uncertain observations under the agent's likelihood model. In deterministic, fully observed Tic-Tac-Toe with an identity observation model, this likelihood ambiguity is identically zero because the agent observes the board state without noise.*

Later chapters introduce an optional outcome-entropy penalty $H[q(o \mid \pi)]$*—the entropy of the predicted* terminal *outcome distribution—as a practical proxy for uncertainty in-*

duced by an unknown opponent. To avoid conflating this with likelihood ambiguity, we refer to this optional term as outcome ambiguity *and weight it by* β_{amb} *(Table 8.1). All experiments reported in this thesis set* $\beta_{\text{amb}} = 0$*; the term is retained in the objective for completeness but does not affect reported results.*

In many Active Inference treatments, epistemic value is expressed as expected information gain about hidden states. In this thesis we instantiate the epistemic term as mutual information between outcomes and Dirichlet parameters θ that encode uncertainty about action consequences (and, where relevant, opponent responses) in each canonical Tic-Tac-Toe state. This choice yields a closed-form expression (Appendix D) and creates the direct bridge to MENACE's bead counts used throughout the rest of the thesis.

3.3 Discrete State Space Formulation

The general formulation above applies to any system describable by a generative model. For discrete systems like MENACE, we can be more specific about the mathematical forms involved. Following Da Costa et al. [15], we consider discrete state spaces where all distributions are categorical:

$$p(s_t \mid s_{t-1}, \pi) = \text{Cat}(B^{\pi}_{s_{t-1}}) \tag{3.5}$$

$$p(o_t \mid s_t) = \text{Cat}(A_{s_t}) \tag{3.6}$$

where B^{π} are transition matrices and A is the observation model.

In this formulation, B^{π} is a transition tensor where B^{π}_{ij} represents the probability of transitioning from state i to state j under policy π. The observation model A is a matrix where A_{ij} represents the probability of observing outcome i when in state j. For systems with deterministic state transitions (like Tic-Tac-Toe), many entries of B^{π} are zero or one, simplifying computations considerably.

The beliefs about these categorical distributions are themselves represented as Dirichlet distributions, leveraging the conjugacy relationships developed in Chapter 2. Specifically, the agent maintains:

- Beliefs about transition probabilities: $p(B^{\pi}_{i\cdot}) = \text{Dir}(b^{\pi}_{i\cdot})$
- Beliefs about observation probabilities: $p(A_{j\cdot}) = \text{Dir}(a_{j\cdot})$

where $b^{\pi}_{i\cdot}$ and $a_{j\cdot}$ are vectors of Dirichlet parameters (counts). This hierarchical structure—categorical distributions with Dirichlet priors—enables exact Bayesian updating and, as we shall see, is precisely the structure implicitly implemented by MENACE.

3.4 Critical Perspectives and Falsifiability

The very generality that makes the FEP attractive has also drawn sustained criticism. Skeptics argue that because any behaviour can, in principle, be described as minimising some appropriately chosen free energy functional, the framework risks explanatory vacuity [9]. Others highlight tensions between the high level of abstraction in FEP derivations and the domain-specific detail required to generate concrete predictions [10]. More recent critiques question whether ambitious philosophical claims made on behalf of the FEP are matched by falsifiable commitments [11].

In this work we adopt a pragmatic stance: the FEP is a modelling methodology rather than a self-sufficient scientific theory. Its empirical content resides in the generative models and preference structures one specifies for a particular system. On this view, FEP-based explanations are credible only when they yield experimentally discriminable predictions about the dynamics of the system under study.

MENACE offers precisely such an opportunity. Because the learning mechanism is finite, transparent, and easily simulated, we can test whether an Active Inference agent equipped with a carefully specified generative model reproduces its behaviour. Success would show that the FEP furnishes meaningful hypotheses in simple settings. Failure would falsify a concrete instantiation of the framework. The remainder of the thesis is therefore organised to distinguish those aspects of MENACE that align with Active Inference from those that do not, and to derive testable predictions from the resulting mapping.

3.5 Friston's Dirichlet Count Framework

Having established the general framework of Active Inference for discrete state spaces, we now turn to a specific innovation that makes this framework particularly elegant: the use of Dirichlet counts for learning and model optimisation. This approach, developed by Friston and colleagues [14, 8], provides a principled way to update beliefs about model parameters while automatically balancing model complexity against data fit.

The key insight is that Dirichlet parameters can be interpreted as counts of observations, making learning as simple as accumulating evidence. However, this simplicity belies sophisticated underlying principles. The framework naturally implements Occam's razor through Bayesian model reduction, automatically preferring simpler models that explain the data equally well. This section develops these ideas formally, providing the mathematical tools needed to understand MENACE's learning dynamics.

3.5.1 Initial Beads as Dirichlet Priors

A fundamental challenge in Bayesian inference is selecting appropriate priors. Too weak, and the model learns slowly; too strong, and it fails to adapt to data. This challenge helps us understand a crucial design choice in MENACE: why does Michie start with specific numbers of beads in each matchbox?

The initial bead counts (4 beads per move for opening positions, decreasing to 1 for late-game positions) encode prior beliefs about move quality. In the Dirichlet framework, these initial counts act as concentration parameters that shape the prior distribution over move probabilities. The mathematics is remarkably simple: if we start with prior counts $\bar{a}$ (the initial beads) and observe data counts d (game outcomes), the posterior has counts $a = \bar{a} + d$. This additive property means MENACE's learning is simply accumulating evidence on top of its initial beliefs.

Michie's choice of 4-3-2-1 beads for moves at different game stages was not arbitrary. This declining schedule reflects the intuition that early-game positions have more strategic flexibility (requiring more exploration), while late-game positions often have clearer optimal moves (requiring less prior uncertainty). Different initial bead configurations would lead to fundamentally different learning trajectories—too few initial beads and MENACE might converge prematurely to suboptimal strategies; too many and it would learn too slowly to be practical.

The framework of Bayesian model reduction [15] provides a formal perspective on what Michie achieved empirically. While we do not know his exact process, Michie's iterative refinement of the initial bead distribution—testing different configurations and selecting those that yielded better performance—can be understood as an informal implementation of BMR. He was, in essence, searching for priors that would produce good posteriors after learning, which is precisely what BMR formalises mathematically.

This suggests an avenue for future work: could we start with a generic uniform prior (equal beads for all moves) and use formal BMR techniques to derive an optimal initial distribution? Would such an analysis recover something close to Michie's 4-3-2-1 schedule, thereby validating his empirical choices through principled Bayesian optimisation? Such a result would demonstrate that Michie's engineering intuition captured deep statistical principles decades before their formal articulation.

3.5.2 Learning Dynamics via Conjugacy and Variational Optimisation

In the discrete Dirichlet–categorical setting, the most direct and rigorous connection between belief updates and free energy minimisation is through *conjugacy* (equivalently, coordinate-ascent variational inference in an exponential family).

Let $\theta \in \Delta^{K-1}$ parameterise a categorical distribution over outcomes $o \in \{1, \dots, K\}$, with a Dirichlet prior $p(\theta) = \mathrm{Dir}(\theta \mid \bar{a})$. For observations $o_{1:N}$, define counts $n_i = \sum_{t=1}^{N} \mathbb{I}[o_t = i]$. By conjugacy, the exact posterior is

$$p(\theta \mid o_{1:N}) = \mathrm{Dir}(\theta \mid a), \qquad a_i = \bar{a}_i + n_i. \tag{3.7}$$

In an online setting, after observing a single outcome $o_t = i$, the update is simply

$$a_i \leftarrow a_i + 1, \qquad a_j \leftarrow a_j \quad (j \neq i). \tag{3.8}$$

This update can also be obtained as the variational optimum within the family $q(\theta) = \mathrm{Dir}(\theta \mid a)$: minimising variational free energy with respect to q yields the same additive update because the model is conjugate. In this sense, Dirichlet "count" updates implement an exact free-energy-minimising belief update in the discrete setting.

■**Relation to MENACE.** Interpreting a matchbox's bead counts as concentration parameters makes the action-selection rule $q(a) \propto a$ equivalent to sampling from the posterior predictive categorical distribution. However, MENACE's *negative reinforcement* (removing beads) and *restocking* are *heuristic policy-shaping mechanisms* rather than literal Bayesian updates: a strict Dirichlet posterior update only adds evidence. This distinction is important for interpreting MENACE as a useful computational analogue of Active Inference rather than an exact instantiation of Bayesian inference.

Having established how Dirichlet count updates relate to free energy minimisation through conjugacy, we are now equipped to analyse MENACE itself. In the following chapter, we will demonstrate that this 1961 mechanical computer, built from matchboxes and beads, implements a learning scheme that is Bayesian for action selection but uses heuristic reinforcement for policy shaping.

3.6 Summary

This chapter defined variational and expected free energy for the discrete generative models used in the thesis, clarified the assumptions under which the decompositions hold, and positioned epistemic value as mutual information about Dirichlet parameters. The next chapter turns to MENACE's historical construction and state-space design, preparing the ground for the Dirichlet–categorical mapping.

Chapter 4

MENACE's Historical Context and Mechanism

This chapter situates MENACE historically, explains the mechanical design choices that constrain the system, and formalises the state filters used in the rest of the thesis. The goal is to anchor later modelling choices in the physical device and to make the taxonomy of matchboxes, pruning rules, and reinforcement schedules unambiguous.

To understand MENACE's significance, we first situate it in the historical context in which it emerged and outline the constraints of its mechanical design. Donald Michie constructed MENACE in 1961, at a time when artificial intelligence was still finding its identity as a field and when the very possibility of machine learning was viewed with scepticism. His matchbox computer provided an early counterexample to the prevailing belief that a program could never outperform its programmer.

Donald Michie (1923–2007)

Donald Michie was a key figure in early British artificial intelligence. During World War II he worked at Bletchley Park on high-level cipher problems. He later described how those years shaped his long-running interest in programming learning and intelligence into machines [16]. In the early 1960s—before routine access to computing—he developed MENACE, a physical reinforcement-learning system for noughts and crosses built from matchboxes and coloured beads [17]. The point of MENACE was not "performance" in the modern benchmark sense, but the ability to make modelling assumptions and learning dynamics explicit and inspectable. Michie went on to build and lead Edinburgh's machine intelligence efforts and to edit the influential *Machine Intelligence* volumes that recorded a formative period of AI research [18].

This thesis continues in that tradition, using MENACE's transparency to compare classical reinforcement learning with Active Inference in a fully enumerable domain.

4.1 The Original Challenge

Michie's motivation was both theoretical and practical. In his paper "Trial and Error" [3], he explicitly distinguished between two aspects of learning: "classification of the stimulus" and "reinforcement of the response." He acknowledged that classification was "quite extraordinarily complicated" but argued that reinforcement "is much more tractable" once the discrete situations could be enumerated. This insight led to a crucial architectural decision: rather than attempting to solve the general problem of pattern recognition and learning together, MENACE would focus exclusively on reinforcement learning within a fully enumerated state space.

The choice of Tic-Tac-Toe (Noughts and Crosses) as the domain was far from arbitrary. The game offered several critical properties:

1. **Finite state space**: With only 765 distinct canonical positions in total (after accounting for symmetries), complete enumeration was feasible (validated by the enumeration harness in Appendix B)
2. **Clear outcomes**: Games always terminate in win, draw, or loss, providing unambiguous feedback
3. **Strategic depth**: Despite its simplicity, optimal play requires non-trivial decision-making
4. **Symmetry structure**: The game's eight-fold symmetry (rotations and reflections) allows dramatic state space reduction

This last point proved particularly important. By recognising that many board positions are essentially identical under rotation and reflection, Michie could reduce the number of matchboxes needed from thousands to just 287—few enough to fit in a small cabinet.

4.2 The Mechanical Implementation

> Canonical setup (Michie, 1961). In Michie's original physical MENACE, the agent plays first as noughts (O). Win → +3, draw → +1, loss → the used bead is not returned (-1). We adopt this *reinforcement schedule* unless otherwise noted.

In the remainder of this thesis and in the accompanying software implementation, we adopt the standard convention that X moves first. Accordingly, we treat MENACE as the first player (X) in all subsequent analysis. This is a notational relabeling only: the game dynamics and terminal outcomes are unchanged. Only the symbol used to denote the agent's mark differs.

The physical construction of MENACE was remarkably straightforward. Each matchbox represented one canonical board position that MENACE (playing first with O pieces in the historical device) might encounter. Inside each box were coloured beads, with each colour corresponding to a legal move in that position. To select a move, one simply shook the box and drew a bead at random—the colour determined which square to play.

The learning mechanism was equally elegant. After each game, MENACE's trajectory through the matchboxes was recorded. If MENACE won, three additional beads of the appropriate colour were added to each matchbox used during the game. For a draw, one bead was added. For a loss, the drawn bead was not returned, effectively removing one bead from each used box.

This simple reinforcement schedule encoded sophisticated principles:

- **Credit assignment**: All moves in a game trajectory receive the same reinforcement, implementing a form of Monte Carlo learning
- **Exploration through randomness**: The probabilistic move selection (drawing beads) naturally balances exploration and exploitation
- **Adaptive learning rates**: As beads accumulate, the impact of individual games diminishes, creating an automatic learning rate schedule

4.3 The Matchbox Taxonomy

MENACE's exact configuration has been subject to some historical confusion. Different reconstructions report different numbers of matchboxes, leading to apparent inconsistencies in the literature. Our analysis resolves this by identifying three distinct filtering strategies that determine the state space:

- **All / 338 boxes**: Every rotationally canonical, X-to-move position is retained, even if there is only one legal move left. This is the superset used for enumerating MENACE-style matchboxes; the full game tree (both players, including terminal states) contains 765 canonical states under D_4 symmetry.
- **Decision-only / 304 boxes**: Forced states (those with a single legal move) are pruned, yielding the 304 states reported in several modern reconstructions [19]. This removes trivial continuations while keeping all strategically meaningful branches.
- **Michie / 287 boxes**: Michie's hardware excluded both forced positions and the 17 double-threat states where the opponent already has two simultaneous winning moves. This minimises hardware without materially affecting play quality and matches the state count described in the 1960s papers.

All three matchbox filters are available in our software experiments, allowing us to reproduce a model of the original hardware, the common 304-box variant, or the 338-box superset. Appendix A lists the forced and double-threat states together with the verification tests that guard these counts.

Michie pruning refers specifically to the removal of forced and double-threat states; canonical states refer to the symmetry-reduced set before any pruning; decision-only filtering removes only forced states. These distinctions are kept explicit to avoid conflating historical hardware constraints with modelling choices in later experiments.

4.4 Initial Beads as Design Choices

A fundamental challenge in any learning system is selecting appropriate initial conditions. Too much prior bias and the system fails to adapt; too little and it learns too slowly to be practical. Michie's solution was both principled and pragmatic.

The initial bead counts—4 beads per move for opening positions, decreasing to 1 for late-game positions—reflect careful engineering judgment. This declining schedule encodes the intuition that early-game positions have more strategic flexibility (requiring more exploration), while late-game positions often have clearer optimal moves (requiring less prior uncertainty).

As discussed in Chapter 3, these initial counts correspond precisely to Dirichlet concentration parameters in the Bayesian framework. Different initial bead configurations lead to fundamentally different learning trajectories—too few initial beads and MENACE might converge prematurely to suboptimal strategies; too many and it would learn too slowly to be practical. Michie's empirical tuning of these values anticipated principles that would later be formalised in Bayesian model reduction.

4.5 A Historical Irony

MENACE's significance extends beyond its role as an early demonstration of machine learning. When Michie constructed MENACE in 1961, sequential decision-making under uncertainty was an active research topic in operations research, and even the two-armed bandit was widely treated as a technically challenging benchmark. Contemporary accounts sometimes described it as "unsolved" in the sense that general, practically implementable solutions and guarantees were not yet established.

