MENACE as a Bayesian Observer

A Technical Analysis through the Free Energy Principle

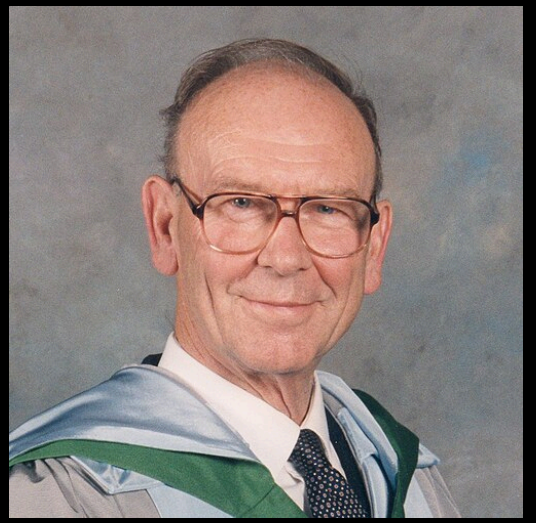
Krzysztof Woś

Nakayama Laboratory

2026-01-23

Michie's Question (1966)

> "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*."
>
> — Donald Michie, "Game-Playing and Game-Learning Automata" (1966)



Donald Michie

Expected free energy provides a Bayes-optimal formalization of the trade-off Michie identified.

Thesis Contributions

1. A mapping from MENACE to Active Inference in Dirichlet-categorical terms
2. Identification of MENACE as an instrumental Active Inference agent
3. Empirical comparison with Active Inference variants and Reinforcement Learning (RL) baselines
4. A generative model in which λ parameterizes the cost of information acquisition

Part I: Background

What is MENACE?

Physical Components

- 287 matchboxes (one per board position requiring a decision)
- Colored beads (9 colors for 9 board positions)
- Each bead represents a possible move

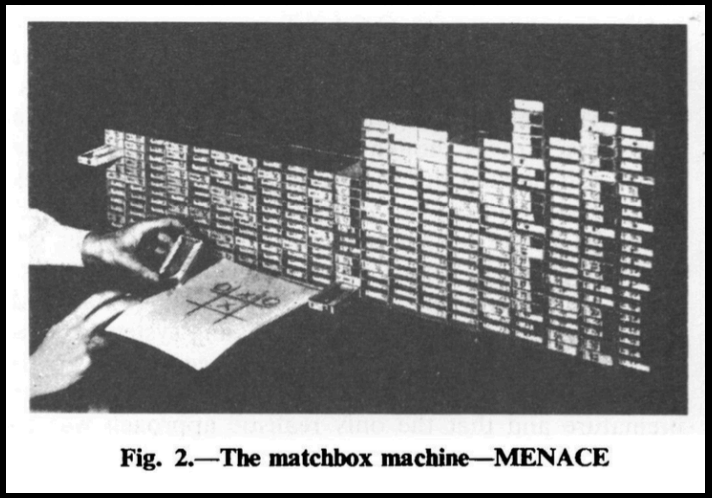


Fig. 2.—The matchbox machine—MENACE

Learning Mechanism

- Draw a bead → make that move
- Win: add 3 beads of that color
- Draw: add 1 bead
- Loss: remove 1 bead

The +3/+1/−1 values encode preferences over outcomes → map to Active Inference priors.

Reducing Complexity Through Symmetry

Without Symmetry

- 5,478 distinct legal positions
- Impractical to build

With Symmetry

- 8 symmetries (4 rotations × 2 reflections)
- 765 canonical positions after symmetry reduction
- 338 X-to-move → 304 (prune forced) → 287 (prune double-threats)
- Manageable physical system

