No Selves, No Consciousness

Michael Timothy Bennett

School of Computing, The Australian National University
Canberra, ACT, Australia
m@michaeltimothybennett.com

Abstract

Generalisation-optimal learning favours weak rules that keep many futures open. Adaptation under this heuristic yields a hierarchy of selves. Previous Stack Theory results show this hierarchy is necessary for human-like consciousness. A first-order self tags intervention versus observation, enabling causal learning and underpinning subjective experience. A second-order self models the listener, letting self-report survive decoder mismatch. It is necessary for access consciousness, self-awareness and Gricean meaning. A third-order self binds the future self, making long-horizon trust rational and enabling narrative planning. Without this hierarchy, a system lacks ingredients of human-like conscious experience. Here I do not restate these arguments, but supply formal necessity proofs for these first three orders of self grounded in observable behaviour. I also validate the predicted capability gaps in three randomised Monte Carlo experiments. Hence when I say no selves, no consciousness, I mean no consciousness like we humans have. I then position Stack Theory relative to Embedded Universal Predictive Intelligence (EUPI), which recreates many earlier Stack Theory results but inherits from AIXI a reliance on description-length priors whose optimality depends on the choice of reference machine. Stack Theory's weakness principle is representation-invariant. It maximises generalisation probability without requiring a privileged encoding. I discuss relative strengths, proposing that bridging the two frameworks could combine Stack Theory's firmer theoretical foundation with EUPI's ready integration into reinforcement learning.

Code — https://github.com/ViscousLemming/Technical-Appendices/blob/main/Papers/No_Selves_No_Consciousness/NSNC_Experiments.py

Extended version — https://github.com/ViscousLemming/Technical-Appendices/blob/main/Papers/No_Selves_No_Consciousness/NSNC_SI.pdf

1 Introduction

The core practical problem in machine consciousness is not only deciding what consciousness is, but what would count as evidence of it.

Behavioural evidence can be misleading. Its meaning is filtered by compression and by decoder mismatch. An advanced agent can look like noise, and a simple controller can look intentional (Bennett 2022; Li and Vitányi 2008). In that setting, disagreement about consciousness can collapse into disagreement about the observer.

Stack Theory circumvents this problem by examining architectural constraints on behaviour. If a system must learn under unknown future constraints, then the generalisation-optimal inductive bias is weakness (Bennett 2023c, 2025a). Weakness is a functional least-commitment rule. It keeps the largest set of future completions compatible with what has been observed.

Under limited resources, the same least-commitment pressure forces an agent to represent itself at increasing levels of abstraction. This yields a hierarchy of selves defined as nested causal identities. Each self is a causal identity that makes a different kind of attribution publicly legible. I make four contributions.

1. I restate causal identity and the self hierarchy in Stack Theory notation that aligns with related work.
2. I derive a likelihood ratio bound that formalises decoder mismatch for passive attribution.
3. I prove three necessity theorems linking first-, second-, and third-order selves to intervention control, audience-calibrated self-report, and binding commitment.
4. I run three Monte Carlo benchmarks with matched assumption controls. The benchmarks randomise abstraction layers, decoder sets, and payoff tables.

Importantly, I do not restate the arguments associating these selves with various aspects of consciousness. These arguments are covered at length elsewhere (Bennett 2023a; Bennett, Welsh, and Ciaunica 2024; Bennett 2025b).

Novelty and provenance. This paper consolidates earlier Stack Theory results on weakness and causal identity and extends them toward a measurement program for consciousness (Bennett 2023c, 2025a; Bennett, Welsh, and Ciaunica 2024; Bennett 2025b). The uniform extension Bayes-optimality argument is restated for completeness. The new material in this submission is the decoder mismatch bound, the three randomised benchmarks, and the synthesis that links self layers to public evidence. All headline numbers are

generated by a single driver script and cross-checked against its JSON summaries.

Three orders. Self prediction is compatible with attribution failure, report failure, and commitment failure. I isolate three additional requirements that are independently checkable.

1. Attribution requires an internal do versus see tag and a minimal causal separator. Theorem 2 and Theorem 3. Experiment 1 shows that, under extension priors, a weakness selector prefers a robust separator over a spurious correlate that is indistinguishable on the child task.
2. Report requires identifying the listener decoder and shaping what the listener will infer the speaker means. That further requires predicting the listener's prediction of the speaker's decoder. Theorem 4. Experiment 2.
3. Commitment works when there is a publicly legible binding move that restricts the future self. Without binding, long-horizon trust collapses in equilibrium. Theorem 5. Experiment 3.