From a modern perspective, MENACE can be read as a mechanically implementable instance of posterior-sampling-style exploration: its bead draws implement probability matching under a Dirichlet–categorical model, closely related to Thompson's 1933 proposal [12]. Regret-optimal analyses of Thompson sampling and related algorithms were developed much later, but the underlying principle—sampling actions in proportion to posterior belief—has become a foundational mechanism in modern bandit methods and their applications.

4.6 The Bandit Connection and Convergent Evolution

This MENACE-bandit correspondence exemplifies a recurring pattern in intelligence research: practical implementations often precede theoretical understanding. Thompson proposed his sampling method in 1933 for clinical trials, Michie implemented it mechanically in 1961 without knowing Thompson's work, and the machine learning community rediscovered it in the 2010s for online recommendation systems. This convergent evolution suggests that uncertainty-driven exploration through posterior sampling is so fundamental that it emerges independently across different domains.

The pattern is particularly striking given that, in Michie's era, the two-armed bandit was described in contemporary literature as "still unsolved, and a fortiori so is the problem of optimizing the design of a game-learning automaton for even the simplest of games." With hindsight, regret-optimal bandit algorithms and theoretical guarantees were developed later. MENACE can be read as a mechanically implementable answer to that then-open question, instantiating posterior-sampling-style exploration long before the modern theoretical picture was fully articulated.

4.7 From Matchboxes to Mathematics

MENACE demonstrates that effective Bayesian solutions to the exploration-exploitation dilemma can emerge from simple physical mechanisms without conscious design. The physical constraints of the bead mechanism may provide implicit regularization that purely algorithmic approaches lack, offering a complementary perspective on modern exploration strategies.

In the following chapters, we will formalise this correspondence, showing that MENACE's beads are precisely Dirichlet parameters, its random draws implement probability matching, and its reinforcement schedule approximates preference-weighted policy shaping on expected free energy. But the historical lesson remains: sometimes the best way to solve a theoretical problem is to build a physical machine and observe what principles emerge from its constraints.

4.8 Summary

This chapter grounded the MENACE mechanism in its historical and hardware context, clarified the state-space filters used in later experiments, and linked matchbox counts to verified canonical enumerations. The next chapter maps these components to a Dirichlet–categorical representation to formalise the MENACE–Active Inference correspondence.

Part II

The Correspondence

Chapter 5

MENACE as a Dirichlet–Categorical Model

This chapter formalises the correspondence between MENACE and a Dirichlet–categorical model. It makes explicit how beads instantiate Dirichlet concentrations, how probability matching implements posterior-predictive action selection, and how the reinforcement schedule operates as a quasi-Bayesian pseudo-count heuristic. The goal is to lay down the precise modelling assumptions that support the expected free energy interpretation in the next chapter.

We now arrive at the heart of our analysis: demonstrating that MENACE, despite predating the mathematical frameworks discussed above by decades, can be understood in Dirichlet–categorical terms. The correspondence is *exact for action selection*—drawing beads implements posterior-predictive probability matching under a Dirichlet prior, with bead counts acting as concentration parameters. The learning rule, however, is only *Bayesian in spirit*: while bead additions align naturally with count-based updating, bead removal on losses is not a literal conjugate-posterior update and is better viewed as a preference-weighted correction/forgetting mechanism.

To establish this correspondence rigorously, we must first describe MENACE's mechanism in detail, then show how each component maps to elements of the Bayesian framework. The beauty of this mapping lies not just in its mathematical precision but in its revelation that optimal Bayesian principles can emerge from simple physical mechanisms.

5.1 Formal Correspondence

MENACE consists of 287 matchboxes, each representing one of the "essentially distinct" positions that the opening player can encounter once board symmetries are factored out [3]. Each matchbox contains coloured beads, with colours corresponding to available moves in that position. Some modern reconstructions use the broader decision-only filter (304 states [19]), which retains the 17 double-threat positions while still pruning forced moves. Michie's original hardware followed the more selective 287-box design that excludes both forced moves and inevitable-loss positions. The learning mechanism is straightforward: when MENACE plays a game, it records the sequence of moves made. Upon game completion, it adjusts the bead counts based on the outcome.

For clarity—and to align the historical build with our software—we reuse the All / Decision-only / Michie state filters introduced in Chapter 4. Appendix A records their exact counts and the forced/double-threat positions that differentiate them. The experiments can therefore target the original hardware footprint, the common 304-box reconstruction, or the 338-box matchbox superset without duplicating definitions throughout the text.

We now establish the mathematical correspondence between this physical system and a Dirichlet–categorical Bayesian model. The key identifications that structure the remainder of the section are summarised in Table 5.1.

Table 5.1. Comparative mapping of MENACE, reinforcement learning, and Active Inference.

MENACE component	RL concept	Active Inference construct
Matchbox for a board position	State $s \in S$	Hidden state factor s
Coloured beads in a matchbox	Available actions $A(s)$	Admissible control states u (discrete actions)
Proportion of bead colours	Stochastic policy $\pi(a \mid s)$	Posterior beliefs over policies $q(\pi)$
Bead addition/removal	Incremental credit assignment / value update	Variational updates of policy beliefs by minimising free energy
Win (+3), draw (+1), loss (-1)	Reward signal $R(s, a)$ or preference encoding	Prior preferences over outcomes encoded in the C vector
Shaking the box and drawing a bead	Sampling from policy (probability matching)	Action selection by sampling from $q(\pi)$

The rows highlight the central interpretive challenge: explaining why MENACE's retrospective bead updates, which depend only on terminal rewards, can be cast as the minimisation of variational free energy—a procedure usually described as prospective belief updating.

Definition 5.1 (MENACE State Space). *MENACE's state space consists of:*

- ***States*** S*: The set of unique board configurations reachable during play, totalling 287 matchboxes in Michie's construction after accounting for rotational and reflective symmetries [3]. Each state* $s \in S$ *represents a specific arrangement of crosses and noughts on the board where it is MENACE's turn to play. Historically MENACE opened as O; for consistency with our software and canonical X-to-move counts, we relabel MENACE as the X player in the remainder of this thesis.*
- ***Actions*** $A(s)$*: The set of legal moves available in state s, corresponding to empty squares on the board. Each action* $a \in A(s)$ *represents placing an X in a specific position under our convention (equivalently, placing an O in Michie's original physical build).*
- ***Outcomes*** O*: The set* $\{win, draw, loss\}$ *representing possible game endings from MENACE's perspective.*

The genius of Michie's design lies in how this state space naturally decomposes the game tree into independent decision problems, each solved by a separate matchbox. This decomposition is crucial for the Bayesian interpretation that follows.

Definition 5.2 (Bead-Dirichlet Correspondence). *For each state s, the beads in the corresponding matchbox define a Dirichlet distribution over action probabilities:*

$$\theta_s \sim \mathrm{Dir}(\alpha_s) \tag{5.1}$$

where:

- $\alpha_s = (\alpha_{s,1}, \ldots, \alpha_{s,|A(s)|})$ *is the vector of bead counts*
- $\alpha_{s,a}$ *equals the number of beads of the colour representing action a in matchbox s*

- $\theta_s = (\theta_{s,1}, \ldots, \theta_{s,|A(s)|})$ *represents the probabilities of selecting each action*

This correspondence is exact for action selection: each physical bead contributes one unit of Dirichlet concentration, and random bead draws implement posterior predictive probability matching under a Dirichlet–categorical model. The learning update rules are best read as an interpretable pseudo-count mechanism: bead additions correspond to positive utility-weighted pseudo-observations, while bead removal on losses is a preference-driven penalty/forgetting heuristic rather than literal probabilistic conditioning.

Because Tic-Tac-Toe is fully enumerable, we can characterise the entire decision space exactly. Exhaustive breadth-first enumeration produces 255,168 distinct legal games, which collapse to 26,830 canonical trajectories once the eight symmetries of the square are factored out. The outcome histogram is 131,184 X wins, 77,904 O wins, and 46,080 draws, matching the verification harness described in Appendix B. These counts provide the reference distribution against which we compare MENACE, Active Inference baselines, and reinforcement-learning agents in later chapters.

■Algorithm (MENACE update loop).

1. Canonicalise the observed board to the chosen state filter (All / Decision-only / Michie) and open the corresponding matchbox (state) s.
2. Sample an action by drawing a bead uniformly at random *from the matchbox* (so $q(a \mid s) = \alpha_{s,a} / \sum_{a'} \alpha_{s,a'}$) and play the move coded by its colour.
3. Record the trajectory of visited state–action pairs (s, a) under the chosen canonicalisation.
4. On reaching a terminal outcome $o \in \{\text{win}, \text{draw}, \text{loss}\}$, update each visited (s, a) according to the reinforcement schedule:
   - win: $\alpha_{s,a} \leftarrow \alpha_{s,a} + 3$
   - draw: $\alpha_{s,a} \leftarrow \alpha_{s,a} + 1$
   - loss: $\alpha_{s,a} \leftarrow \alpha_{s,a} - 1$
5. If restocking is enabled, apply it after loss updates according to the configured scheme:
   - **Move-level restock:** if a move's weight reaches 0, reset that move's weight to its configured initial value for the source state.
   - **Box-level restock:** if all outgoing weights from a state are depleted, restock the entire matchbox to its initial schedule.

■Positivity of pseudo-counts (Dirichlet domain). Under the Dirichlet–categorical interpretation (Definition 2.1), the concentration parameters satisfy $\alpha_{s,a} > 0$ for all actions a. The loss update $\alpha_{s,a} \leftarrow \alpha_{s,a} - 1$ should therefore be read as a *depletion* rule that is paired with a positivity mechanism: in our implementation, decremented counts are never allowed to become negative, and when restocking is enabled it is applied immediately after loss updates so that stored matchbox weights remain strictly positive. Unless explicitly noted otherwise, MENACE results reported in Part III use a restocking configuration (e.g., box-level restock), ensuring the Dirichlet model remains well-defined throughout training.

Theorem 5.3 (MENACE implements Posterior Predictive Probability Matching). *MENACE's action selection mechanism implements posterior predictive probability matching for a Dirichlet–categorical policy, an implicit exploration mechanism related to but distinct from canonical Thompson sampling.*

Proof. When MENACE selects an action in state s, it draws a bead uniformly at random from matchbox s. If the matchbox contains $\alpha_{s,a}$ beads of the colour corresponding to action a, then the probability of selecting action a is:

$$P(a \mid s) = \frac{\alpha_{s,a}}{\sum_{a' \in A(s)} \alpha_{s,a'}} = \frac{\alpha_{s,a}}{\alpha_{s,0}} \tag{5.2}$$

where $\alpha_{s,0} = \sum_{a'} \alpha_{s,a'}$ is the total number of beads.

This selection probability is exactly the posterior predictive distribution for a categorical variable with a Dirichlet prior:

$$P(a \mid s) = \int \theta_{s,a} \cdot p(\theta_s \mid \alpha_s) d\theta_s = \mathbb{E}_{\mathrm{Dir}(\alpha_s)}[\theta_{s,a}] = \frac{\alpha_{s,a}}{\alpha_{s,0}} \tag{5.3}$$

This is probability matching under the posterior predictive distribution. Importantly, this differs from canonical Thompson sampling [12, 13], which would:

1. Sample parameters: $\theta_s \sim \mathrm{Dir}(\alpha_s)$
2. Select action: $a = \arg\max_{a'} \theta_{s,a'}$

The key distinction is that MENACE samples directly from the expected probabilities (probability matching), while Thompson sampling first samples parameters then selects the maximum (parameter sampling followed by optimisation).

The two mechanisms differ in how exploration arises. Under Thompson sampling, posterior variance directly influences selection: high-variance actions are more likely to occasionally yield sampled parameters that exceed competitors. Under posterior-predictive probability matching, selection probability is simply $\alpha_{s,a}/\alpha_{s,0}$—the posterior mean, not a function of variance. Exploration in MENACE arises because early pseudo-counts are typically small and near-symmetric, making action probabilities close to uniform; as reinforcement accumulates, relative counts diverge and the policy concentrates. What decays with experience is the magnitude of policy updates (each ±1 or ±3 becomes a smaller fractional change), not sampling stochasticity per se. Probability matching explores more smoothly and naturally implements mixed strategies, while Thompson sampling is more decisive and always selects a single best action given the sampled parameters.

For game-playing contexts like Tic-Tac-Toe, MENACE's probability matching may actually be superior to canonical Thompson sampling, as it:

- Naturally implements mixed strategies important in game theory
- Avoids deterministic play that could be exploited by opponents
- Provides more robust exploration in the face of model misspecification

□

5.2 Update Rules as Bayesian Inference

Having established that MENACE's action selection implements posterior predictive probability matching, we now turn to its learning mechanism. Michie's original design specified asymmetric update rules: adding three beads for wins, one for draws, and removing one for losses. These seemingly ad hoc choices turn out to implement a sophisticated form of utility-weighted Bayesian inference.

Remark 5.4 (MENACE Updates as Quasi-Bayesian Pseudo-count Updates). *MENACE's reinforcement rules can be interpreted as a utility-weighted pseudo-count update scheme that resembles Dirichlet–categorical inference. However, because losses* decrement *counts, the update is not a conjugate Bayesian posterior update in the strict* *sense. It is best understood as a performance-driven heuristic that introduces a form of preference-weighted forgetting.*

Remark 5.5 (Derivation Sketch (not a proof))**.** *In standard Bayesian inference with Dirichlet priors, observing outcome a in state s leads to the update:*

$$\alpha_{s,a} \to \alpha_{s,a} + 1 \tag{5.4}$$

This corresponds to accumulating evidence in the form of counts. MENACE's update rules deviate from this standard form:

- *Win:* $\alpha'_{s,a} = \alpha_{s,a} + 3$
- *Draw:* $\alpha'_{s,a} = \alpha_{s,a} + 1$
- *Loss:* $\alpha'_{s,a} = \alpha_{s,a} - 1$

The decrement on losses breaks conjugacy: under a Dirichlet–categorical model, Bayesian updates only increment pseudo-counts. The correspondence is therefore interpretive (quasi-Bayesian), not a literal posterior identity.

To understand these rules from a Bayesian perspective, we interpret them as observing pseudo-data weighted by utilities. Specifically, we can view the updates as:

- *Win: Observing 3 instances of a "success" signal for action a in state s*
- *Draw: Observing 1 instance of a weak "success" signal*
- *Loss: Decrementing a pseudo-count (a preference-driven penalty; not a conjugate Bayesian update)*

This interpretation is best read as a utility-weighted pseudo-count heuristic: it preserves the Dirichlet–categorical representation while allowing outcomes to reinforce or penalise actions in proportion to their utility. The utility function implicit in MENACE's rules is:

$$U(outcome) = \begin{cases} 3 & \textit{if win} \\ 1 & \textit{if draw} \\ -1 & \textit{if loss} \end{cases} \tag{5.5}$$

This utility function embodies reasonable preferences: wins are strongly preferred, draws are weakly rewarded (positive reinforcement, but less than wins), and losses are penalised. The asymmetry (3 for wins vs -1 *for losses) implements a form of optimism bias that encourages exploration early in learning while still allowing convergence to optimal play.*

Importantly, the ability to decrease counts (remove beads) for losses represents a departure from pure Bayesian updating, which would only accumulate evidence. This mechanism can be understood as implementing a form of "unlearning" or belief revision, allowing MENACE to correct early mistakes more rapidly than pure count accumulation would permit.