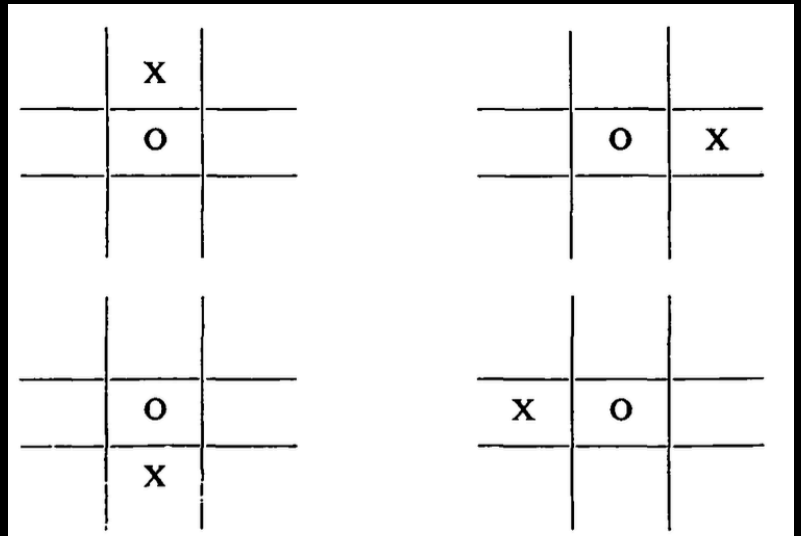


Fig. 3.—Four positions which are in reality variants of a single position

Part II: The Correspondence

MENACE Implements Dirichlet-Categorical Inference

Matchbox State

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

where α_s = bead counts

Action Selection

1. Bead counts → move probabilities (Dirichlet posterior predictive)
2. Uniform bead draw = sample from $\mathrm{Cat}(\alpha_s / \alpha_{s,0})$

Posterior Predictive Probability Matching

Drawing a bead uniformly = sampling from the posterior predictive

$$P(a|s) = E[\theta_{s,a}] = \frac{\alpha_{s,a}}{\alpha_{s,0}}$$

Thesis contribution: This mapping formalizes MENACE's implicit Bayesian structure.

θ_s: move probability vector for state s • α_s: bead counts (Dirichlet parameters) • $\alpha_{s,0} = \sum_a \alpha_{s,a}$: total beads • Dir: Dirichlet distribution • Cat: categorical distribution

Bayesian Updates Through Bead Manipulation

Standard Bayesian update

$$\text{Prior} + \text{Data} = \text{Posterior}$$

In Dirichlet terms

$$\text{Dir}(\alpha) + \text{observation} = \text{Dir}(\alpha + 1)$$

MENACE's version

$$\alpha_{s,a} \leftarrow \alpha_{s,a} + U(o)$$

where $U(o) \in \{3, 1, -1\}$ is a utility-weighted pseudo-count update.

Loss updates are a heuristic penalty/forgetting step, not a literal conjugate posterior.

The Free Energy Principle and Active Inference

Core idea: Adaptive systems minimize surprise — but surprise is intractable.

$$\text{Surprise} = -\ln p(o) \quad \text{(requires marginalizing over hidden states)}$$

Solution: Minimize free energy F, a computable upper bound on surprise.

$$F = \underbrace{D_{KL}[q(s|o) \| p(s|o)]}_{\geq 0} + \underbrace{(-\ln p(o))}_{\text{surprise}} \geq \text{surprise}$$

Active Inference operationalizes the FEP: update beliefs to match observations, or act to make observations match beliefs.

F: free energy • D_{KL}: KL divergence • $q(s|o)$: approximate posterior • $p(s|o)$: true posterior • $p(o)$: evidence

Expected Free Energy with λ Parameter

The epistemic weight λ prices information in outcome units:

$$G_\lambda(\pi) = \underbrace{\text{Risk}(\pi)}_{\text{instrumental cost}} - \lambda \underbrace{I(o;\theta)}_{\text{epistemic value}}$$

- Risk: $D_{KL}(q(o|\pi)\|p(o|C))$ — distance from preferred outcomes
- Epistemic Value: $I(o;\theta)$ — expected information gain about parameters

λ is Michie's "exchange rate" between information and immediate gain.

Decision objective: Expected free energy makes the utility + information trade-off explicit.

$G_\lambda(\pi)$: expected free energy • π: policy • λ: epistemic weight • o: outcome • θ: model parameters • $I(o;\theta)$: information gain about θ from observing o • $q(o|\pi)$: predicted outcomes • C: prior preferences • $p(o|C)$: preferred outcomes

MENACE's Implicit Generative Model

What MENACE "believes"

$$p(o, s_{0:T}, a_{0:T}) = p(o|s_T) \prod_t p(s_{t+1}|s_t, a_t) p(a_t|s_t)$$

where:

- $p(o|s_T)$: Outcome model — final position determines outcome (deterministic)
- $p(s_{t+1}|s_t, a_t)$: Transition model — game rules (fixed)
- $p(a_t|s_t)$: Policy to learn (the beads)

All of MENACE's learning focuses on optimizing the policy component.

---

o: outcome • s_t: state at time t • a_t: action at time t • T: terminal time • $p(o|s_T)$: outcome model • $p(s_{t+1}|s_t, a_t)$: transition model • $p(a_t|s_t)$: policy

Preference-Weighted Policy Shaping

MENACE corresponds to the instrumental objective

$$G_\lambda(\pi) = \mathrm{Risk}(\pi) - \lambda I(o;\theta), \quad \lambda = 0$$

- Preferences are encoded by $U(o) \in \{+3, +1, -1\}$
- Bead updates shift Dirichlet counts toward preferred outcomes
- Negative updates are heuristic penalties; restocking keeps $\alpha_{s,a} > 0$

The update is directionally aligned with reducing instrumental risk, not an exact gradient step.