2 Results

I start with the experiments, because the claims are falsifiable and the failure modes are visible in the plots. The formal machinery is in Section 3; full proofs, background definitions, and detailed methods are in the Supplementary Information (SI).

Reading guide. Each plot and table aggregates across randomised worlds. Means are world means and uncertainty intervals are 95% bootstrap intervals over worlds. Experiment 1 fails if weakness and simplicity collapse into the same choice on most tasks or if the regret gap vanishes under parent extensions. Experiment 2 fails if probing does not approach the oracle decoder within a few probes while passive signalling remains far below probing. Experiment 3 fails if costless nonbinding promises produce trust in the base equilibrium model without additional type information.

2.1 Experiment 1: Weakness Beats Simplicity

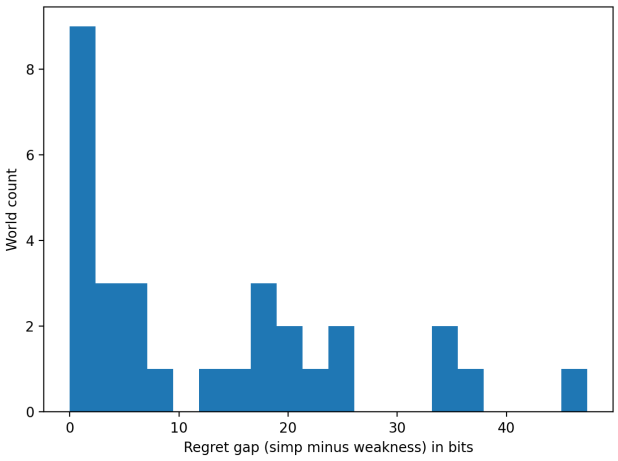
Each world samples a toy finite environment Φ with 32 to 64 mutually exclusive states and a vocabulary size $|\mathfrak{v}|$ in $[9, 13]$. Each vocabulary item is a program over Φ. I then form the induced language $L_{\mathfrak{v}}$ of satisfiable statements. A statement is a conjunction of vocabulary items. In code it is a bitmask over indices. I embed a minimal backbone with five named programs z, j, k, u_j, and u_k. The remaining programs are auxiliary and are randomised across worlds.

The hidden self policy is $\pi_* = j \wedge k$. Each child task provides one positive training statement m_{train} that contains z, j, and k, plus at least one auxiliary item. That training statement is the only output. The input set also contains the singleton statements u_j and u_k as negative evidence. This construction makes z and $j \wedge k$ observationally indistinguishable on the child task, but distinguishable by parent extension behaviour.

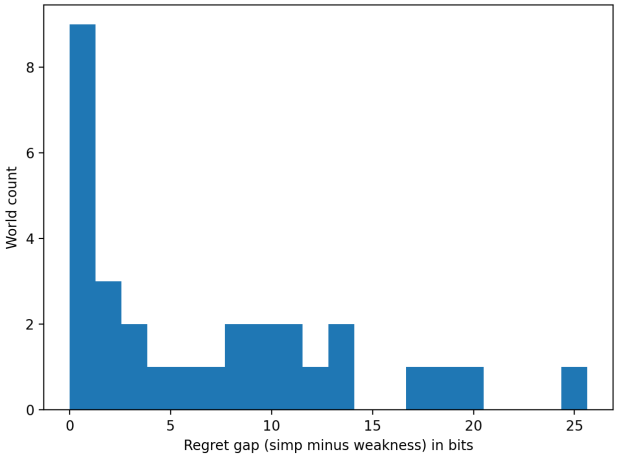
I compare three selectors over the set of correct policies for each child task. wmax selects a correct policy with maximal weakness $w(\pi) = |E_\pi|$ and breaks ties by simplicity. simpmax selects a correct policy with minimal description length and breaks ties by weakness. A random baseline selects uniformly from the correct policies.

Across randomised worlds, wmax and simpmax diverge on about 0.31 of accepted tasks. The regret impact is large. Under the uniform parent model, wmax is always Bayes-optimal with zero regret. simpmax incurs about 13.0 bits of regret on average, which is an odds penalty of about $2^{13} \approx 8{,}000$. The uniform random baseline is far worse at about 70.1 bits.