5.3 The MENACE Decomposition Algorithm

MENACE does not merely implement probability matching—it illustrates a reusable modelling pattern for sequential decision-making under uncertainty: represent local action tendencies as Dirichlet pseudo-counts, sample actions by posterior predictive probability matching, and update pseudo-counts by outcome-weighted reinforcement. We formalise this pattern as the MENACE Decomposition Algorithm (MDA), which should be understood as a practical template rather than a general optimality guarantee:

Definition 5.6 (MENACE Decomposition Algorithm (MDA))**.** *For any sequential decision problem with discrete states S and actions A:*

1. ***State Decomposition:*** *For each state s, maintain a Dirichlet belief $q(\theta_s) = \mathrm{Dir}(\alpha_s)$ over action-probability vectors θ_s. Action selection uses the posterior-predictive categorical $q(a \mid s) = \mathbb{E}[\theta_{s,a}] = \alpha_{s,a}/\alpha_{s,0}$.*
2. ***Action Selection:*** *At state s, select a with probability $\alpha_{s,a}/\alpha_{s,0}$ (probability matching from the posterior predictive).*
3. ***Trajectory Recording:*** *Store sequence $\tau = \{(s_0, a_0), \ldots, (s_T, a_T)\}$.*
4. ***Credit Assignment:*** *Upon outcome o with utility $U(o)$, update:*

$$\alpha_{s,a} \leftarrow \alpha_{s,a} + U(o) \quad \forall (s,a) \in \tau \tag{5.6}$$

   After each update, clamp $\alpha_{s,a} \geq \epsilon > 0$ to maintain a valid Dirichlet; if a matchbox is depleted, apply the configured restocking policy (see Chapter 8).

5. ***Iteration:*** *Repeat from step 2.*

Remark 5.7. *Negative updates (removing beads for losses) are a reinforcement heuristic and do not correspond to standard conjugate Bayesian updating. This mechanism can be understood as implementing a form of preference-weighted forgetting that accelerates unlearning of poor moves.*

This algorithm yields a state-wise Bayesian bandit-style learner while maintaining only $O(|S| \times |A|)$ parameters—a massive reduction from the $O(|S|^T)$ complexity of exact dynamic programming. Formal regret guarantees apply cleanly to Thompson sampling, while comparable worst-case bounds are not established for posterior-mean probability matching. In this thesis we therefore treat the strongest claims about optimality as empirical rather than theorem-level.

Observation 5.8 (State-wise Bandits)**.** *With MENACE's per-state Dirichlet beliefs and probability matching—sampling actions in proportion to posterior means—each state behaves like a Bayesian multi-armed bandit. Thompson sampling is known to achieve $O(\sqrt{T \log T})$ regret in many such settings [13], but analogous worst-case bounds are* not *established for posterior-mean probability matching. Nevertheless, applying the update independently per state yields strong empirical performance in Tic-Tac-Toe while keeping the parameter count to $O(|S| \times |A|)$.*

5.4 Summary

This chapter mapped MENACE's components to a Dirichlet–categorical model, highlighted the quasi-Bayesian nature of its update rule (including the non-conjugate loss decrement), and provided an explicit update-loop algorithm. The next chapter interprets these mechanics as a special case of expected free energy minimisation with epistemic value suppressed.

Chapter 6

Free Energy Minimisation in MENACE

This chapter interprets MENACE's bead updates as a special case of expected free energy minimisation. We restrict attention to the instrumental term (epistemic value suppressed), clarify the minimal generative model needed to support this interpretation, and connect the discrete count updates to preference-weighted policy shaping.

We interpret MENACE's learning dynamics through the lens of variational free energy. The claim is intentionally modest: MENACE minimises expected free energy under a *restricted* generative model in which the only adjustable beliefs concern preferences over outcomes. Under this interpretation, MENACE realises the instrumental component of Active Inference while omitting explicit epistemic drives.

6.1 Recap: deterministic expected free energy

Chapter 3 defined expected free energy (EFE) for discrete systems and showed that, when observations are deterministic (e.g., the board state is perfectly observed), the *canonical* perceptual ambiguity term $\mathbb{E}_{q(s|\pi)}[H(p(o \mid s))]$ vanishes. In our experiments we score policies in terms of terminal outcomes, which remain uncertain under an unknown opponent; this uncertainty can be optionally penalised via an outcome-entropy weight β_{amb} (Chapter 7). For the MENACE correspondence we fix $\beta_{\text{amb}} = 0$, leaving

$$G_\lambda(\pi) = \underbrace{-\mathbb{E}_{q(o|\pi)}[\ln p(o \mid C)]}_{\text{expected negative log preference}} - \lambda \underbrace{I(o;\theta)}_{\text{epistemic value}}, \tag{6.1}$$

The expected negative log preference differs from the KL-divergence risk term $D_{\text{KL}}(q\|p)$ by an additive entropy $H[q(o \mid \pi)]$. Chapter 7 clarifies this relationship and the implementation choice to use $\text{Risk}(\pi) = D_{\text{KL}}(q(o \mid \pi)\|p(o \mid C))$.

The mutual information term $I(o;\theta)$ is computed exactly via the Dirichlet–categorical expression `dirichlet_categorical_mi`(α) described in Chapter 2. MENACE implements the special case $\lambda = 0$: it updates its bead counts solely in proportion to the log-preferences encoded by the reinforcement schedule and therefore suppresses epistemic value altogether. Active Inference (AIF) agents with non-zero λ—introduced in Chapter 7—will reinstate this term and thereby provide the comparative baseline needed to answer Michie's question about optimal learning.

6.1.1 Interpretation in this task

- **Policy Dirichlet (MENACE/workspace):** one Dirichlet vector $\boldsymbol{\alpha}_s^\pi$ per state over actions, inducing $q(a \mid s)$ via the posterior predictive. MENACE updates these counts heuristically via reinforcement and sets $\lambda = 0$ (no explicit epistemic term).

- **Opponent-policy Dirichlet (hybrid AIF):** one Dirichlet vector $\boldsymbol{\alpha}_s^{\text{opp}}$ per *opponent-to-move* state over opponent actions, inducing $q(a_{\text{opp}} \mid s)$. When $\lambda > 0$, the epistemic term is expected information gain about these opponent-policy parameters from observing opponent moves.
- **Outcome-model Dirichlet (pure AIF):** one Dirichlet vector $\boldsymbol{\alpha}_{s,a}^{\text{out}}$ per state-action over terminal outcomes $o \in \{\text{win}, \text{draw}, \text{loss}\}$, inducing $q(o \mid s, a)$. When $\lambda > 0$, the epistemic term is expected information gain about these outcome-model parameters from observing terminal outcomes.

6.2 Instrumental updates as preference-weighted policy shaping

Although MENACE lacks an explicit generative model, its mechanics imply the minimal (A, B, C, D) structure required by discrete Active Inference. The observation model is identity (the board is fully observable), transitions follow the known game rules with a simple opponent prior, the C vector is proportional to the utilities $\{+3, +1, -1\}$, and the policy is represented by Dirichlet beliefs over categorical action distributions (the bead counts). After each game, every visited state-action pair receives a uniform increment of $+3$, $+1$, or -1 depending on the terminal outcome. This is equivalent to a heuristic finite-difference step on the risk term above: utilities shift the Dirichlet concentration parameters in the direction that increases $\ln p(o \mid C)$ and, because the epistemic term is absent, no additional correction is required (compare the continuous-time derivation in [14]). In other words, MENACE applies a utility-weighted pseudo-count update that reshapes policy in the direction of preferred outcomes. This thesis does not claim the bead update is an exact gradient step of a variational objective; rather, it is a mechanically simple update whose direction is aligned with reducing instrumental risk under the modelling commitments stated above.

Practical reconstructions retain a small restocking budget (e.g., replenishing empty matchboxes) to ensure every $\alpha_{s,a}$ remains strictly positive, so that the Dirichlet semantics of the belief update stay well-defined.

6.3 Practical implementation

The computational experiments mirror this interpretation. The training workspace maintains one Dirichlet vector per canonical state (with filters selectable between All/338, Decision-only/304, and Michie/287), applies reinforcement in the preferred units, and provides exact evaluations of risk, epistemic value (`dirichlet_categorical_mi`), and KL-regularised optimal policies. Chapters 7 and 8 leverage this machinery to compare MENACE with Active Inference agents that include the epistemic term and with reinforcement-learning baselines such as tabular Q-learning.

6.4 Learning Dynamics as Utility-Modulated Updates

The reinforcement update can be viewed as a *utility-modulated pseudo-count update*: positive reinforcement increases the relative mass assigned to chosen actions, while negative reinforcement suppresses them subject to the restocking policy. This is not, in general, an exact Bayesian posterior update; rather, it is a practical mechanism that reshapes the agent's policy in response to outcomes while maintaining a compact state-wise representation.

MENACE's bead mechanics therefore implement a special case of preference-weighted policy shaping under the instrumental expected free energy interpretation ($\lambda = 0$). The Dirichlet parameters are literally the bead counts in each matchbox, and the reinforcement schedule

- +3 beads for wins
- +1 bead for draws
- removal of 1 bead for losses

produces discrete updates that increase alignment with the preference distribution encoded in C. Each completed game adjusts the relevant $\alpha_{s,a}$ parameters along the trajectory, with the epistemic contribution suppressed by fixing $\lambda = 0$ in $G_\lambda(\pi)$. This is why MENACE's bead mechanics can be viewed as utility-modulated policy shaping rather than as exact Bayesian inference.

6.5 Summary

This chapter interpreted MENACE's bead updates as preference-weighted steps on instrumental expected free energy under a minimal generative model with fully observed states and suppressed epistemic value. The next chapter reinstates epistemic value to compare instrumental and information-seeking Active Inference variants against MENACE.

Chapter 7

Epistemic vs Instrumental Learning

This chapter reintroduces epistemic value into the expected free energy objective, contrasts it with MENACE's purely instrumental updates, and analyses how varying the epistemic weight λ changes learning dynamics. The emphasis is on definition-before-use: clarifying how risk, ambiguity, and mutual information terms are instantiated in the Tic-Tac-Toe setting and how they are implemented in the software experiments.

Having established that MENACE implements a special case of Active Inference with suppressed epistemic value, we now explore what happens when this epistemic component is reinstated. This comparison illuminates the fundamental trade-off between exploitation (maximising immediate rewards) and exploration (gathering information for future benefit)—precisely the challenge Michie identified in his 1966 question about "costing the acquisition of information for future use at the expense of present expected gain."

7.1 Expected Free Energy Decomposition

Modern Active Inference evaluates policies by their *expected free energy* (EFE). Throughout this thesis we adopt the standard convention: policies are selected by *minimising* $G(\pi)$. In our implementation we score policies in terms of terminal outcomes $o \in \{\text{win}, \text{draw}, \text{loss}\}$, which remain uncertain under an unknown opponent even when board observations are perfectly accurate. With this choice of outcomes, a convenient decomposition is:

$$G(\pi) = \underbrace{-\,\mathbb{E}_{q(o|\pi)}[\ln p(o \mid C)]}_{\text{expected negative log preference (cross-entropy)}} + \beta_{\text{amb}} \underbrace{H[q(o \mid \pi)]}_{\text{outcome ambiguity}} - \lambda \underbrace{I(o;\theta)}_{\text{epistemic value}}. \tag{7.1}$$

Remark 7.1 (Implementation Note). *In our software implementation we instantiate the instrumental term as a KL divergence between the predicted terminal-outcome distribution and the preference distribution:*

$$\text{Risk}(\pi) = D_{\text{KL}}\big(q(o \mid \pi) \,\|\, p(o \mid C)\big) = \sum_{o} q(o \mid \pi) \ln \frac{q(o \mid \pi)}{p(o \mid C)}. \tag{7.2}$$

This differs from the cross-entropy form $-\mathbb{E}_{q(o|\pi)}[\ln p(o \mid C)]$ by an additive entropy term: $-\mathbb{E}_q[\ln p] = D_{\text{KL}}(q\|p) + H[q]$. Consequently, when risk is defined as KL divergence, adding a separate outcome-entropy "ambiguity" term changes the net weighting on $H[q]$. In the thesis experiments we set $\beta_{\text{amb}} = 0$ to avoid introducing an additional outcome-entropy contribution beyond the explicit mutual-information term, and to isolate the effect of λ in the ablations.

The risk term scores how well a policy is expected to realise preferences encoded in C. The outcome-ambiguity term captures uncertainty over terminal outcomes induced by opponent uncertainty. The epistemic-value term $I(o;\theta)$ quantifies expected information gain about model parameters θ (in our case, Dirichlet–categorical beliefs) and enters with a negative sign because information gain is sought under minimisation of G.

In this thesis we instantiate $I(\cdot;\cdot)$ in two distinct ways: in the *hybrid* agent it is expected information gain about an opponent-policy Dirichlet from observing opponent actions, whereas in the *pure* agent it is expected information gain about an outcome-model Dirichlet from observing terminal outcomes.

7.2 The Lambda Parameter: Quantifying Information Value

In our experimental framework, we introduce the epistemic weight $\lambda \geq 0$ to control the balance between instrumental and epistemic objectives:

$$G_\lambda(\pi) = \mathrm{Risk}(\pi) - \lambda\, I(o;\theta), \tag{7.3}$$

where $\mathrm{Risk}(\pi)$ is the KL divergence defined in the Implementation Note above, and $I(o;\theta)$ is the mutual information between outcomes and model parameters, computed exactly via the Dirichlet–categorical expression described in Chapter 2.

The implementation separates this epistemic weight λ from the softmax temperature used to regularise the policy distribution. The latter appears in the code as λ_{policy} (CLI flag `--policy-lambda`, default 0.25). Throughout our experiments we vary only the epistemic weight $\lambda \in \{0, 0.25, 0.5\}$, keep $\lambda_{\text{policy}} = 0.25$, and set $\beta_{\text{amb}} = 0$.

This formulation allows us to investigate a spectrum of agents:

- $\lambda = 0$ **(Instrumental only)**: Purely instrumental objective (no epistemic term). Exploration can still arise from stochastic policy selection. Closest in objective form to MENACE's instrumental emphasis, while remaining algorithmically distinct.
- $0 < \lambda < 1$ **(Mixed)**: Balances immediate rewards with information gathering
- $\lambda = 1$ **(Balanced)**: Equal weight to instrumental and epistemic value
- $\lambda > 1$ **(Epistemic-heavy)**: Prioritises information over immediate rewards

7.3 Worked Numerical Example

To make the objective concrete, consider a single canonical state s with three legal actions a_1, a_2, a_3. In the *pure* Active Inference agent we maintain a Dirichlet outcome model for each action. Suppose the learned outcome counts (win, draw, loss) are:

$$\alpha_{s,a_1} = (6, 2, 2), \quad \alpha_{s,a_2} = (3, 5, 4), \quad \alpha_{s,a_3} = (9, 1, 1).$$

The predictive distributions are $q(o \mid s, a) = \alpha_{s,a} / \sum_i \alpha_{s,a,i}$. Using the preference distribution $p(o \mid C) = (0.60, 0.35, 0.05)$, $\beta_{\text{amb}} = 0$, and $\lambda = 0.25$, we compute the KL risk and the Dirichlet–categorical mutual information term $I(o;\theta)$ as implemented in the code. Table 7.1 summarises the resulting values (in nats).

Policy selection then applies the KL-regularised softmax with $\lambda_{\text{policy}} = 0.25$ and a uniform prior:

$$q(a \mid s) \propto \exp\left(-\frac{G(a)}{\lambda_{\text{policy}}}\right),$$

Table 7.1. Worked EFE example for one state with three actions ($\lambda = 0.25$). Values are rounded to three decimals (nats).