Thesis contribution: This model is instantiated for MENACE and Tic-Tac-Toe, with explicit computation of all quantities.

$G_\lambda(\pi)$: expected free energy • π: policy • o: outcome • θ: model parameters (bead proportions) • $I(o;\theta)$: information gain about θ from o • $U(o)$: utility • $\alpha_{s,a}$: bead count

Part III: Empirical Analysis

Experimental Setup

Agents Compared

- MENACE (Michie filter, box restock)
- Instrumental AIF ($\lambda = 0$)
- Hybrid AIF ($\lambda = 0.5$)
- Pure AIF ($\lambda \in \{0, 0.25, 0.5\}$)
- Q-Learning / SARSA baselines

Protocol

- Training: 500 games (5,000 for RL)
- Validation: 100 games vs. optimal
- Seeds: 10 independent runs
- State filters: Michie (287 states)
- MENACE curriculum: mixed; AIF baseline: optimal-only

Thesis contribution: First systematic empirical comparison of MENACE against Active Inference variants.

Key Finding 1: Instrumental Equivalence

Algorithm	Draw Rate (%)	Loss Rate (%)
MENACE (box restock)	84.5 ± 8.1	15.5 ± 8.1
Instrumental AIF ($\lambda = 0$)	88.1 ± 3.9	11.9 ± 3.9

- When $\lambda = 0$, the EFE-based policy reduces to a purely instrumental objective, closely matching MENACE's pseudo-count reinforcement
- However, MENACE has implicit exploration: low concentration → high variance → naturally exploratory early behavior

Result: MENACE ≈ Instrumental Active Inference.

Key Finding 2: The Value of Information

Algorithm	Draw Rate (%)
Pure AIF ($\lambda = 0.0$)	79.7 ± 6.5
Pure AIF ($\lambda = 0.25$)	79.1 ± 4.8
Pure AIF ($\lambda = 0.5$)	77.0 ± 3.7

Paradox: Epistemic variants ($\lambda > 0$) do NOT outperform instrumental baseline ($\lambda = 0$) within 500 games.

Resolution: This reflects Michie's trade-off in practice: "acquisition of information for future use *at the expense of present expected gain*".

Key Finding 3: Robustness vs. Specialization

Algorithm	Training	Draw Rate (%)
Q-learning	random	98.0 ± 1.2
Q-learning	defensive	10.2 ± 30.9
SARSA	random	97.9 ± 1.9
SARSA	defensive	20.5 ± 40.3

- Box-level restocking preserves full support
- Provides implicit “insurance” against distributional shift
- Q-learning can drive Q-values to extremes, zeroing out actions
- MENACE achieves competitive performance with 10× fewer games

Dirichlet advantage: MENACE keeps all action probabilities strictly positive.

Answering Michie's Question

> "The difficulty lies in *costing the acquisition of information* for future use at the expense of present expected gain."

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

The scalar $\lambda \geq 0$ is the exchange rate Michie asked for — not a single algorithm, but a family:

Active Inference (General)

- Epistemic value priced via λ
- Trade-off is a design parameter

MENACE (Special Case)

- Instrumental objective ($\lambda = 0$)
- Exploration emerges from uncertainty

Thesis contribution: This framework answers Michie's question with MENACE as a concrete, mechanizable special case.

$G_\lambda(\pi)$: expected free energy • λ: exchange rate • $I(o;\theta)$: information gain • Risk: instrumental cost

Conclusions

Key Findings

- Dirichlet-categorical mapping formalizes MENACE's Bayesian structure
- MENACE ≈ instrumental Active Inference ($\lambda = 0$) confirmed empirically
- λ quantifies information cost with measurable short-horizon effects
- Dirichlet representation provides robustness that tabular RL lacks

MENACE is a historically remarkable, mechanizable special case of Active Inference.

Thank you for your attention

Code, data, and thesis: github.com/krzysztofwos/masters-thesis

Appendix

A1: Donald Michie (1923-2007)

The Pioneer Behind MENACE

- Codebreaker at Bletchley Park during WWII
- Worked alongside Alan Turing on Colossus
- MA, DPhil, DSc in biological sciences (Oxford)
- Professor of Machine Intelligence, Edinburgh

Key insight: “Programming human intelligence into machines” — inspired by wartime cryptanalysis.