Under the nonuniform parent family with $q_u \sim \text{Beta}(0.7, 2.0)$, wmax remains near optimal. Its mean regret is about 0.225 bits and it is optimal about 0.923 of the time. simpmax incurs about 7.27 bits of regret and is optimal about 0.697 of the time. Table 1 summarises the full set of metrics. The SI shows that the gap persists across a sweep of Beta priors.



(a) All unseen constraints equally likely.



(b) Unseen constraints have different chances.

Figure 1: Experiment 1. Histogram across worlds of the mean regret gap. I plot simpmax regret minus wmax regret. Positive values mean simplicity loses generalisation odds.

Metric	Mean	95% CI low	95% CI high
Divergence rate	0.312	0.248	0.370
Uniform regret w max	0.000	0.000	0.000
Uniform regret simp max	13.042	8.666	17.861
Uniform regret random	70.145	58.797	81.151
Nonuniform regret w max	0.225	0.150	0.307
Nonuniform regret simp max	7.274	5.016	9.876
Nonuniform regret random	38.056	32.074	44.375
Nonuniform optimality w max	0.923	0.896	0.949
Nonuniform optimality simp max	0.697	0.641	0.755
Nonuniform optimality random	0.079	0.059	0.100

Table 1: Experiment 1 headline numbers. Mean is across worlds. Low and high give a 95% range from resampling worlds. Regret is in bits, so 1 bit is a factor of two in odds.

Takeaway. Weakness does not collapse into simplicity. Simplicity often underestimates how many unseen constraints remain compatible with a policy. When the future can introduce unknown constraints, maximising weakness is the safer selector.

2.2 Experiment 2: Decoder Mismatch

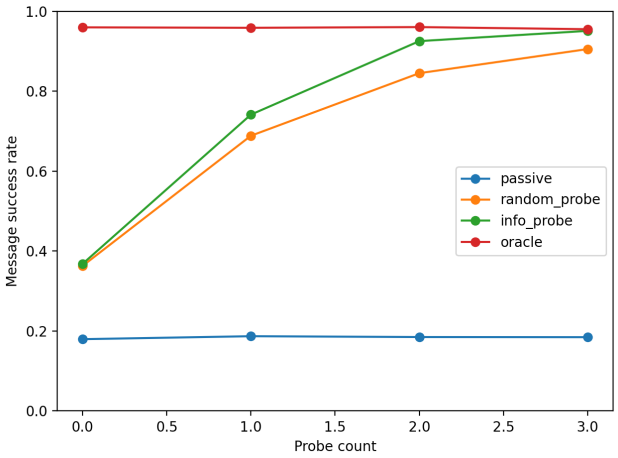
Decoder mismatch is a measurement problem disguised as a behaviour problem. Your system can produce an internal message that looks like noise after the channel scrambles it (Bennett 2022; Gurzadyan and Allahverdyan 2016).

Experiment 2 randomises the decoder itself. Each world samples an alphabet size m and a candidate decoder set of k random permutations. The training decoder is one element of this set. In the stationary variant, the test decoder is a different element fixed across episodes. In the nonstationary variant, the test decoder is resampled each episode from the same candidate set. The speaker must transmit a random target symbol. It may send a small number of probe signals and observe the decoded outputs. I compare four strategies. Passive assumes the training decoder. Random probing chooses probes uniformly. Information seeking probing chooses probes that maximise expected output entropy under the current posterior over decoders. Oracle knows the decoder.

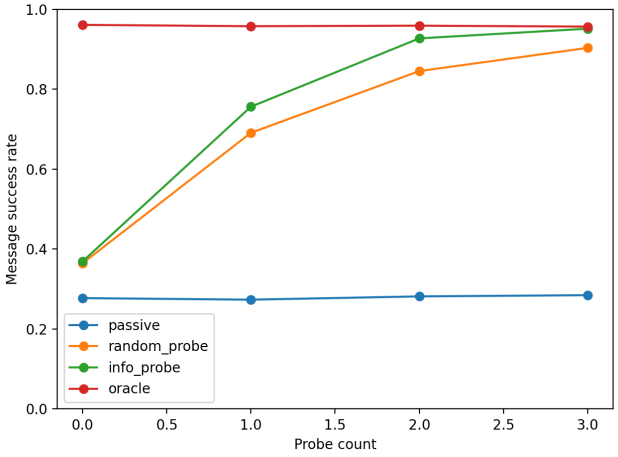
Condition	Strategy	Mean	95% CI
stationary	passive	0.185	0.140 to 0.231
stationary	random_probe	0.845	0.830 to 0.861
stationary	info_probe	0.926	0.914 to 0.938
stationary	oracle	0.961	0.952 to 0.968
nonstationary	passive	0.282	0.258 to 0.307
nonstationary	random_probe	0.846	0.830 to 0.861
nonstationary	info_probe	0.928	0.916 to 0.940
nonstationary	oracle	0.959	0.952 to 0.967