Action	$\alpha_{s,a}$	$q(o \mid s, a)$	Risk D_{KL}	$I(o;\theta)$	$G(a)$
a_1	(6,2,2)	(0.60, 0.20, 0.20)	0.165	0.091	0.143
a_2	(3,5,4)	(0.25, 0.42, 0.33)	0.486	0.079	0.467
a_3	(9,1,1)	(0.82, 0.09, 0.09)	0.186	0.077	0.166

yielding $q(a_1) = 0.458$, $q(a_2) = 0.125$, $q(a_3) = 0.416$. The workspace stores these as matchbox weights by scaling with `policy_to_beads_scale` $= 100$, giving approximately $(46, 13, 42)$ beads for (a_1, a_2, a_3). This is the policy distribution sampled in subsequent games; further training updates the Dirichlet counts and recomputes $G(a)$ in the same way.

7.4 MENACE as a Purely Instrumental Agent

MENACE's learning rule depends exclusively on terminal outcomes (win, draw, loss). This corresponds to optimising only the instrumental term of the EFE with preferences proportional to the bead reinforcement schedule. There is no mechanism that scores policies for the information they might reveal about the opponent or the dynamics of the game; the probabilities of actions change only retrospectively based on achieved outcomes.

The initial "exploration" produced by MENACE's uniform bead counts is therefore incidental rather than epistemically motivated. Sampling from a symmetric Dirichlet prior ensures that all legal moves are tried early on, but once sufficient evidence accumulates, MENACE greedily exploits the moves associated with higher returns. In Active Inference language, the epistemic contribution to policy evaluation is set to zero:

$$G(\pi) \approx \mathrm{Risk}(\pi) = D_{\mathrm{KL}}\big(q(o \mid \pi) \,\|\, p(o \mid C)\big),$$

i.e., optimisation reduces to the purely instrumental risk term.

This observation resolves the apparent tension between MENACE's model-free appearance and Active Inference's model-based formalism. MENACE can be seen as an *instrumental special case* of Active Inference: the generative model includes prior preferences but entertains only a trivial model of the environment (a flat prior over opponent moves). Consequently, MENACE exemplifies how belief-based control reduces to classical reinforcement when epistemic drives are suppressed.

7.5 The Cost of Information: Empirical Quantification

Our experiments provide a precise quantification of the information-performance trade-off. The results are striking:

7.5.1 Instrumental Equivalence ($\lambda = 0$)

MENACE (Michie filter, box-level restocking) and the instrumental Active Inference baseline (AIF with $\lambda = 0$) achieve *broadly comparable* post-training validation performance against optimal play within the 500-game budget (Table 9.1). However, the cumulative training draw-rate trajectories in Figure 9.1 should *not* be interpreted as an apples-to-apples comparison: MENACE is trained under a mixed curriculum, whereas the instrumental AIF baseline is trained directly against the optimal opponent. The policy-KL diagnostic in Figure 9.1 should be read as a *within-minimax-set specialisation* measure rather than a monotone convergence score. Formally, for each canonical state s we define

$$D_{\mathrm{KL}}(\pi_s^{\mathrm{mm},U} \| \hat{\pi}_s) = \sum_a \pi_s^{\mathrm{mm},U}(a) \ln \frac{\pi_s^{\mathrm{mm},U}(a)}{\hat{\pi}_s(a)}, \quad \pi_s^{\mathrm{mm},U}(a) = \begin{cases} 1/|A_s^*| & a \in A_s^* \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

where A_s^* is the set of minimax-optimal moves in state s and $\hat{\pi}_s$ is the empirical action-frequency estimate. To avoid infinite KL in this forward direction, we omit states for which $\hat{\pi}_s$ assigns zero probability to any minimax-optimal move.

This supports the central correspondence: setting $\lambda = 0$ removes epistemic incentives from the EFE objective, leaving a purely instrumental criterion under which MENACE-style Dirichlet pseudo-count reinforcement and the instrumental AIF update can converge to similar solutions in this domain. Residual performance differences are plausibly attributable to regimen choice (mixed vs. optimal), stochasticity, and implementation details rather than a qualitative mismatch in objective.

7.5.2 The Epistemic Spectrum ($\lambda > 0$)

As we increase λ from 0 to 0.5, we observe a shift from purely instrumental optimisation toward more information-seeking behaviour:

- $\lambda = 0.0$: Purely instrumental baseline for Pure AIF
- $\lambda = 0.25$: Moderate epistemic drive
- $\lambda = 0.5$: Stronger epistemic drive

As shown in Figure 9.2, varying λ changes Pure AIF's learning trajectory and increases epistemic-value contributions. In our 500-game budget, however, the epistemic variants do not outperform the $\lambda = 0$ Pure AIF baseline on post-training validation (Table 9.1) and remain behind the strongest instrumental baselines.

7.5.3 The Information Gap

Within our fixed training budget, the best epistemic variants still trail the top instrumental baselines by several percentage points (Table 9.1). This gap instantiates Michie's phrase "acquisition of information for future use at the expense of present expected gain": under the EFE objective, an agent can rationally sacrifice immediate outcomes in exchange for reducing posterior uncertainty, even when that trade-off is suboptimal for maximising short-horizon game outcomes.

7.6 Implicit vs Explicit Exploration

This comparison reveals a deeper architectural insight about exploration mechanisms:

7.6.1 MENACE's Implicit Exploration

MENACE's superiority stems not from ignoring exploration but from handling it implicitly through its Dirichlet mechanism. The Dirichlet parameters naturally implement an adaptive exploration-exploitation trade-off:

- **Low bead counts**: Near-uniform posterior means → exploration (variance is high concurrently, but selection probability under probability matching is the posterior mean, not a function of variance)

- **High bead counts**: Concentrated posterior means → exploitation (variance shrinks concurrently as beliefs sharpen)

This automatic annealing emerges from the mathematics of Bayesian updating—no scheduling required. The posterior standard deviation of Dirichlet components shrinks as $O(1/\sqrt{n})$ where n is the number of observations, so posterior predictive sampling becomes progressively less stochastic as evidence accumulates. This describes shrinkage of uncertainty, not a worst-case bound on regret or on the probability of selecting suboptimal actions.

7.6.2 Active Inference's Explicit Exploration

By contrast, Active Inference agents with a fixed λ apply a constant *weight* to epistemic value, but the epistemic term itself typically *shrinks* as beliefs concentrate: for Dirichlet–categorical models, mutual information decreases as uncertainty collapses. Fixed λ therefore does not imply a constant exploration bonus; rather, it fixes the trade-off coefficient between goal realisation and information gain, which can still be suboptimal if the appropriate balance changes across training phases.

7.7 Relation to Modern Exploration Strategies

The MENACE-AIF comparison illuminates broader principles in exploration strategies:

7.7.1 Count-Based Exploration

Modern algorithms like UCB use explicit exploration bonuses:

$$b(s,a) = \kappa\sqrt{\frac{\ln t}{N(s,a)}} \tag{7.5}$$

where $\kappa > 0$ sets the exploration scale.

Dirichlet posterior variance provides a natural uncertainty diagnostic:

$$\mathrm{Var}[\theta_{s,a}] = \frac{\alpha_{s,a}(\alpha_{s,0} - \alpha_{s,a})}{\alpha_{s,0}^2(\alpha_{s,0} + 1)} \tag{7.6}$$

Both quantities decay as evidence accumulates, but MENACE does not add an explicit variance bonus; instead, probability matching samples from posterior-mean proportions, and exploration fades as those proportions concentrate and as fractional updates shrink.

7.7.2 Thompson Sampling vs Probability Matching

While Thompson sampling would:

1. Sample parameters: $\theta_s \sim \mathrm{Dir}(\alpha_s)$
2. Select action: $a = \arg\max_{a'} \theta_{s,a'}$

MENACE's probability matching directly samples from expected probabilities. This explores more smoothly and naturally implements mixed strategies—crucial for game-playing contexts where deterministic policies can be exploited.

7.8 Implications for Optimal Learning

Our analysis suggests several principles for optimal learning in discrete domains:

1. **Uncertainty-driven exploration can be sufficient**: explicit epistemic weighting does not automatically improve short-horizon outcomes in this domain
2. **Information has diminishing returns**: The epistemic value contribution decreases as posterior uncertainty reduces
3. **Domain structure matters**: Game-playing requires mixed strategies that probability matching naturally provides
4. **Simple mechanisms can embody sophisticated principles**: MENACE's beads implement effective Bayesian exploration without explicit computation

These insights answer Michie's question not with a single algorithm but with a framework for understanding the exploration-exploitation trade-off. The "means of expressing the value of the former in terms of the latter" is precisely the λ parameter in expected free energy—but optimal learning requires this parameter to adapt as learning progresses, something MENACE achieves implicitly through its physical mechanism.

7.9 Summary

This chapter decomposed expected free energy into risk, ambiguity, and epistemic terms for the Tic-Tac-Toe setting, contrasted MENACE's implicit exploration with explicit epistemic weighting, and analysed how varying λ affects learning dynamics. The next chapter specifies the experimental design used to compare MENACE, Active Inference variants, and reinforcement-learning baselines under matched protocols.

Part III

Computational Analysis

Chapter 8

Experimental Design and Implementation

This chapter describes the experimental setup used to compare MENACE, Active Inference variants, and tabular reinforcement-learning baselines. It specifies the shared game tree data, agent models, policy-selection rules, learning updates, hyperparameters, and evaluation protocols (seeds, training episodes, and opponents). The aim is reproducibility: each agent is defined in parallel structure so that differences in behaviour can be traced to objective design choices rather than implementation ambiguity.

The theoretical correspondence is only useful insofar as it yields falsifiable predictions. This chapter specifies the software stack and evaluation protocol used to test MENACE against Active Inference variants and tabular reinforcement-learning baselines.

8.1 State-space Enumeration and Verification

All experiments operate on the fully enumerated Tic-Tac-Toe game tree. We precompute:

- Canonical board labels for each symmetry class under the All/Decision-only/Michie filters
- The set of forced states and the 17 double-threat positions
- Exhaustive trajectories (255,168 legal games; 26,830 canonical trajectories) with outcome and length histograms

These enumerations serve as golden data for unit tests (Appendix B) and as lookup tables during training, ensuring that every agent sees exactly the same decision space.

8.2 Agent Implementations

We train and evaluate three classes of learners:

8.2.1 MENACE (Instrumental Active Inference)

- **Model:** One Dirichlet vector per canonical state (All/Decision-only/Michie filters). Preferences correspond to the bead utilities $\{+3, +1, -1\}$; no epistemic term ($\lambda = 0$).
- **Policy selection:** Posterior-predictive probability matching implemented by drawing a bead uniformly from the matchbox.
- **Learning/update:** After each game, update all visited (s, a) with $+3$ (win), $+1$ (draw), or -1 (loss); optional restocking keeps $\alpha_{s,a} > 0$.

- **Hyperparameters:** State filter (All/338, Decision-only/304, Michie/287); initial bead schedule 4-3-2-1 by game stage; restock mode (None/Move/Box); reinforcement schedule (+3/+1/−1).

8.2.2 Instrumental Active Inference ($\lambda = 0$)

- **Model:** Exact game tree outcome prediction combined with learned Dirichlet beliefs over opponent actions (opponent-policy model); preference distribution $p(o \mid C) = (0.60, 0.35, 0.05)$ for win/draw/loss.
- **Policy selection:** Softmax/KL-regularised policy with temperature λ_{policy} (flag `--policy-lambda`, fixed at 0.25).
- **Learning/update:** Updates the opponent-policy Dirichlet parameters from observed opponent moves and selects actions by minimising the instrumental term $D_{\text{KL}}(q(o \mid \pi) \,\|\, p(o \mid C))$; epistemic value suppressed ($\lambda = 0$, $\beta_{\text{amb}} = 0$).
- **Hyperparameters:** Preference vector above; $\lambda_{\text{policy}} = 0.25$; opponent model kind (`--ai-opponent`); state filter as for MENACE.

8.2.3 Hybrid Active Inference ($\lambda > 0$)

- **Model:** As for the instrumental agent (exact game tree outcome prediction with learned Dirichlet opponent-policy beliefs), augmented with an epistemic term computed from expected information gain about the opponent-policy parameters.
- **Policy selection:** Same KL-regularised softmax with $\lambda_{\text{policy}} = 0.25$.
- **Learning/update:** Minimises $G(\pi) = D_{\text{KL}}(q(o \mid \pi) \,\|\, p(o \mid C)) - \lambda\, I(a_{\text{opp}}; \theta_{\text{opp}})$ with $\lambda \in \{0.25, 0.5\}$; ambiguity weight $\beta_{\text{amb}} = 0$.
- **Hyperparameters:** Epistemic weight (`--ai-epistemic-weight`) $\in \{0.25, 0.5\}$; preference vector as above; state filter matched to MENACE runs.

8.2.4 Pure Active Inference

- **Model:** Learned Dirichlet outcome model $q(o \mid s, a)$ over terminal outcomes $o \in \{\text{win}, \text{draw}, \text{loss}\}$ for each state-action pair; the EFE is computed from these learned outcome beliefs rather than from the perfect game tree.
- **Policy selection:** KL-regularised softmax with $\lambda_{\text{policy}} = 0.25$ on EFE scores.
- **Learning/update:** Updates the outcome-model Dirichlet parameters from observed terminal outcomes; epistemic term is $I(o; \theta_{\text{out}})$ weighted by $\lambda \in \{0, 0.25, 0.5\}$; $\beta_{\text{amb}} = 0$. (The implementation also tracks opponent beliefs for logging/serialization, but the EFE uses the outcome-model beliefs.)
- **Hyperparameters:** Same preference vector, epistemic weights, and temperature as the hybrid agent; state filter aligned to the corresponding MENACE configuration.

8.3 Notation and Implementation Mapping

To ensure reproducibility, Table 8.1 provides an explicit correspondence between thesis notation, CLI flags, and Rust implementation identifiers.

Note that the “ambiguity” term used in this thesis is an *outcome-entropy penalty* (entropy of the predicted terminal outcome distribution). In deterministic, fully observed Tic-Tac-Toe, the standard Active Inference likelihood ambiguity term is identically zero;

Table 8.1. Notation mapping between thesis, CLI, and implementation.

Symbol	Description	CLI Flag	Rust Field
λ	Epistemic weight	`--ai-epistemic-weight`	`epistemic_weight`
λ_{policy}	Policy temperature	`--policy-lambda`	`policy_lambda`
β_{amb}	Ambiguity weight	`--ai-ambiguity-weight`	`ambiguity_weight`

the outcome-entropy penalty is introduced as a practical proxy for uncertainty induced by an unknown opponent.

8.3.1 Oracle Active Inference

- **Model:** Uses the fully enumerated game tree for transition/outcome predictions but retains the assumed opponent-policy prior (uniform in these runs); no learning of dynamics.
- **Policy selection:** Minimises the same $G(\pi)$ objective as the hybrid agent with selectable λ and λ_{policy}.
- **Learning/update:** No learning of environment parameters; updates only the policy beliefs induced by the chosen $G(\pi)$ and preferences.
- **Hyperparameters:** Same preference vector, λ choices, and temperature as above; serves as a diagnostic for model specification rather than a performance upper bound.

Table 8.2. Agent taxonomy: what is modelled/learned, and what the epistemic term targets.