Contributions to AI

- Founded Edinburgh’s Machine Intelligence unit
- Editor-in-Chief, “Machine Intelligence” series
- Developed machine learning into “industrial-strength tool”
- 1996 Feigenbaum Medal for ML applications

MENACE (1961): Physical demonstration that learning could be mechanized.

A2: Dirichlet-Categorical Conjugacy Proof

Theorem: If $\theta \sim \text{Dir}(\alpha)$ and n outcomes are observed, then:

$$\theta|\text{data} \sim \text{Dir}(\alpha + n)$$

Proof:

$$\begin{aligned} p(\theta|\text{data}) &\propto p(\text{data}|\theta)p(\theta) \\ &= \prod_{i=1}^{k} \theta_i^{n_i} \cdot \frac{1}{B(\alpha)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{k} \theta_i^{\alpha_i + n_i - 1} \end{aligned}$$

This is the kernel of $\text{Dir}(\alpha + n)$ ∎

A3: Posterior Predictive Probability Matching

Theorem: Drawing beads = posterior predictive probability matching.

Proof: For a Dirichlet–categorical model,

$$P(a) = \int \theta_a \cdot \mathrm{Dir}(\theta; \alpha)\, d\theta = \frac{\alpha_a}{\sum_i \alpha_i}$$

MENACE's probability

$$P(a) = \frac{\#\text{ beads of color } a}{\text{total beads}} = \frac{\alpha_a}{\sum_i \alpha_i}$$

Identical ∎

$P(a)$: probability of action a • θ_a: move probability for action a • $\mathrm{Dir}(\theta; \alpha)$: Dirichlet distribution • α_a: bead count for action a • $\sum_i \alpha_i$: total beads

A4: Dirichlet–Categorical Mutual Information

Epistemic value in the thesis is the mutual information between observations and parameters.

$$I(o;\theta) = H\left[\mathrm{Cat}\left(\frac{\alpha}{\alpha_0}\right)\right] - \left[\psi(\alpha_0+1) - \sum_i \frac{\alpha_i}{\alpha_0}\psi(\alpha_i+1)\right]$$

Equivalent form:

$$I(o;\theta) = \sum_i \frac{\alpha_i}{\alpha_0}\left[\psi(\alpha_i{+}1) - \psi(\alpha_0{+}1) - \ln\frac{\alpha_i}{\alpha_0}\right] \geq 0$$

As total concentration α_0 becomes large, $I(o;\theta) \to 0$.

$I(o;\theta)$: mutual information • o: observation (action or outcome) • $\theta \sim \mathrm{Dir}(\alpha)$: probability vector • α_i: concentration parameter for category i • $\alpha_0 = \sum_i \alpha_i$: total concentration • ψ: digamma function • H: entropy

A5: Deriving Variational Free Energy

Goal: Approximate intractable posterior $p(s|o)$ with tractable $q(s|o)$.

$$
\begin{aligned}
D_{KL}[q(s|o)\|p(s|o)] &= \mathbb{E}_q\left[\ln q(s|o) - \ln \frac{p(o|s)p(s)}{p(o)}\right] \\
&= \mathbb{E}_q[\ln q(s|o) - \ln p(o|s) - \ln p(s)] + \ln p(o) \\
&= \underbrace{D_{KL}[q(s|o)\|p(s)] - \mathbb{E}_q[\ln p(o|s)]}_{F\ =\ \text{Free Energy}} + \ln p(o)
\end{aligned}
$$

Since $D_{KL} \geq 0$: $\quad F \geq -\ln p(o) = \text{Surprise}$

Key insight: Minimizing F simultaneously approximates the posterior and bounds surprise.

$q(s|o)$: approximate posterior • $p(s|o)$: true posterior • $p(o)$: evidence/marginal likelihood • F: variational free energy

A6: Three Equivalent Forms of Free Energy

Energy − Entropy: Fit data while maintaining uncertainty

$$F = \underbrace{-\mathbb{E}_q[\ln p(o,s)]}_{\text{Energy}} + \underbrace{H[q(s \mid o)]}_{\text{Entropy}}$$

Complexity − Accuracy: Simple models that explain data well

$$F = \underbrace{D_{KL}[q \parallel p(s)]}_{\text{Complexity}} - \underbrace{\mathbb{E}_q[\ln p(o \mid s)]}_{\text{Accuracy}}$$

Divergence + Surprise: Approximate inference while minimizing surprise

$$F = \underbrace{D_{KL}[q \parallel p(s \mid o)]}_{\text{Divergence}} + \underbrace{(-\ln p(o))}_{\text{Surprise}}$$