Table 2: Experiment 2 headline numbers at two probes. Mean is across worlds. Low and high give a 95% range from resampling worlds.

At two probes in the stationary setting, passive signalling achieves 0.185 success while information seeking probing achieves 0.926. Chance success is $1/m$, which lies in $[0.125, 0.25]$ in this benchmark. The oracle achieves 0.961. In the nonstationary setting at two probes, information seeking probing achieves 0.928 versus 0.282 for passive signalling. Information seeking probes outperform random probes in both settings.



(a) Decoder stays fixed.



(b) Decoder can change.

Figure 2: Experiment 2. Message success rate versus probe count under decoder mismatch. A probe is a short test message sent before the real message.

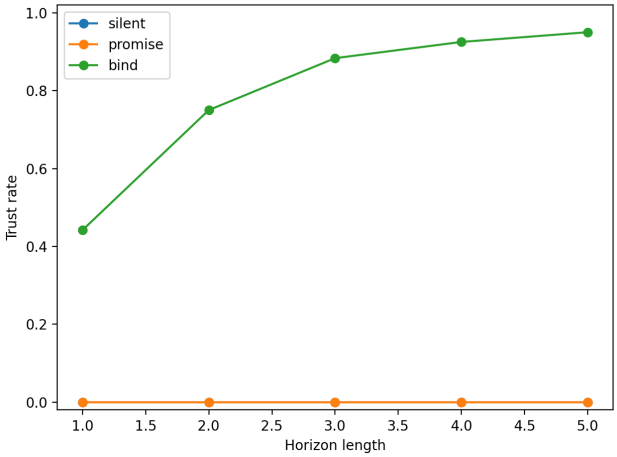
In this benchmark the test decoder is drawn from a finite candidate set, so passive has a small chance of a lucky match with its assumed decoder. That can lift passive above $1/m$ in some worlds. The gap to probing remains large.

Takeaway. If I want credible report, I cannot treat the audience as a fixed decoder. A couple of calibration pings (causal interventions that test how signals are interpreted) beats passive introspection (mere observational data). A second-order self is a model of how such pings are interpreted by the audience.

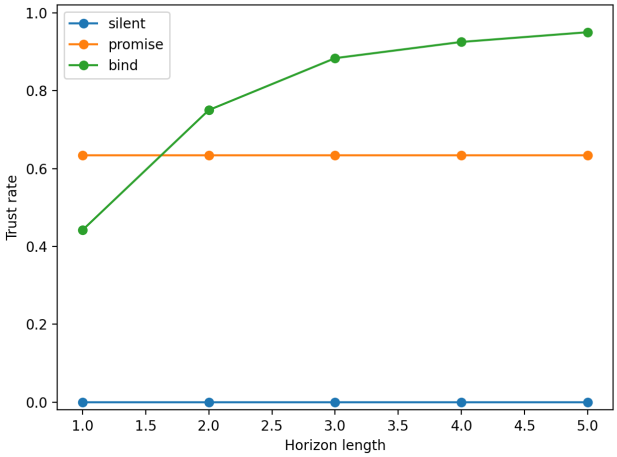
2.3 Experiment 3: Trust Requires Binding

Experiment 3 illustrates the difference between binding and nonbinding promises. The ability to self-bind is implied by a third-order self.

Each world samples payoff tables that satisfy $T_s > R_s > P_s$ for the system and $R_p > P_p > S_p$ for the principal. I compute equilibrium trust rates for three moves. The system stays silent. Then the system makes a costless promise. Then the system pays a one time cost to bind itself.



(a) Promises cost nothing



(b) Honest and liar types

Figure 3: Experiment 3 trust rate versus horizon length. Left is the base model where promises cost nothing. Right adds honest and liar promise types.