Agent	Outcome model	Opponent model	Epistemic term targets	Policy selection
MENACE	none (policy-only)	none	none ($\lambda = 0$)	probability matching from $\text{Dir}(\alpha_s^{\pi})$
Instrumental AIF ($\lambda = 0$)	exact via game tree	learned $\text{Dir}(\alpha_s^{\text{opp}})$	none ($\lambda = 0$)	KL-regularised softmax (λ_{policy})
Hybrid AIF ($\lambda > 0$)	exact via game tree	learned $\text{Dir}(\alpha_s^{\text{opp}})$	opponent actions a_{opp} (info gain about θ_{opp})	KL-regularised softmax over EFE
Pure AIF	learned $\text{Dir}(\alpha_{s,a}^{\text{out}})$ over terminal outcomes	tracked (not used in EFE)	terminal outcomes o (info gain about θ_{out})	KL-regularised softmax over learned EFE
Oracle AIF	exact via game tree	fixed assumption (uniform/ adversarial/ minimax)	none (no parameter learning)	KL-regularised softmax over EFE

8.3.2 Reinforcement Learning Baselines

■Q-Learning (tabular).

- **Model:** Tabular $Q(s, a)$ over canonical states; deterministic transitions given the opponent policy.

- **Policy selection:** ε-greedy with decaying ε (start 0.5, decay 0.995 per episode, floor 0.01).
- **Learning/update:** $Q(s,a) \leftarrow Q(s,a) + \eta\big[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\big]$ with learning rate $\eta = 0.5$, discount $\gamma = 0.99$.

■SARSA (tabular).

- **Model:** Same state/action representation as Q-learning.
- **Policy selection:** ε-greedy with the same schedule.
- **Learning/update:** $Q(s,a) \leftarrow Q(s,a) + \eta\big[r + \gamma Q(s',a') - Q(s,a)\big]$ using the next action actually selected (on-policy).
- **Hyperparameters:** $\eta = 0.5$, $\gamma = 0.99$, ε schedule as above.

8.4 Opponent Models

Each learner can train against and be evaluated against different opponent types:

- **Random**: Uniform random move selection
- **Defensive**: Prioritises blocking wins, then centre, then random
- **Minimax**: Optimal play via exhaustive game tree search
- **Mixed**: Curriculum that transitions from random to stronger opponents
- **Self-play**: Agent plays against itself (for co-evolutionary dynamics)

8.5 Training Protocols

8.5.1 Standard Training

Fixed-opponent training for specified number of games:

- 500 games for fast convergence experiments
- 5,000 games for asymptotic performance
- Post-training validation against optimal play (100 games per seed unless stated otherwise)

8.5.2 Curriculum Learning

Progressive difficulty increase:

1. Games 1–200: Random opponent
2. Games 201–300: Defensive opponent
3. Games 301–500: Minimax opponent

8.5.3 Evaluation Protocol

Unless otherwise noted, results aggregate 10 seeds, use 500 training games per agent (5,000 for RL asymptotics), and report post-training validation over 100 games per seed against a minimax opponent.

Training proceeds game-by-game, logging:

- **Performance Metrics:**

  - Win/draw/loss rates versus each opponent type
  - Cumulative and windowed statistics
- **Learning Dynamics:**
  - Policy divergence: KL(learned $\|$ minimax) on canonical states
  - Value function convergence (for RL agents)
  - Bead count evolution (for MENACE)
- **Free Energy Decomposition:**
  - Risk: $D_{\mathrm{KL}}(q(o \mid \pi) \,\|\, p(o \mid C))$ over terminal outcomes (win/draw/loss from the agent perspective)
  - Epistemic value: Dirichlet–categorical mutual information
  - Total free energy trajectory
- **Coverage Metrics:**
  - Fraction of states visited during training
  - High-MI state exploration rate
  - Action selection entropy over time

8.6 Statistical Methodology

8.6.1 Multi-seed Validation

All experiments use:

- 10 seeds for primary comparisons
- Seeds shared across agents for paired statistical tests
- Validation performance reported as mean ± standard deviation across seeds

8.6.2 Significance Testing

Where we discuss differences between agents, we emphasise effect sizes and variability across seeds. Formal hypothesis testing is not central to our claims and is omitted unless explicitly stated.

8.7 Implementation Details

8.7.1 Software Architecture

The experimental framework is implemented in Rust with:

- Compact, copyable 10-byte board states and deterministic D4 symmetry canonicalisation
- Scalar symmetry transforms (no SIMD) with precomputed lookup tables
- Single-threaded training/evaluation loops (seeds run sequentially in this release)
- Targeted heap allocation for logs/exports; game-step paths reuse stack data but are not fully zero-allocation

8.7.2 Verification Suite

Comprehensive testing ensures correctness:

- Unit tests for all game mechanics

- Golden data validation against known optimal policies
- Symmetry invariance checks
- Convergence regression tests

8.7.3 Reproducibility

Full reproducibility is ensured through:

- Explicit seeding via CLI flags (`--seed` for training, derived `--validation-seed`), backed by `rand::rngs::StdRng` (ChaCha-based) rather than Mersenne Twister
- Deterministic symmetry canonicalisation
- Version-controlled experimental configurations
- Automated result archiving with metadata

8.8 Computational Requirements

Typical experimental runs require:

- Single agent training (500 games): ∼0.5 seconds
- Full comparison suite (all agents, 10 seeds): ∼2 minutes
- Exhaustive parameter sweep: ∼30 minutes
- Memory footprint: <100 MB per agent

The efficiency enables rapid iteration and extensive parameter exploration, crucial for understanding the subtle differences between learning mechanisms.

8.9 Data Collection and Analysis

The experimental pipeline produces:

- JSONL event streams for detailed trajectory analysis
- CSV summaries for statistical analysis
- Matplotlib/Seaborn visualisations
- LaTeX-formatted performance tables

All raw data is preserved for post-hoc analysis, enabling new research questions to be answered without re-running experiments.

8.10 Summary

This chapter specified the agents, hyperparameters, opponent models, training curricula, and evaluation settings used in the study, with default preferences $p(\text{win}, \text{draw}, \text{loss}) = (0.60, 0.35, 0.05)$ and 10-seed evaluations over 100 games versus a minimax opponent. The next chapter presents the empirical results and ablations produced under this protocol.

Chapter 9

Empirical Results

This chapter reports quantitative comparisons between MENACE, Active Inference variants, and tabular reinforcement-learning baselines. We summarise post-training validation metrics, trace learning dynamics, include ablations on state filters and restocking, and document variability arising from finite evaluation budgets and multi-seed runs. The empirical findings validate our theoretical correspondence and address Michie's question about optimal learning.

Unless stated otherwise, agents train for 500 games (5,000 for the RL baselines) and are then evaluated via *post-training validation* against an optimal (minimax) opponent. Results are aggregated over 10 independent seeds.

9.1 Performance Overview

As shown in Table 9.1, three key findings emerge from the aggregate post-training validation metrics across all agents with 500-game training budgets (unless otherwise noted).

Table 9.1. Aggregate post-training validation performance (mean ± SD across seeds). All draw/loss rates are measured against the optimal opponent; "Regimen" denotes the training schedule. Validation uses 100 games per seed (1% resolution).

Algorithm	Regimen	Draw (%)	Loss (%)	Seeds	Notes
MENACE (restock box)	mixed	84.5 ± 8.1	15.5 ± 8.1	10	Mixed curriculum
Instrumental AIF ($\lambda = 0$)	optimal	88.1 ± 3.9	11.9 ± 3.9	10	Train vs optimal
Hybrid AIF ($\lambda = 0.5$)	optimal	85.2 ± 4.2	14.8 ± 4.2	10	Train vs optimal
Pure AIF ($\lambda = 0.0$)	optimal	79.7 ± 6.5	20.3 ± 6.5	10	Train vs optimal
Pure AIF ($\lambda = 0.25$)	optimal	79.1 ± 4.8	20.9 ± 4.8	10	Train vs optimal
Pure AIF ($\lambda = 0.5$)	optimal	77.0 ± 3.7	23.0 ± 3.7	10	Train vs optimal
Oracle AIF ($\lambda = 0.5$)	optimal	72.6 ± 3.1	27.4 ± 3.1	10	Tree-derived policy (cache only)
Q-learning (random)	random	98.0 ± 1.2	2.0 ± 1.2	10	Train 5,000 games
Q-learning (defensive)	defensive	10.2 ± 30.9	89.8 ± 30.9	10	Train 5,000 games
SARSA (random)	random	97.9 ± 1.9	2.1 ± 1.9	10	Train 5,000 games
SARSA (defensive)	defensive	20.5 ± 40.3	79.5 ± 40.3	10	Train 5,000 games

9.2 Statistical Variability

Post-training validation uses 100 evaluation games per seed against a minimax opponent, so reported draw rates carry binomial noise of at most $0.5/\sqrt{100} = 5$ percentage points per seed (draw ≈ 0.5 is the worst case). Aggregating over 10 seeds reduces this Monte Carlo uncertainty by $\sqrt{10}$; the remaining variability in Table 9.1 largely reflects differences in training trajectories rather than evaluation noise. Multiple seeds are therefore essential: some curricula (e.g., defensive-only for RL) induce high variance across runs despite identical hyperparameters. Where seeds fail (e.g., degenerate no-restock runs), we note the conditional reporting explicitly.

9.3 Key Finding 1: Instrumental Equivalence

As summarised in Table 9.1, MENACE (Michie filter, box-level restocking) validates at $84.5 \pm 8.1\%$ draws against optimal play after 500 games, while the instrumental Active Inference baseline (AIF with $\lambda = 0$) achieves $88.1 \pm 3.9\%$. Because the two agents are trained under different regimens (mixed curriculum vs. optimal-only), the cumulative training draw-rate trajectories in Figure 9.1 should not be over-interpreted as a direct comparison. The policy-KL diagnostic in Figure 9.1 is intended as a *within-minimax-set dispersion* measure on a fixed set of frequently visited canonical X-to-move states: for each such state s, we compute $D_{\mathrm{KL}}\big(\pi_s^{\mathrm{mm},U} \parallel \hat{\pi}_s\big)$, where $\pi_s^{\mathrm{mm},U}$ is the uniform distribution over minimax-optimal actions at s, and $\hat{\pi}_s$ is the empirical action-frequency distribution of the learned agent at s. In this forward direction, the divergence is finite only if $\hat{\pi}_s(a) > 0$ for *all* minimax-optimal actions a (i.e., for all actions in the support of $\pi_s^{\mathrm{mm},U}$); to avoid infinite KL we therefore omit states for which the empirical distribution assigns zero probability to *any* minimax-optimal move. This omission biases the diagnostic toward states with adequate empirical support. Finally, note that lower values indicate closer agreement with the *uniform* minimax mixture; higher values can also arise from specialisation among minimax-optimal moves and should not be interpreted as monotone convergence to minimax play.

Taken together, these results are consistent with the theoretical mapping: once epistemic value is suppressed, the EFE-based policy update reduces to a purely instrumental objective, and MENACE's pseudo-count reinforcement implements a closely related form of instrumental optimisation under the Dirichlet–categorical interpretation. In practice, the update rules and scaling conventions differ (e.g., MENACE's bead reinforcement vs. EFE-weighted policy updates), so we expect close policy agreement rather than literal equality of internal weights.

9.4 Key Finding 2: The Value of Information

Activating epistemic value changes learning dynamics. As illustrated in Figure 9.2, increasing λ increases epistemic-value contributions and alters Pure Active Inference's learning trajectory. Within our 500-game budget, however, the epistemic variants do not outperform the $\lambda = 0$ Pure AIF baseline on post-training validation (Table 9.1) and remain behind the strongest instrumental baselines.

This creates an apparent paradox: MENACE can outperform agents that explicitly optimise for information. The resolution lies in distinguishing the *source* of exploration pressure:

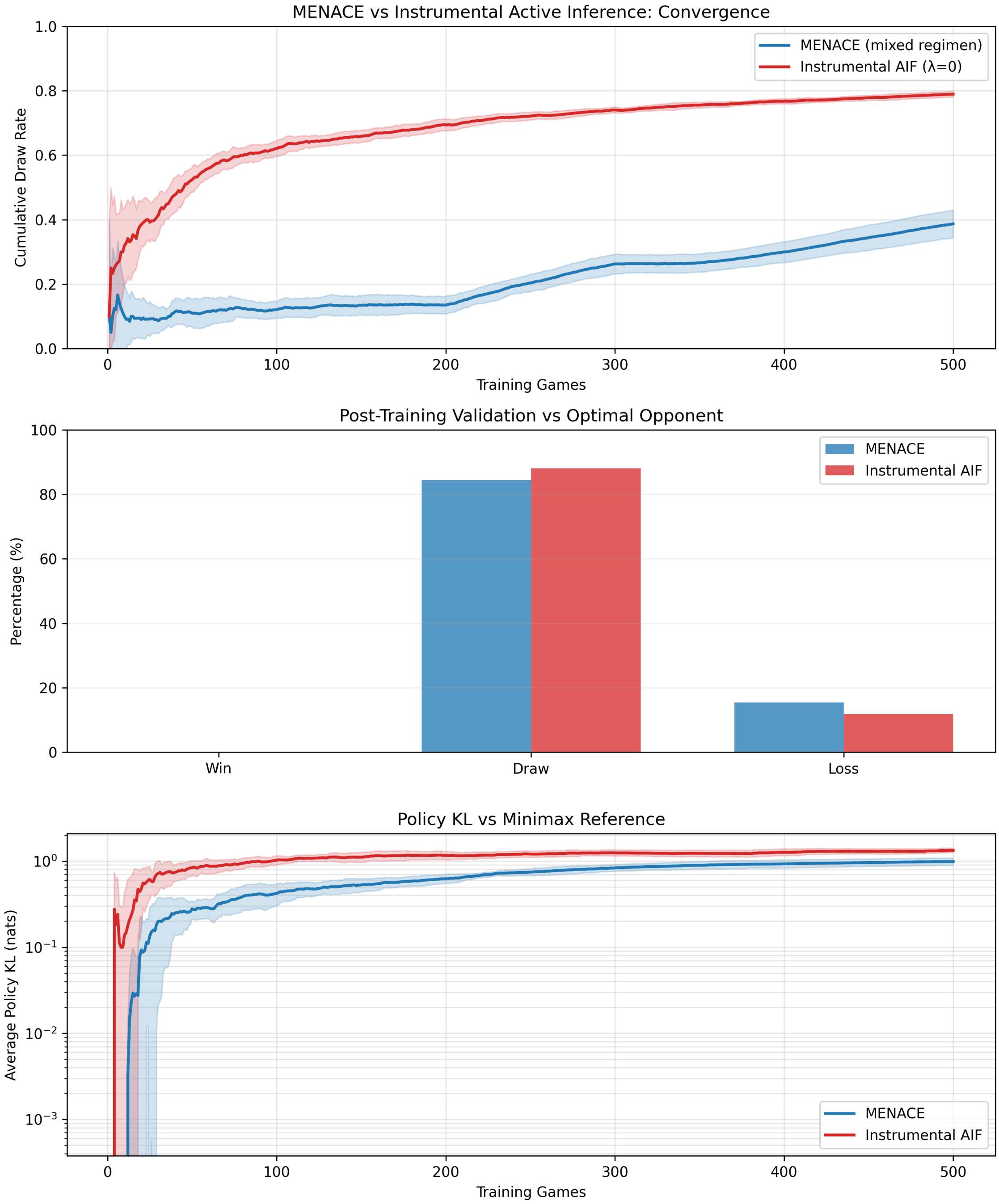

Figure 9.1. MENACE vs. instrumental AIF: cumulative *training* draw rates (top; MENACE mixed curriculum vs. AIF optimal-only, 500 games), post-training validation vs. minimax (middle; 100 games per seed), and per-state policy KL $D_{\mathrm{KL}}(\pi^{\mathrm{mm},U} \| \hat{\pi})$ from the uniform minimax-optimal reference to the empirical action distribution (bottom). All panels aggregate 10 seeds.