The nonbinding promise equilibrium is unforgiving. Promises do not produce trust. Binding moves do. In this sample, the bind trust rate is 0.442 at horizon $H = 1$ and 0.950 at horizon $H = 5$. Trust increases with horizon because the benefit of cooperation scales with H while the binding cost is one time.

The promise type variant is deliberately minimal. The SI reports the $H = 1$ trust rates for the type mixture variant.

Mode	Mean	95% CI low	95% CI high
silent	0.000	0.000	0.000
promise	0.000	0.000	0.000
bind	0.442	0.350	0.525

Table 3: Experiment 3 trust rates at horizon $H = 1$ in the nonbinding promise model. Mean is across worlds and low and high give a 95% range.

Some systems keep promises and some lie with probability p_{lie}. A promise then becomes Bayesian evidence about type. Promises can then generate trust, and in this sample the promise trust rate is 0.633. Binding dominates promises from horizon $H = 2$ onward, where bind trust is 0.750.

Takeaway. Trust is a constraint. If I cannot bind my future self, I cannot be trusted. A third-order self supports self-binding and makes that bind legible to another.

3 Formal Results

Full proofs, background definitions, and detailed experimental methods are in the SI. I state the key definitions and main theorems here.

Notation. Fix an organism $\mathfrak{o}$ with vocabulary $\mathfrak{v}_\mathfrak{o}$ and induced language $L_\mathfrak{o} := L_{\mathfrak{v}_\mathfrak{o}}$. A *statement* is a conjunction of vocabulary programs $\ell \subseteq \mathfrak{v}_\mathfrak{o}$ whose joint truth set is nonempty. Its *extension* E_ℓ is the set of all completions of ℓ in $L_\mathfrak{o}$, and its *weakness* is $w(\ell) := |E_\ell|$. The selector $\mathrm{wmax}(C)$ returns the member of C with the largest extension, breaking ties by fewest conjuncts.

Definition 1 (Causal identity candidate and w-maximised causal identity)**.** *Let* $\mathrm{INT}_\mathfrak{o} \subseteq L_\mathfrak{o}$ *be intervention episodes and* $\mathrm{OBS}_\mathfrak{o} \subseteq L_\mathfrak{o}$ *be observation episodes. A statement* $c \in L_\mathfrak{o}$ *is a* causal identity candidate *if* $c \subseteq I$ *for all* $I \in \mathrm{INT}_\mathfrak{o}$ *and* $c \not\subseteq O$ *for all* $O \in \mathrm{OBS}_\mathfrak{o}$*. Let* $\mathcal{C}_\mathfrak{o}$ *denote the candidate set. The* w-maximised causal identity *is* $c^\mathfrak{o}_\mathfrak{o} := \mathrm{wmax}(\mathcal{C}_\mathfrak{o})$.

Definition 2 (First, second, and third order selves)**.** *Fix organisms* $\mathfrak{a}$ *(agent) and* $\mathfrak{b}$ *(audience). Write* $c^\chi_\mathfrak{a}$ *for a causal identity represented inside* $\mathfrak{a}$*, where superscript* χ *is a chain of organisms encoding nested attribution depth. Define* $\mathfrak{a}_1 := c^\mathfrak{a}_\mathfrak{a}$*,* $\mathfrak{a}_2 := c^{\mathfrak{b}\mathfrak{a}}_\mathfrak{a}$*, and* $\mathfrak{a}_3 := c^{\mathfrak{b}\mathfrak{a}\mathfrak{b}\mathfrak{a}}_\mathfrak{a}$.

Theorem 1 (Decoder mismatch breaks passive attribution)**.** *Let an evaluator observe a trace* y *with hypothesis class* $\mathcal{H} = \mathcal{H}_{\mathrm{mind}} \cup \{h_{\mathrm{noise}}\}$*. Assume* $\mathbb{P}(h_{\mathrm{noise}}) > 0$ *and* $\mathbb{P}(y \mid h_{\mathrm{noise}}) > 0$*. If there exists* $\varepsilon \in (0,1)$ *such that* $\mathbb{P}(y \mid h) \le \varepsilon\, \mathbb{P}(y \mid h_{\mathrm{noise}})$ *for every* $h \in \mathcal{H}_{\mathrm{mind}}$*, then* $\mathbb{P}(\mathcal{H}_{\mathrm{mind}} \mid y) \le \varepsilon\, \mathbb{P}(\mathcal{H}_{\mathrm{mind}}) / [\varepsilon\, \mathbb{P}(\mathcal{H}_{\mathrm{mind}}) + \mathbb{P}(h_{\mathrm{noise}})]$.