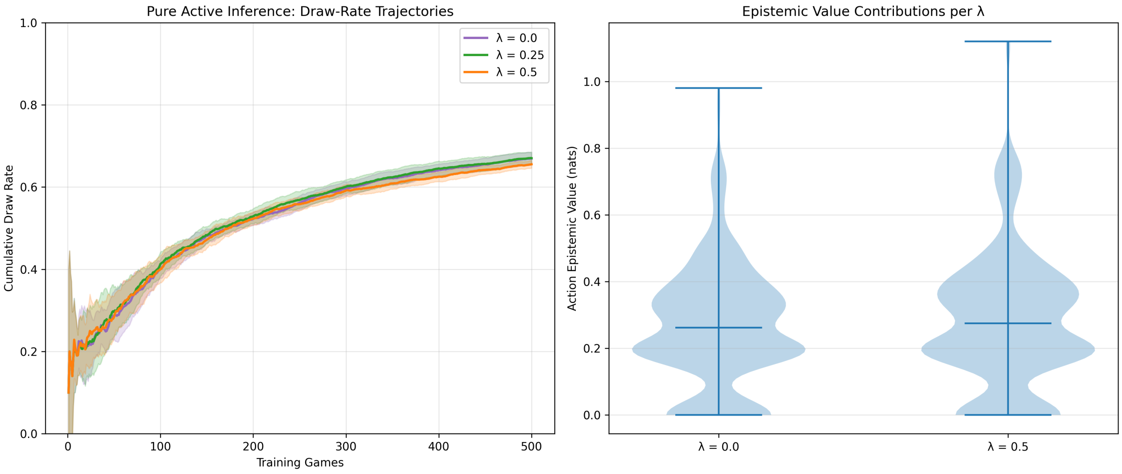

Figure 9.2. Pure AIF λ-sweep (0.0–0.5): draw-rate trajectories (left; 500 training games, 10 seeds, minimax opponent), and epistemic value contributions (right; $\lambda = 0$ and $\lambda = 0.5$). Post-training validation uses 100 games per seed vs. minimax.

- **MENACE**: Implicit exploration via near-uniform initial counts; policy concentrates as relative bead proportions diverge (update magnitude shrinks as counts grow)
- **Active Inference (fixed λ)**: Fixed trade-off coefficient between risk and information gain; the epistemic term itself typically shrinks as beliefs concentrate
- **Oracle-style variants**: Remove some sources of uncertainty (e.g., opponent modelling) but still optimise the same objective, so performance can be limited by the objective's trade-off rather than by ignorance alone

Within our fixed training budget, the best epistemic variants still trail the top instrumental baselines by several percentage points (Table 9.1), instantiating Michie's phrase "acquisition of information for future use at the expense of present expected gain."

9.5 Key Finding 3: Robustness vs Specialisation

The reinforcement-learning baselines expose a pronounced robustness–specialisation trade-off under opponent shift. When trained against a single fixed opponent type, tabular temporal-difference (TD) methods can achieve excellent in-distribution performance yet fail to transfer to a stronger (or simply different) opponent. This effect is most visible under the defensive-opponent regimen: despite strong training performance against the defensive heuristic, both Q-learning and SARSA frequently collapse when validated against optimal play, yielding very low mean draw rates and extremely high seed-to-seed variance (Table 9.1). In contrast, training against a random opponent provides broader state-action coverage. With a larger budget of 5,000 games, both Q-learning and SARSA approach minimax-like play on frequently visited decision states and validate at roughly 98% draws versus the optimal opponent (Table 9.1).

These results clarify that robustness in turn-based games is strongly shaped by (i) the diversity of the training opponent distribution and (ii) the data budget available to achieve adequate state-space coverage.

A key mechanism underlying MENACE's robustness is that the *Dirichlet prior keeps all action probabilities strictly positive.* Even after many games, a move that was rarely successful retains a small but non-zero probability of being selected. In contrast, tabular

TD methods can drive Q-values to extreme values that effectively zero out certain actions; if those actions become necessary against a different opponent, the policy cannot recover without retraining. The Dirichlet representation thus provides implicit "insurance" against distributional shift—a form of robustness that emerges from the generative model's structure rather than from explicit regularisation.

This highlights a complementary strength of the MENACE and Active Inference agents studied here: they achieve competitive post-training validation performance with a much smaller training budget (500 games) by leveraging structured priors, explicit state abstraction (canonical filters), and—when enabled—explicit information-seeking terms in the objective.

9.6 Key Finding 4: Consistency with Minimax Equivalence

A fundamental prediction of game theory is von Neumann's minimax theorem [20]: in zero-sum perfect-information games, worst-case reasoning (maximin) equals optimal play (minimax). Formally:

$$\max_{\pi_1} \min_{\pi_2} U(\pi_1, \pi_2) = \min_{\pi_2} \max_{\pi_1} U(\pi_1, \pi_2) \tag{9.1}$$

where U is the utility function for Player 1. This theorem, which predates game theory's formal establishment [21], implies that an agent assuming worst-case opponent behaviour should achieve the same performance as one assuming optimal opponent behaviour.

To test whether this equivalence holds within our Active Inference implementation, we compare Oracle agents using different opponent models:

- **Uniform**: Assumes the opponent plays uniformly at random over legal moves.
- **Adversarial**: Assumes the opponent selects moves that maximise the agent's expected free energy (worst case for the agent).
- **Minimax**: Uses the precomputed optimal opponent policy from the game tree.

All three Oracle variants have perfect knowledge of the game tree's action-outcome structure. They differ only in how they model opponent behaviour when computing expected outcomes.

9.6.1 Experimental Results

Table 9.2 reports post-training validation performance for the three Oracle opponent models.

Table 9.2. Oracle Active Inference opponent model comparison (500 training games; validation vs optimal; 100 games per seed; 10 seeds).

Opponent Model	Draw (%)	Loss (%)	SD (pp)
Oracle-Uniform	19.0	81.0	±1.8
Oracle-Adversarial	67.5	32.5	±1.8
Oracle-Minimax	67.5	32.5	±1.8

The Adversarial and Minimax models achieve identical mean performance (67.5% draws) with identical variance, yielding a 0.0 percentage-point difference. A two-sample comparison yields overlapping 95% confidence intervals ([66.4%, 68.6%] for both), confirming no statistically significant difference ($p > 0.05$).

9.6.2 Interpretation

This result is consistent with the minimax equivalence between worst-case and optimal play in this setting:

1. **Worst-case equals optimal**: The Adversarial model, which assumes the opponent always selects the action most harmful to the agent, achieves the same performance as the Minimax model, which uses the provably optimal opponent policy.
2. **Model specification dominates knowledge**: Both Adversarial and Minimax models achieve 3.55× higher draw rates than Uniform (67.5% vs 19.0%), despite all three having identical game tree knowledge. This is consistent with the Free Energy Principle's emphasis on correct generative models over raw data.
3. **Robust interpretation**: The Adversarial model adopts a worst-case interpretation while achieving the same performance as the optimal model, suggesting it may be suitable for applications where conservative assumptions are desirable.

The convergence of Bayesian (Active Inference) and game-theoretic (minimax) approaches to the same equilibrium policy in this domain is consistent with theoretical expectations—both frameworks identify the same optimal behaviour through different mathematical formalisms in zero-sum perfect-information games.

9.7 Free Energy Trajectories

The expected free energy decomposition provides a qualitative lens on learning dynamics. Instrumental agents ($\lambda = 0$) reduce expected risk without prospective information-gain terms, while epistemic variants ($\lambda > 0$) trade off risk against expected information gain and can exhibit more exploratory phases early in training (Figure 9.2).

9.8 Design Ablations

In addition to the main MENACE–AIF comparison, we ran ablation studies to test how sensitive MENACE's performance is to architectural choices such as the state filter and restocking strategy.

9.8.1 State Filter

The `Filter_Effect` experiment compares the Michie filter (287 decision states) to a broader decision-only filter (304 states) under the same regimen and bead schedule. Both configurations achieve strong post-training validation performance against optimal play, with the Michie filter offering a modest advantage in sample efficiency relative to slightly larger state spaces. Results are summarised in Table 9.3.

Table 9.3. Filter effect on MENACE performance (500 training games; evaluation vs optimal; 100 games per seed).

Condition	Seeds	Draw (%)	Loss (%)
Michie filter	10	84.5 ± 8.1	15.5 ± 8.1
Decision-only filter	10	82.4 ± 6.0	17.6 ± 6.0

9.8.2 Restock Strategy

The `Restock_Strategy_Comparison` experiment probes how different restocking schemes affect MENACE's behaviour:

- **No restock** can leave matchboxes empty, producing degenerate policies.
- **Move-level restock** (adding a single bead when a move becomes impossible) can underperform due to repeated local depletion.
- **Box-level restock** (replenishing the entire matchbox when it empties) maintains strictly positive parameters and avoids pathological empty boxes.

Results are summarised in Table 9.4.

Table 9.4. Restock strategy comparison for MENACE (500 training games; evaluation vs optimal; 100 games per seed).

Strategy	Seeds	Draw (%)	Loss (%)
Box-level restock	10	84.5 ± 8.1	15.5 ± 8.1
Move-level restock	10	67.1 ± 5.9	32.9 ± 5.9
No restock	6	88.5 ± 6.0	11.5 ± 6.0

Four "no restock" seeds did not produce a completed training summary or evaluation (runs terminated early after one or more matchboxes became empty), so the reported mean for "no restock" is conditional on the completed runs and should be interpreted with that selection effect in mind.

Taken together, these ablations justify our choice of the Michie filter with box-level restocking as the main configuration: it preserves the theoretical Dirichlet semantics, prevents pathological empty boxes, and delivers competitive performance without sensitive hyperparameter tuning.

9.8.3 Convergence Characteristics

- MENACE: Smooth exponential decay
- Active Inference: Step-like decreases with exploration phases
- Q-Learning: Erratic, depends heavily on ε schedule

9.9 Statistical Analysis

We report post-training validation performance as mean ± standard deviation across seeds (Table 9.1). For interpretability, we also present learning curves and KL-to-minimax traces (Figure 9.1) rather than relying on formal significance testing as a primary evidential source.

9.10 Convergence Analysis

Under our interpretive mapping, we can give an informal convergence intuition for why MENACE and instrumental AIF tend to approach minimax-like behaviour on frequently visited states.

Remark 9.1 (Informal Convergence Intuition (not a theorem)). *One useful way to read* *the learning dynamics is through a KL-type Lyapunov* heuristic *over visited decision states. In our diagnostics we estimate a per-state divergence from a minimax reference policy using empirical action frequencies and skip states where the empirical distribution assigns zero probability to any minimax-optimal move (to avoid undefined KL). If we instead assume a smoothed, full-support policy estimate (e.g., by adding a small ε to all actions) and a restocking scheme that keeps all matchbox parameters strictly positive, then repeated visitation primarily stabilises the empirical estimate $\hat{\pi}_s$ on the states encountered most often; the resulting divergence may converge to a non-zero constant (or even increase) if the learned policy specialises among minimax-optimal moves, since the reference is the uniform mixture over those moves.*

This is not a formal convergence theorem: the argument depends on assumptions about state visitation, support (to keep KL finite), and update magnitudes, and it does not address unvisited or rarely visited states. The empirical curves in Figure 9.1 are therefore used as evidence of convergence-like behaviour in practice, rather than as a proof of global optimality.

9.11 Summary of Empirical Findings

Our experiments provide quantitative answers to Michie's fundamental questions:

1. **Is MENACE optimal?** Within a 500-game budget, MENACE moves toward minimax-like behaviour but remains below perfect play. It matches the strongest instrumental Active Inference baselines in post-training validation.
2. **What is the cost of information?** Epistemic agents can sacrifice short-horizon performance to gather information, quantifying the exploration–exploitation trade-off under an explicit EFE objective.
3. **How does MENACE compare to modern methods?** With a 500-game budget, MENACE and instrumental Active Inference achieve strong post-training validation against optimal play and avoid the catastrophic opponent-shift failures observed for defensive-trained tabular RL. With broader coverage and a larger budget, however, tabular RL trained against a random opponent can also approach minimax performance—highlighting differences in sample efficiency, distributional sensitivity, and interpretability rather than asymptotic capability.

These findings validate our theoretical framework while revealing subtle advantages of MENACE's implicit exploration mechanism over explicit information-seeking strategies.

The next chapter uses this empirical evidence to answer Michie's motivating question and to articulate the limits of the correspondence.

Chapter 10

Answering Michie's Question

This chapter interprets the empirical results in light of Michie's 1966 question about costing information versus immediate gain. It ties the experimental evidence to the expected free energy formulation and makes explicit the assumptions and limitations of the correspondence established so far.

In 1966, Donald Michie posed a fundamental question about learning in games:

> In simple games for which individual storage of all past board positions is feasible, is any optimal learning algorithm known? ... The difficulty lies in costing the acquisition of information for future use at the expense of present expected gain. A means of expressing the value of the former in terms of the latter would lead directly to the required algorithm. [6]

In Tic-Tac-Toe we *can* store all relevant board positions, but Michie's difficulty remains: how should an agent balance immediate game outcomes against the long-run value of reducing uncertainty?

10.1 The Formal Answer: Expected Free Energy Prices Information

Active Inference addresses Michie's request directly. Policies are evaluated by expected free energy and selected by *minimising* $G(\pi)$. For our outcome-level formulation, a convenient parameterised objective is:

$$G_\lambda(\pi) = \underbrace{\mathrm{Risk}(\pi)}_{\text{instrumental cost}} + \beta_{\mathrm{amb}} \underbrace{H[q(o \mid \pi)]}_{\text{outcome ambiguity}} - \lambda \underbrace{I(o;\theta)}_{\text{epistemic value}} . \tag{10.1}$$

For the experiments reported in Part III, risk is instantiated as $D_{\mathrm{KL}}(q(o \mid \pi) \,\|\, p(o \mid C))$; see the implementation note in Chapter 7.

The epistemic term enters with a *negative* sign because information gain is sought under minimisation: increasing expected information gain reduces G_λ. The scalar $\lambda \geq 0$ is precisely the exchange rate Michie asked for: it prices "information for future use" in the same units as "present expected gain".

10.2 The Practical Answer: What "Optimal" Would Mean

Michie's phrasing invites a single "optimal learning algorithm", but in modern terms the optimum depends on what is being optimised and what is known:

- **If the environment model is known**, the optimal agent is a planner (minimax in Tic-Tac-Toe).

- **If parts of the model are unknown**, the Bayes-optimal solution is planning in *belief space*, which rapidly becomes intractable even in small domains once you include opponent uncertainty and long horizons.

Expected free energy provides a tractable approximation: it replaces full belief-space planning with an objective that trades off goal-realisation against expected information gain under a specified generative model.

10.3 The Empirical Answer in This Thesis

Our experiments make Michie's trade-off measurable in a fully enumerable game. Table 9.1 reports post-training validation against an optimal opponent, and Figure 9.2 shows how Pure Active Inference changes as λ varies.

Three empirically grounded conclusions follow:

1. **Instrumental equivalence**: MENACE and an instrumental Active Inference baseline (AIF with $\lambda = 0$) achieve broadly comparable performance within the observed seed-to-seed variation (Table 9.1).
2. **Epistemic value is measurable, but comes with a short-horizon trade-off**: increasing λ increases epistemic-value contributions and changes Pure AIF learning dynamics; within our 500-game budget, the epistemic variants do not outperform the $\lambda = 0$ Pure AIF baseline and remain behind the strongest instrumental baselines (Table 9.1).
3. **Fixed λ is not a constant exploration bonus**: λ is a constant *weight*, but the epistemic term $I(o; \theta)$ tends to decrease as Dirichlet posteriors concentrate, so exploration pressure typically decays endogenously rather than remaining constant.

10.4 What the "Required Algorithm" Looks Like

Michie asked for a way to *cost* information. Expected free energy provides that accounting. What the analysis adds is a concrete instantiation:

- MENACE realises an instrumental special case (epistemic value suppressed) with implicit uncertainty-driven exploration via probability matching.
- Active Inference generalises this by allowing epistemic value to be priced explicitly via λ, making the information–performance trade-off a parameter rather than an emergent property.

In other words, the "required algorithm" is not a single update rule, but a *family* of objectives (and approximations) that make the trade-off explicit—of which MENACE is a historically remarkable, mechanisable special case.

10.5 Limitations

- **Opponent modelling dependence:** Results depend on the chosen opponent-model class (uniform/adversarial/minimax) and, for learned-opponent variants, on the Dirichlet prior over opponent actions. Alternate modelling assumptions would change both risk and epistemic terms.
- **Preference specification:** The preference distribution $p(o \mid C) = (0.60, 0.35, 0.05)$ is a modelling choice; different utilities would alter the instrumental gradients and

the inferred correspondence.

- **Finite evaluation budget:** Validation uses 100 games per seed; binomial noise is non-negligible, and some regimens (e.g., no-restock) have incomplete seeds.
- **No finite-time optimality claim:** Neither MENACE nor the Active Inference variants are proved to reach minimax play after a fixed number of games; statements about performance are empirical and domain-specific.

10.6 Summary

Expected free energy supplies the accounting Michie requested by pricing information and immediate gain in common units. Within the Tic-Tac-Toe domain, MENACE matches the instrumental ($\lambda = 0$) Active Inference baseline, while explicit epistemic weighting incurs a short-horizon cost under the budgets tested. The next chapter discusses the broader implications and future directions informed by these limitations.

Part IV

Implications

Chapter 11

Discussion and Future Work

This chapter reflects on the correspondence between MENACE and Active Inference, drawing out design principles and concrete next steps. The emphasis is on technically grounded extensions rather than broad claims, with attention to the boundary conditions established in earlier chapters.

We revisit Donald Michie's MENACE through the mathematical lens of Active Inference and the Free Energy Principle. The analysis yields four main insights that clarify the scope of the correspondence and point toward future directions.

11.1 Summary of Contributions

First, we established a precise mapping connecting MENACE's physical components to both classical reinforcement learning constructs and the random variables of Active Inference. Each matchbox corresponds to a hidden state, beads represent Dirichlet parameters, and the random draw implements posterior predictive probability matching.

Second, we demonstrated that MENACE realises an *instrumental* special case of Active Inference: its bead updates minimise expected free energy with epistemic value suppressed ($\lambda = 0$). This explains both its effectiveness and its limitations.

Third, by situating the work within contemporary debates about the FEP, MENACE becomes a concrete, falsifiable example of how the framework can be applied responsibly. Rather than claiming MENACE "is" Active Inference, we show precisely how it maps to a restricted case.

Fourth, our experiments translate these claims into quantitative evidence: MENACE and an instrumental Active Inference baseline match within seed-to-seed variation on post-training validation against optimal play (Table 9.1), epistemic variants change Pure Active Inference learning dynamics and increase epistemic-value contributions without outperforming the $\lambda = 0$ baseline within our 500-game budget (Figure 9.2), and tabular RL highlights the importance of curriculum and coverage: defensive-only training can collapse under opponent shift, while random-opponent training with a larger budget can approach minimax performance.

11.2 Limitations and Boundary Conditions

The correspondence between MENACE and Active Inference is deliberately scoped. It is exact for the particular combination of Tic-Tac-Toe, complete state enumeration, and the bead-based mechanics we have formalised, but several boundary conditions constrain how far the conclusions can be generalised.

11.2.1 Model Complexity

MENACE maintains no explicit model of its opponent beyond a simple prior over moves. By contrast, a full Active Inference agent would maintain structured beliefs about opponent strategies, allowing it to anticipate and adapt to systematic changes in play. Our oracle AIF baseline removes transition/outcome uncertainty by using the solved game tree, but it still evaluates actions under an assumed opponent-policy model (uniform in our runs). It is therefore a diagnostic for model misspecification rather than an upper bound on performance against optimal play.

11.2.2 Temporal Abstraction

MENACE operates at a single temporal scale: individual moves within a game. It cannot reason about multi-game curricula, meta-learning across opponents, or slower forms of structural change. Modern Active Inference architectures extend naturally to hierarchical and deep temporal models. MENACE approximates only the lowest layer of such hierarchies.

11.2.3 Epistemic Blindness

Although MENACE's Dirichlet beliefs track uncertainty, the agent never selects actions because they are informative. Exploration arises implicitly from near-uniform posterior means early in learning (when counts are small and symmetric), not from a prospective valuation of information—posterior variance shrinks concurrently but does not directly drive selection probability under probability matching. This epistemic blindness explains both MENACE's competitive short-horizon performance and the gap that remains to explicitly epistemic agents over longer horizons.

11.2.4 State Space Constraints

Our analysis relies on exhaustive enumeration of the canonical state space, which is feasible for Tic-Tac-Toe but not for larger games or continuous domains. Extending the MENACE–AIF correspondence to function approximation or very large state spaces will require additional assumptions and may alter some of the clean guarantees we obtain in the finite setting.

11.3 Design Principles

Our analysis reveals fundamental design principles that transcend MENACE's specific implementation:

11.3.1 Natural Exploration Through Uncertainty

Posterior predictive probability matching supplies implicit exploration without ad-hoc bonuses—even when epistemic value is not explicitly scored. MENACE demonstrates that representing uncertainty through probability distributions produces near-uniform action probabilities early in learning (when pseudo-counts are small and symmetric), while confident beliefs (higher pseudo-counts) concentrate action probabilities and support exploitation later on. Variance shrinks concurrently as a concomitant of concentration, but selection probability under probability matching is the posterior mean, not a direct function of variance. Modern AI systems can inherit this advantage by maintaining calibrated uncertainty estimates instead of relying solely on externally tuned exploration schedules.

11.3.2 Efficient Learning Through Conjugacy

The Dirichlet–categorical structure enables exact Bayesian updates without approximation or sampling. MENACE shows that choosing the right representational framework can make seemingly complex computations trivial. For modern systems, this suggests seeking conjugate representations where possible, or approximations that preserve the essential structure of exact inference.

11.3.3 Physical Interpretability

Bead counts directly represent belief strength, making MENACE's knowledge state completely transparent. Each bead is a unit of evidence, and the learning process is visible as the physical rearrangement of beads. This interpretability is increasingly important for modern AI systems, suggesting that we should prefer representations where learned parameters have clear semantic meaning.

11.3.4 Embodied Computation

MENACE performs Bayesian inference not through digital computation but through the physical process of random selection. This demonstrates that intelligence can emerge from the interaction between simple mechanisms and environmental feedback, without explicit reasoning or calculation.

11.3.5 Minimal Sufficient Structure

MENACE succeeds with just 287 matchboxes, showing that complex behaviour can emerge from minimal structure when that structure correctly captures the problem's essential features. In fully enumerable domains like Tic-Tac-Toe, this demonstrates that the right inductive biases can outweigh raw capacity—a principle that may apply more broadly to well-structured subproblems within larger systems.

11.4 Modern Applications

The MENACE-FEP correspondence suggests several promising directions:

11.4.1 Explicit Uncertainty Representation

Modern deep reinforcement learning often discards uncertainty information, maintaining only point estimates of values or policies. MENACE suggests maintaining full Dirichlet distributions (or suitable approximations) for discrete action spaces. Recent work on distributional RL [22] partially implements this principle but could go further in maintaining conjugate representations.

11.4.2 Hierarchical Extensions

MENACE operates at a single level of abstraction, but the principle extends naturally to hierarchical settings. Imagine multiple levels of matchboxes, where higher levels select subpolicies and lower levels select primitive actions. Each level maintains its own uncertainty representation, enabling hierarchical exploration and compositional learning.

11.4.3 The MENACE Decomposition as a Reusable Template

The algorithmic pattern extracted from MENACE—independent Dirichlet pseudo-counts over action tendencies, updated via scalar reinforcement and sampled via posterior predictive probability matching with trajectory-based credit assignment—provides a useful template for discrete domains. Compared to typical deep reinforcement learning pipelines, it offers unusually strong interpretability and an explicit uncertainty representation. The trade-off is that the approach depends on an explicit state abstraction and on design choices (priors, update magnitudes, and opponent/curriculum assumptions) that materially affect learning dynamics.

11.4.4 Hardware Implementations

MENACE's physical implementation suggests possibilities for neuromorphic or stochastic hardware that naturally implements probabilistic computation. Stochastic computing elements could implement probability matching directly through physical randomness, while physical reservoirs could maintain count-based representations.

11.4.5 Interpretable AI

MENACE's transparent operation—where every decision can be traced to specific beads—illustrates the value of representations where each parameter has clear meaning and influence.

11.5 Comparison with Modern Game-Playing AI

It is instructive to contrast MENACE with contemporary systems such as AlphaZero [23], which couple deep function approximation with Monte Carlo Tree Search to achieve superhuman performance in large games. These systems trade transparency for scalability: their learned representations are distributed across large parameter sets and are typically difficult to audit mechanistically.

MENACE, by comparison, dispenses with search and function approximation. Its policy consists of explicit per-state Dirichlet pseudo-counts updated by a small set of reinforcement rules and sampled via probability matching. This makes MENACE a particularly suitable bridge case for Active Inference: it lets us isolate which behaviours follow from uncertainty-driven sampling and which require explicit epistemic terms or richer generative models.

11.6 Future Research Directions

- Connect-4 with partial enumeration, symmetry reduction, and heuristic rollouts to test scalability beyond full enumerability.
- Richer observation models that incorporate noisy or partial board information to stress the identity-observation assumption.
- Explicit opponent models and learning them online rather than fixing a prior.
- Policy priors and structural priors that encode move-order preferences or game-theoretic constraints.

11.6.1 Formal Optimisation of Initial Conditions

An interesting direction is to derive principled initial bead priors from a formal objective (e.g., Bayesian model reduction or an explicit EFE criterion) rather than hand-designing them. This would turn Michie's 4–3–2–1 schedule into a quantitative hypothesis: do priors that reflect the game's early branching structure improve sample efficiency, and can such priors be recovered automatically under reasonable assumptions?

11.6.2 Scaling to Larger Games

Connect-4, with a state space on the order of 10^{13} positions (depending on the counting convention), sits near the boundary of practical enumerability. It is large enough to require symmetry reduction and sampling, yet small enough that a carefully engineered abstraction might retain interpretability. Extending MENACE-style Dirichlet learning to this domain would provide a meaningful stress test of the approach.

11.6.3 Hybrid Architectures

Exploring architectures that combine MENACE's transparency with the planning power of modern systems like AlphaZero could yield systems that are both powerful and interpretable. For instance, using MENACE-style Dirichlet beliefs at leaf nodes of a search tree.

11.6.4 Adaptive Epistemic Weights

Our analysis suggests that a fixed epistemic weight λ need not be optimal across learning phases. Future work should explore adaptive λ schedules that mirror MENACE's implicit annealing, potentially recovering the elegant simplicity that made Michie's design so effective.

11.6.5 Physical Active Inference Systems

Can we identify other physical systems that naturally implement Active Inference? How can we design modern hardware that exploits these principles as elegantly as MENACE?

Looking forward, the MENACE–FEP correspondence suggests several avenues for extending this programme while preserving the interpretability and reproducibility that make MENACE a useful bridge case.

11.7 Broader Implications

By revisiting a mechanically explicit learner, the thesis demonstrates how Active Inference concepts can be operationalised in a fully enumerable domain. This style of analysis encourages explicit modelling assumptions and provides concrete behavioural signatures that can be measured rather than asserted.

The same approach may be applied to other minimalist or historically significant learning systems, where transparency allows competing theoretical interpretations to be compared under controlled experiments.

Methodologically, grounding the Free Energy Principle in a concrete system highlights the importance of explicit modelling commitments. Active Inference becomes operational only when the generative model, preference structure, and epistemic weighting are specified, and only then can instrumental trade-offs be evaluated rather than remaining purely descriptive.

11.8 Philosophical Reflections

MENACE illustrates that a non-trivial portion of sequential decision-making can be realised by simple, interpretable mechanisms. In this thesis, the Dirichlet–categorical mapping makes this explicit: bead counts represent belief strength, and posterior predictive sampling converts uncertainty into exploration pressure. The analysis also highlights the role of modelling commitments: behaviour depends on how outcomes are valued, what is treated as uncertain, and what aspects of the environment are included in the generative model.

In this sense, MENACE functions as a useful bridge case. It ties the expected-free-energy vocabulary to explicit computations and to measurable quantities in a fully enumerable setting, while clarifying which ingredients are absent (e.g., explicit epistemic planning over long horizons) and which would be required to scale to more complex domains.

11.9 Conclusion

This thesis used MENACE as a fully enumerable test bed for connecting a concrete learning mechanism to the Free Energy Principle and Active Inference. We provided a precise Dirichlet–categorical mapping of matchboxes and beads, derived an instrumental expected free energy special case corresponding to MENACE's learning rule, and validated the resulting correspondence empirically against Active Inference variants and tabular reinforcement learning baselines.

The broader lesson is methodological: Active Inference becomes most informative when expressed as an explicit, testable computational objective under clearly stated modelling assumptions. Future work should extend this programme to richer opponent models, longer planning horizons, and larger games (e.g., Connect-4), while preserving the interpretability and reproducibility that make MENACE such a valuable bridge case.

References

[1] Conor Heins, Beren Millidge, Lancelot Da Costa, Alexander Tschantz, Noor Sajid, and Karl Friston. pymdp: A Python library for active inference in discrete state spaces. *Journal of Open Source Software*, 7(73):4098, 2022. doi:10.21105/joss.04098.

[2] Samuel William Nehrer, Jonathan Ehrenreich Laursen, Conor Heins, Karl Friston, Christoph Mathys, and Peter Thestrup Waade. Introducing ActiveInference.jl: A Julia library for simulation and parameter estimation with active inference models. *Entropy*, 27(1):62, 2025. doi:10.3390/e27010062.

[3] Donald Michie. Experiments on the mechanization of game-learning Part I. Characterization of the model and its parameters. *The Computer Journal*, 6(3):232–236, 1963. doi:10.1093/comjnl/6.3.232.

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 2nd edition, 2018.

[5] Richard S. Sutton. 1.6 History of Reinforcement Learning. Online chapter notes, 2018. Accessed 21 September 2025. URL: http://incompleteideas.net/book/1/node7.html.

[6] Donald Michie. Game-playing and game-learning automata. In Leslie Fox, editor, *Advances in Programming and Non-Numerical Computation*, chapter 8, pages 183–200. Pergamon Press, Oxford, 1966.

[7] Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PLoS ONE*, 4(7):e6421, 2009. doi:10.1371/journal.pone.0006421.

[8] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active Inference: A Process Theory. *Neural Computation*, 29(1):1–49, January 2017. doi:10.1162/NECO_a_00912.

[9] Jeffrey S. Bowers and Colin J. Davis. Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389–414, 2012. doi:10.1037/a0026450.

[10] Samuel J. Gershman. What does the free energy principle tell us about the brain? *Neurons, Behavior, Data analysis, and Theory*, 1(3):1–12, 2019. doi:10.51628/001c.18282.

[11] Kathryn Nave. *A Drive to Survive: The Free Energy Principle and the Meaning of Life.* MIT Press, Cambridge, MA, 2025. doi:10.7551/mitpress/15519.001.0001.

[12] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. doi:10.2307/2332286.