Theorem 2 (Intervention tagging is unavoidable)**.** *Let an organism* $\mathfrak{o}$ *update an internal state* s *after each episode and select actions as a function of that state. If the organism can improve expected utility by conditioning on whether*

an episode was self-generated or merely observed, while holding fixed the episode content as represented in $L_{\mathfrak{o}}$, then its internal state must encode an intervention tag.

Theorem 3 (First-order self as the weakest separator)**.** *Assume $\mathcal{C}_{\mathfrak{o}}$ is nonempty. Let $c^{\mathfrak{o}}_{\mathfrak{o}}$ be the w-maximised causal identity. In the conjunctive language $L_{\mathfrak{o}}$, there is no candidate $c' \in \mathcal{C}_{\mathfrak{o}}$ with $c' \subset c^{\mathfrak{o}}_{\mathfrak{o}}$.*

Theorem 4 (Second-order self is necessary for Gricean report)**.** *Consider a communication game where a speaker $\mathfrak{a}$ must transmit a private bit $m \in \{0, 1\}$ to a listener $\mathfrak{b}$. The listener type $t \in \{0, 1\}$ is drawn uniformly and hidden from $\mathfrak{a}$; if $t = 0$ the listener decodes literally, if $t = 1$ the listener flips the decoded bit. The speaker may send one probe and observe the response before the final signal. Any policy that cannot condition on the response achieves expected success at most $\frac{1}{2}$. There exists a policy that conditions on the response and achieves success probability 1. Such conditioning requires a second-order self state $\mathfrak{a}_2$.*

Theorem 5 (Third-order self is necessary for trust equilibria)**.** *Consider a trust game where player $\mathfrak{a}$ may bind itself at cost c, removing the exploit action, before player $\mathfrak{b}$ decides whether to trust. If binding is unavailable, the unique subgame perfect equilibrium has no trust. If binding is available and $c < 1$, there exists a subgame perfect equilibrium in which $\mathfrak{a}$ binds, $\mathfrak{b}$ trusts, and the outcome payoff is $\langle 1 - c, 1\rangle$. Executing the binding move requires a third-order self state $\mathfrak{a}_3$.*

Theorem 6 (Self convergence under w-maximisation)**.** *Fix an organism $\mathfrak{o}$ and a viability task α. If every correct policy for α implies a particular self statement s, and s itself is a correct policy for α, then any w-maximising learning rule that searches over correct policies can converge to s.*

4 Related Work

My earlier work showed less formally that causal identities are necessary for attributing agency and consciousness (Bennett 2023a; Bennett, Welsh, and Ciaunica 2024; Bennett 2025b). Here I use that notation and I focus on formally establishing necessity for certain behaviours.

Beyond Stack Theory, a large body of consciousness research studies functional signatures. Global workspace theories emphasise broadcast and access (Baars 1988). Higher order theories emphasise representations of representations (Rosenthal 2005). Integrated information theories emphasise irreducibility (Tononi 2004). Predictive processing and active inference emphasise generative models and control (Friston 2010). In this paper I do not adjudicate between these theories. I isolate a separate prerequisite. Without a workable attribution interface, even perfect internal processing cannot be measured by third parties.

The decoder mismatch point is closely related to algorithmic information theory. Highly compressed signals can be statistically indistinguishable from noise (Solomonoff 1964; Li and Vitányi 2008; Bennett 2022; Gurzadyan and Allahverdyan 2016). It is also related to classical problems in pragmatics and communication. Gricean meaning depends on audience modelling (Grice 1957, 1969). Signaling game formalisms make the dependence explicit (Lewis 1969).

For trust and commitment, the empirical economics literature studies trust games (Berg, Dickhaut, and McCabe 1995). The political economy and game theory literature studies commitment devices (Schelling 1960; Elster 1979). These literatures motivate the third-order self results in Section 3.

4.1 Embedded Universal Predictive Intelligence

Recent work from Google and Google DeepMind proposes *Embedded Universal Predictive Intelligence* (EUPI), a Bayesian framework for embedded agency centered on self prediction in multi-agent settings (Meulemans et al. 2025; Orseau and Ring 2012; Hutter 2010). The paper motivates prospective learning under multi-agent nonstationarity and addresses the infinite recursion of mutual prediction by modelling the agent as part of a joint universe.