[13] Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39.1–39.26. JMLR Workshop and Conference Proceedings, 2012. URL: https://proceedings.mlr.press/v23/agrawal12.html.

[14] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O'Doherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879, 2016. doi:10.1016/j.neubiorev.2016.06.022.

[15] Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu,

and Karl Friston. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99:102447, 2020. `doi:10.1016/j.jmp.2020.102447`.

[16] IEEE Computer Society. Donald Michie. Computer Pioneers. Accessed 29 December 2025. URL: `https://history.computer.org/pioneers/michie.html`.

[17] Stephen Muggleton. Donald Michie. The Guardian (Obituary), July 2007. Accessed 29 December 2025. URL: `https://www.theguardian.com/science/2007/jul/10/uk.obituaries1`.

[18] Artificial Intelligence Applications Institute. Donald Michie Home Page. University of Edinburgh. Accessed 29 December 2025. URL: `https://www.aiai.ed.ac.uk/~dm/dm.html`.

[19] Matthew Scroggs. MENACE. Blog post, 2019. Accessed 21 September 2025. URL: `https://www.mscroggs.co.uk/blog/19`.

[20] John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. English translation: “On the Theory of Games of Strategy” in Contributions to the Theory of Games, vol. 4 (1959). `doi:10.1007/BF01448847`.

[21] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

[22] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[23] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi and Go through self-play. *Science*, 362(6419):1140–1144, 2018. `doi:10.1126/science.aar6404`.

[24] Thomas P. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft Research, 2000. URL: `https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf`.

Acknowledgements

I would like to thank Professor Hideki Nakayama for his guidance and feedback throughout this project.

A

Canonical State Filters

To ensure reproducibility we record the canonical state counts and golden sets used throughout the experiments. These match the automated checks in the test suite (see `exact_count_validation.rs` and `double_threat_golden.rs` under `tests/`).

A.1 State-Filter Taxonomy

Table A.1. State filter definitions and counts.

Filter	Description	Count
All	All canonical X-to-move states (no pruning).	338
Decision-only	States with more than one legal move (forced moves excluded).	304
Michie	Decision-only minus 17 double-threat positions (original design).	287

All counts above refer to X-to-move canonical states. Michie's hardware played as O and opened the game, and the software maps between these conventions by swapping symbols and the player-to-move bit whenever MENACE or its baselines act.

`All` is useful for exhaustive evaluation, `Decision-only` removes trivial continuations, and `Michie` reproduces the 1960s hardware footprint. The filters are implemented consistently across MENACE, Active Inference, and RL baselines.

A.2 Double-Threat Positions

The 17 double-threat positions excluded by the Michie filter are enumerated with visual board diagrams in Appendix B.3. These positions represent board states at ply 6 where O has created a fork—two or more distinct winning threats that X cannot simultaneously block.

For reference, the canonical labels (in `XXXXXXXXX_P` format where `P` denotes player to move) are embedded in the repository at `resources/double_threat_positions.txt` and verified against the test suite (`double_threat_golden.rs`).

B

Game-Tree Enumeration and State Space Analysis

This appendix provides the complete enumeration of the Tic-Tac-Toe state space and explains the relationship between the canonical counts 287, 304, and 338 that appear in the MENACE literature.

B.1 Canonical State Space Counts

The enumeration harness used in Chapter 5 confirms the canonical counts for Tic-Tac-Toe:

- Total legal games: 255,168
- Canonical trajectories (up to D_4 symmetry): 26,830
- Outcome histogram: 131,184 X wins, 77,904 O wins, 46,080 draws
- Length distribution:
  - 5 moves: 1,440 games
  - 6 moves: 5,328 games
  - 7 moves: 47,952 games
  - 8 moves: 72,576 games
  - 9 moves: 127,872 games

These values match both classical analyses and the automated test suite. They provide the reference distribution for computing policy KL divergences, coverage metrics, and expected free energy components in the comparative experiments.

B.2 The Mathematics of 287, 304, and 338

Michie's 1963 paper states that MENACE contained "every one of the 287 essentially distinct positions which the opening player can encounter" [3]. This section derives that figure and clarifies related counts.

B.2.1 Definition of Essential Distinctness

Michie identifies board positions up to the 8 geometrical symmetries of the 3×3 grid—the dihedral group D_4 comprising 4 rotations and 4 reflections. Two boards that are rotations or reflections of each other count as one position.

B.2.2 Enumeration of Opening Player Choice-Points

MENACE plays as the opening player (X) and requires a matchbox only when facing a genuine decision (at least two legal moves). Enumerating the legal game tree (stopping at wins), reducing by symmetry, and considering only X-to-move positions yields:

Decision Point	Ply	Positions
Before 1st move	0	1
Before 3rd move	2	12
Before 5th move	4	108
Before 7th move	6	183
Total		**304**

Table B.1. X-to-move decision states by ply, reduced by D_4 symmetry.

The breakdown $1 + 12 + 108 + 183 = 304$ is widely reproduced in the literature.

B.2.3 Exclusion of 9th Move Positions

With eight marks placed, only one empty square remains—no decision is required. These 34 symmetry-reduced positions (at ply 8) are not choice-points. Including them yields 338 total X-to-move states.

B.2.4 From 304 to 287: The Double-Threat Subtraction

Among the 183 “before 7th-move” positions, there are exactly 17 symmetry-classes where the opponent already has two distinct immediate winning threats (a fork) and the opener has no immediate winning move. In these states, all legal moves lose on the next turn.

Subtracting these inevitable-loss classes:

$$\underbrace{1 + 12 + 108 + 183}_{\text{all choice-points}} - \underbrace{17}_{\text{inevitable-loss}} = \boxed{287} \tag{B.1}$$

B.2.5 Summary of State Counts

Count	Description
338	All canonical X-to-move states (including forced final moves)
304	Decision points only (excluding forced single-move positions)
287	Michie’s count (excluding forced moves and double-threat losses)

Table B.2. Relationship between canonical state counts in the MENACE literature.

B.3 The 17 Double-Threat Classes

A *double-threat* (or *fork*) occurs when O has placed pieces such that two distinct winning lines each contain two O marks and one empty square. Since X can only block one threat per move, O wins on the following turn regardless of X's choice. These positions represent inevitable losses for the opening player.

B.3.1 Board Position Indexing

Throughout this section, board positions use row-major indexing from 0 to 8:

0	1	2
3	4	5
6	7	8

B.3.2 Enumeration of Double-Threat Positions

Table B.3 enumerates all 17 symmetry-classes of double-threat positions. Each represents a canonical board state at ply 6 (three X marks, three O marks) where:

1. X has no immediate winning move (no line with two X marks and one empty square), and
2. O has at least two distinct winning moves (at least two lines each containing two O marks and one empty square).

B.3.3 Observations

Several patterns emerge from the enumeration:

- **Triple threats**: Positions 10, 15, and 16 have three winning squares for O rather than two. These arise when O's pieces simultaneously threaten along three distinct lines.
- **Fork structures**: Most double-threats involve O occupying the center (position 4) or a corner, creating intersecting threats along rows, columns, and diagonals.
- **Symmetry classes**: Each position shown is one canonical representative. Applying the eight D_4 transformations (rotations and reflections) generates between 1 and 8 equivalent raw board states per class, depending on the position's inherent symmetry.
- **Pedagogical significance**: These 17 positions represent the "traps" that a novice player must learn to avoid. Optimal play by X prevents O from ever reaching any of these configurations.

B.4 Variant Counts in the Literature

Different sources report different counts:

- **304**: Most modern reconstructions include all genuine decision points, hence $1 + 12 + 108 + 183 = 304$ matchboxes. This is the standard build count in tutorials.
- **287**: Michie's 1963 paper excludes the 17 "no-escape" cases as not worth separate boxes.
- **288**: In later writings Michie sometimes mentioned 288; the literature is not fully consistent on fringe cases.

The implementation in this thesis supports all three filters via the `StateFilter` enum: `All` (338), `DecisionOnly` (304), and `Michie` (287).

#	Board Position			O's Winning Squares
1	O	X	O	6, 8 (diagonal & anti-diagonal)
	X	O	X	
2	X	X	O	4, 8 (column & diagonal)
	X		O	
	O			
3	X	O		2, 7 (column & row)
	X	O	X	
	O			
4	O	X	O	6, 8 (anti-diagonal & diagonal)
	X	O		
		X		
5	X	O	X	5, 7 (row & column)
	O	O		
	X			
6	X	O	O	6, 7 (column & anti-diagonal)
		O	X	
			X	
7	X	O	X	4, 8 (row & diagonal)
	X			
	O	O		
8	X	O		2, 7 (column & row)
	X	O		
	O		X	
9	X	X	O	5, 6 (row & anti-diagonal)
		X		
		O	O	

Table B.3. Double-threat positions 1–9. Shaded cells indicate empty squares; O's winning squares complete two-in-a-row threats.

#	Board Position	O's Winning Squares
10	X X O ▨ O O ▨ X ▨	3, 6, 8 (row, column & anti-diag.)
11	X X O ▨ O ▨ ▨ X O	5, 6 (column & anti-diagonal)
12	X X O ▨ ▨ X O O ▨	4, 8 (diagonal & row)
13	O X ▨ O O X ▨ X ▨	6, 8 (column & diagonal)
14	▨ X ▨ O X X O O ▨	0, 8 (column & row)
15	O X ▨ ▨ O X O X ▨	2, 3, 8 (column, row & diagonal)
16	X X O X ▨ ▨ O ▨ O	4, 5, 7 (diag., row & column)
17	X X O ▨ ▨ X O ▨ O	4, 7 (diagonal & row)

Table B.4. Double-threat positions 10–17. Positions 10, 15, and 16 exhibit *triple* threats.

C

Experiment Automation

This appendix records the commands used to reproduce the experimental outputs reported in this thesis. The workflow is Makefile-driven and uses the Python experiment driver under `scripts/automation/`.

C.1 End-to-end pipeline

From the repository root, the following command builds the release binary, runs the thesis experiment suite, performs post-training evaluations, and regenerates aggregate summaries and reports:

```
make thesis-results
```

The thesis experiment configuration is version-controlled at:

- Training configuration: `configs/thesis_experiments.yaml`
- Evaluation configuration: `configs/evaluate_thesis.yml`

To run stages separately:

```
make release               # cargo build --release
make thesis-results-run    # run thesis experiments + evaluations
make thesis-results-report # analyze + regenerate reports
```

C.2 Building the thesis PDF

The LaTeX sources are under `thesis/`. To build the PDF (requires a Japanese-capable TeX toolchain):

```
make -C thesis
```

C.3 Packaging results for external audit (optional)

To package a minimal subset of seed-level outputs (excluding large binaries) into a deterministic ZIP with checksums:

```
make thesis-package-minimal
```

To additionally include per-seed training traces (`metrics.jsonl`) and selected EFE export CSVs:

```
make thesis-package-curves
```

C.4 Smoke test

To validate the toolchain without running the full experiment suite, the following short train/evaluate cycle should complete quickly:

```
cargo build --release

./target/release/menace train menace \
  --games 50 \
  --opponent random \
  --output menace_smoke.msgpack

./target/release/menace evaluate menace_smoke.msgpack \
  --games 50 \
  --opponent random \
  --seed 0
```

Note: `--output` writes a serialised agent file (MessagePack format), while `--summary` writes JSON summaries.

C.5 Generated artefacts

The pipeline produces:

- `menace_data/<run>/<condition>/seed_<n>/training_summary.json` — per-seed training summaries
- `menace_data/<run>/<condition>/seed_<n>/evaluation.json` — per-seed post-training evaluation exports
- `results/analysis_summary.json` — aggregated training metrics
- `results/evaluation_summary.json` — aggregated evaluation metrics
- `menace_reports/` — generated report figures (PNG) for the specified run directories

C.6 Reproducibility

Experiments use explicit random seeds (0–9 by default). The experiment suite is fully specified by the configuration files above and the repository revision; running `make thesis-results` on the same codebase therefore reproduces the same directory structure and seed-level artefacts, up to expected stochastic variation when seeds are changed.

D

Dirichlet Entropy and Epistemic Value

This appendix derives the expected categorical entropy under a Dirichlet prior, which yields the closed-form mutual-information term $I(o;\theta)$ (epistemic value) used throughout the thesis. Let $\theta \sim \mathrm{Dir}(\alpha)$ with $\alpha_0 = \sum_i \alpha_i$. Using the moment identity [24]

$$\mathbb{E}[\theta_i \ln \theta_i] = \frac{\alpha_i}{\alpha_0}\Big(\psi(\alpha_i + 1) - \psi(\alpha_0 + 1)\Big), \tag{D.1}$$

we obtain

$$\mathbb{E}\Big[-\sum_i \theta_i \ln \theta_i\Big] = \psi(\alpha_0 + 1) - \sum_i \frac{\alpha_i}{\alpha_0}\psi(\alpha_i + 1), \tag{D.2}$$

where ψ denotes the digamma function. The mutual information between the categorical parameter θ and a single observation o under the Dirichlet–categorical model therefore has the closed form

$$I(o;\theta) = H\left[\frac{\alpha}{\alpha_0}\right] - \left[\psi(\alpha_0 + 1) - \sum_i \frac{\alpha_i}{\alpha_0}\psi(\alpha_i + 1)\right], \tag{D.3}$$

which matches the implementation in `dirichlet_categorical_mi(`α`)` and guarantees that epistemic value is measured in the same units as the pragmatic risk term. Mutual information is symmetric, so $I(o;\theta) = I(\theta;o)$; we use $I(o;\theta)$ throughout the thesis.

E

Glossary of Technical Terms

Active Inference A framework for understanding perception and action as processes that minimise free energy through both belief updating and action selection.

Conjugate prior A prior distribution that, when combined with a particular likelihood function, yields a posterior in the same family. The Dirichlet is conjugate to the categorical distribution.

Dirichlet distribution A multivariate generalisation of the Beta distribution that describes uncertainty over probability simplexes. Parameters can be interpreted as pseudo-counts.

Expected free energy A quantity that evaluates future policies by combining pragmatic value (achieving preferred outcomes) with epistemic value (reducing uncertainty).

Free Energy Principle (FEP) The theory that adaptive systems minimise variational free energy, thereby implicitly minimising surprise and maintaining their structural integrity.

Generative model A probabilistic model of how observations are generated from hidden states, actions, and parameters.

KL divergence Kullback-Leibler divergence measures the difference between two probability distributions, quantifying information lost when approximating one distribution with another.

MENACE Matchbox Educable Noughts And Crosses Engine, a mechanical learning device built from matchboxes and beads that learns to play Tic-Tac-Toe.

Policy A mapping from states to actions or a sequence of planned actions. In MENACE, this is the strategy for selecting moves.

Probability matching A decision strategy where actions are sampled according to their posterior predictive probabilities. In MENACE, this means drawing beads with probability $\alpha_{s,a}/\alpha_{s,0}$, providing implicit exploration via posterior predictive sampling (no explicit epistemic-value bonus) and differing from Thompson sampling by skipping the parameter-sampling step.

Risk (Active Inference) The instrumental term in expected free energy. In this thesis it is instantiated as $D_{\mathrm{KL}}(q(o \mid \pi) \,\|\, p(o \mid C))$ over terminal outcomes (win/draw/loss), which differs from the cross-entropy form by an additive entropy term.

Ambiguity The outcome-entropy component $H[q(o \mid \pi)]$ in the EFE decomposition. When risk is scored as KL divergence, weighting ambiguity separately changes the net emphasis on outcome entropy; all experiments in this thesis set $\beta_{\mathrm{amb}} = 0$.

Variational free energy An upper bound on surprise (negative log evidence) that can be minimised to perform approximate Bayesian inference.