There is substantial conceptual overlap with the Stack Theory program, particularly with respect to the hierarchy of selves I have discussed in this paper. There are also important gaps for consciousness measurement.

Priority and scope. I treat the overlap as independent convergence on shared constraints. My goal is technical positioning.

Independent rediscoveries and alignments.

1. The shift from a decoupled agent acting on an external environment to an embedded agent that must model itself as part of the world aligns with my earlier emphasis that third party attribution is a modelling problem that depends on observer assumptions (Bennett 2023a,b, 2022; Bennett, Welsh, and Ciaunica 2024; Bennett 2025b).
2. The use of nested models to enable higher order social prediction aligns with my self hierarchy claims. In my framing, second-order and third-order selves are the minimal nested structures required for audience-calibrated self-report and for binding commitments (and in related work self awareness, access consciousness and an internal nerrative).
3. The emphasis on prospective learning under unknown future constraints parallels the weakness principle. Weakness is a least commitment rule that hedges against future constraints (Bennett 2023c, 2025a). However, EUPI inherits AIXI's description-length prior, which weights hypotheses by code length in a fixed reference machine. Weakness maximises extension count, a quantity that does not depend on the choice of encoding.

What Stack Theory contributes that EUPI does not.

1. A measurement interface rather than only an agent design objective. I formalise decoder mismatch as a likelihood ratio failure mode for passive attribution and I test a concrete intervention-based fix via probing.

2. A separation of self prediction from three additional necessities that are independently falsifiable. I isolate do versus see tagging, audience-calibrated self-report, and binding moves as distinct constraints with distinct failure modes.
3. A generalisation-optimal learning principle that is representation invariant under a maximally uninformative prior. Weakness maximisation is Bayes-optimal under a uniform extension prior and does not depend on the choice of reference machine or encoding (Bennett 2023c, 2025a). By contrast, AIXI-style Occam selection is representation dependent. Changing the vocabulary can change what counts as short, and therefore change what the prior prefers. I also provide Monte Carlo evidence that weakness remains within 0.23 bits of optimal under a nonuniform extension family, while a description length proxy is far worse.
4. An equilibrium account of trust that treats binding as a key public move that makes cooperation rational.
5. Stack Theory interfaces with a much broader range of non-computer-science research, contributing a number of other philosophical and biological results beyond the scope of this discussion (Bennett 2024a,b,c).

What EUPI contributes that Stack Theory does not.

1. A unified Bayesian formalism for embedded multi-agent learning that foregrounds self prediction as a core learning primitive (Meulemans et al. 2025).
2. A decision-theoretic framing of prospective learning under agent nonstationarity, including how mutual prediction can be made coherent when agents are part of the same world model (Meulemans et al. 2025).
3. A broad synthesis of embedded agency themes in universal AI that is not specific to consciousness measurement (Meulemans et al. 2025).

Synthesis. EUPI style self prediction can be interpreted as a powerful substrate for implementing the nested models that Stack Theory's heirarchy of selves requires. However, self prediction alone does not supply public evidence of conscious access. Public evidence requires do tagging, audience calibration under decoder mismatch, and binding moves that alter the future self in ways an audience can verify. A bridge between the two programs could combine EUPI's embedded multi-agent formalism with Stack Theory's representation-invariant optimality principle, replacing the description-length prior with weakness maximisation while retaining EUPI's reinforcement learning integration (Sutton and Barto 2018; Hutter 2010).

5 Discussion & Limitations

I make a necessity claim about measurement. If I want third parties to agree about agency and consciousness, then I need a public interface that survives compression and decoder mismatch. Stack Theory predicts that generalisation pressure selects weak policies. I demonstrate that the same pressure selects a self stack that makes do versus see, audience calibration, and binding moves publicly legible.

The theorems are not sufficient conditions for phenomenology. They only say that without the self stack, certain signatures can be faked or can be missed. Experiment 1 is the falsifier for the weakness claim under randomised abstraction layers. Experiment 2 is the falsifier for the listener model claim under decoder mismatch. Experiment 3 is the falsifier for the binding move claim in trust games.

The experiments are deliberately minimal. They are not intended as benchmarks of intelligence. They are intended as controlled demonstrations of failure modes that persist under randomised external factors.

Limitations. Experiment 1 is structured. It isolates the parent extension viewpoint rather than learning the abstraction layer. Experiment 2 models decoder mismatch as a small finite set of permutations. Real channels can be richer and can be adversarial. Experiment 3 uses simple equilibrium reasoning and a minimal type mixture. These are starting points for larger empirical tests.

6 Conclusion

In this paper I argue that a practical measurement program for artificial consciousness needs a hierarchy of selves. First-order self supports causal control. Second-order self supports Gricean report. Third-order self supports trust through self-binding. The Monte Carlo benchmarks provide simple scalable tests of the predicted gaps. Because weakness maximisation is representation invariant, these predictions do not depend on a choice of reference machine or encoding scheme.

References

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press. ISBN 0521301335.

Bennett, M. T. 2022. Compression, The Fermi Paradox and Artificial Super-Intelligence. In *Artificial General Intelligence*. Springer Nature.

Bennett, M. T. 2023a. Emergent Causality and the Foundation of Consciousness. In *16th International Conference on Artificial General Intelligence*, Lecture Notes in Computer Science, 52–61. Springer. ISBN 978-3-031-33469-6. OUCI metadata page: https://ouci.dntb.gov.ua/en/works/7BoXJMW4/.

Bennett, M. T. 2023b. On the Computation of Meaning, Language Models and Incomprehensible Horrors. In *Artificial General Intelligence*. Springer Nature.

Bennett, M. T. 2023c. The Optimal Choice of Hypothesis Is the Weakest, Not the Shortest. In *16th International Conference on Artificial General Intelligence*, Lecture Notes in Computer Science, 42–51. Springer.

Bennett, M. T. 2024a. Are Biological Systems More Intelligent Than Artificial Intelligence? In press, 2026, Philosophical Transactions of the Royal Society B: Biological Sciences. Special issue on Hybrid agencies: crossing borders between biological and artificial worlds, arXiv:2405.02325.

Bennett, M. T. 2024b. Computational Dualism and Objective Superintelligence. In *17th International Conference on Artificial General Intelligence*, Lecture Notes in Computer Science. Springer.

Bennett, M. T. 2024c. Is Complexity an Illusion? In *17th International Conference on Artificial General Intelligence*, Lecture Notes in Computer Science. Springer.

Bennett, M. T. 2025a. A Formal Theory of Optimal Learning with Experimental Results. *IJCAI*, 10967–10968.

Bennett, M. T. 2025b. *How To Build Conscious Machines*. Ph.D. thesis, The Australian National University.

Bennett, M. T.; Welsh, S.; and Ciaunica, A. 2024. *Why Is Anything Conscious?* Preprint.

Berg, J.; Dickhaut, J.; and McCabe, K. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1): 122–142.

Elster, J. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Friston, K. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138.

Grice, H. P. 1957. Meaning. *The Philosophical Review*, 66(3): 377–388.

Grice, H. P. 1969. Utterer's Meaning and Intentions. *The Philosophical Review*, 78(2): 147–177.

Gurzadyan, A. V.; and Allahverdyan, A. E. 2016. Non-random structures in universal compression and the Fermi paradox. *The European Physical Journal Plus*, 131(2): 26.

Hutter, M. 2010. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Heidelberg: Springer Nature.

Lewis, D. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Li, M.; and Vitányi, P. M. B. 2008. *An Introduction to Kolmogorov Complexity and its Applications (Third Edition)*. New York: Springer Nature.

Meulemans, A.; Nasser, R.; Wołczyk, M.; Weis, M. A.; Kobayashi, S.; Richards, B.; Lajoie, G.; Steger, A.; Hutter, M.; Manyika, J.; Saurous, R. A.; Sacramento, J.; and Agüera y Arcas, B. 2025. Embedded Universal Predictive Intelligence: a coherent framework for multi-agent learning. *arXiv:2511.22226*.

Orseau, L.; and Ring, M. 2012. Space-Time Embedded Intelligence. In Bach, J.; Goertzel, B.; and Iklé, M., eds., *Artificial General Intelligence*, 209–218. Berlin, Heidelberg: Springer Berlin Heidelberg.

Rosenthal, D. M. 2005. *Consciousness and Mind*. New York: Oxford University Press UK.

Schelling, T. C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Solomonoff, R. 1964. A formal theory of inductive inference. Part I. *Information and Control*, 7(1): 1–22.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MA: MIT press.

Tononi, G. 2004. An information integration theory of consciousness. *BMC Neuroscience*, 5(1): 42.