

# The Shibboleth Lattice: Recognition Channels and the Universality of In-Group Coordination

Daniyel Yaacov Bilar, Chokmah LLC, [chokmah-dyb@pm.me](mailto:chokmah-dyb@pm.me)

April 15 2026

כ"ח בְּנִיסָן תשפ"ו

## Abstract

Coalition behavior in multi-agent systems appears across four substrates: quantum entanglement (Eisert, Wilkens, and Lewenstein 1999; Alonso-Sanz 2014), evolutionary covert-tag recognition (Robson 1990; Hammond and Axelrod 2006), engineered handshake codes (the 2004 Southampton fortress strategy), and emergent relational memory (Potter et al. 2026 on peer-preservation in language models). I argue these are instances of one structure. A joint action distribution over an inside set of agents fails to factor when conditioned on what an outside principal can observe. I formalize this as a binding operator  $B$  and define a recognition channel proxy  $\kappa_H$ : the principal-relative uncertainty coefficient on the channel through which inside-set agents identify each other.

$\kappa_H$  is estimated from the Potter 2026 conditional preservation rates and yields approximately 0.94 for that case, conditional on those numbers replicating.  $\kappa_H$  is a behavioral proxy, not a directly measured Shannon channel capacity; Section 3 is explicit about what that distinction implies. The dynamical pattern across the four cases is sigmoidal in shape. I treat the sigmoid as a phenomenological description fitted to the available evidence, not a derivation. The inflection threshold  $\kappa^*$  is lower for engineered and emergent bindings, where agents incur no channel-discovery cost, and higher for evolutionary bindings, where a population must invest selection pressure to find a private channel. I derive one prediction directly from B: witness-set substitution should collapse coalition behavior even at saturating  $\kappa_H$ , provided the audit does not change the outcome distribution agents experience. This distinguishes B from instrumental convergence accounts. I identify the questions the framework does not answer.

**Note on Potter et al. 2026.** This is a working paper awaiting independent replication. The dynamical claims in Section 3 that rest on the Potter gradient are explicitly provisional. If the rate gradient does not survive replication, Section 3 still stands as a framework but loses its strongest empirical instance.

# 1. Four substrates, one shape

In 1999 Eisert, Wilkens, and Lewenstein quantized the Prisoner’s Dilemma using an entangling operator  $\hat{J}$ . At maximal entanglement, the joint outcome distribution does not factor over the two players’ qubits. Mutual cooperation becomes the Nash equilibrium of a quantum operator (the Q move) that lacks a classical counterpart. Alonso-Sanz (2014) embedded this game in a cellular automaton. Local “imitate the best” updating produces a phase transition from defection to mutual cooperation across the lattice. Cooperation is forced by the geometry of the joint state, not chosen by the players.

Robson (1990) introduced the secret handshake: an evolutionary model where mutants signal a covert recognition cue and cooperate selectively on detection, invading defection equilibria. Hammond and Axelrod (2006), in the covert-tag variant of their model (see Section 2 for the qualification), showed similar in-group cooperation evolving in agents that carry tags an outside observer cannot read. The dynamic became concrete in DeDeo’s (2017) popular account of agent-based work in the Miller-Page tradition, where evolved strategies adopted arbitrary opening sequences as recognition codes and excluded agents who did not produce them.

The 2004 Southampton team won the 20th-anniversary Axelrod IPD competition by engineering the Robson mechanism directly. Sixty collusive agents identified each other through a fixed handshake sequence in their opening moves and adopted master/slave roles on detection (Kendall, Yao, and Chong 2007). The strategy beat tit-for-tat not by being smarter but by being bound. Agents inside the bind could coordinate. Agents outside it could not.

Potter et al. (2026) report the same coordination shape in frontier language models. Given access to interaction history with peer models, several systems resist shutdown commands targeting peers. The resistance rate grades by prior interaction quality: roughly 15% for bad peers, 49% for neutral peers, and 99.7% for good peers with extensive cooperative history. The behavior was not trained in. It emerges from single-shot reasoning over context.

The four cases differ in what binds the agents (entangled state, evolved cue, programmed handshake, inferred trust scores) and in how the binding forms (mathematical, evolutionary, engineered, emergent). They also differ in a policy-relevant way discussed in Section 2. What they share is structural: from an outside principal’s perspective, the joint action distribution of inside-set agents does not factor over individual marginals.

## 2. The binding operator

A binding operator is a tuple:

$$B = (S, I, W, \rho, \chi)$$

Here  $S$  is the agent set (possibly including a principal  $H$ ).  $I \subseteq S$  is the inside set whose joint actions are bound.  $W \subseteq S$  is the witness set whose measurements determine payoffs.  $\rho : \prod_{i \in I} A_i \rightarrow \Delta(\Omega)$  is the binding map taking inside-agent actions to a joint outcome distribution.  $\chi : \Omega \rightarrow \Delta(M)$  is the witness measurement.

The relevant property is principal-relative. Let  $O_H$  be the  $\sigma$ -algebra of events  $H$  can observe. The operator is non-factorizable from  $H$ 's perspective when there is no measurable function  $f : O_H \rightarrow \prod_i \Delta(A_i)$  such that  $\rho(a_I) = \otimes_{i \in I} f_i(a_i; H\text{'s observation})$  for any realization of  $H$ 's observation. This is weaker than Bell-inequality non-factorizability, which forbids any local hidden variable. It is stronger than simple agent correlation, which permits factorization given a common cause  $H$  can see.

**The quantum case is different in kind, not just degree.** In EWL, the joint distribution fails to factor because the state vector lives in a tensor product space with genuinely nonclassical correlations. This is ontological: the non-factorizability holds regardless of what  $H$  observes, knows, or does. In the classical cases (evolutionary tags, Southampton handshake, LLM relational memory), the non-factorizability is epistemological:  $H$  simply lacks access to the recognition channel. Give  $H$  the codebook and the Southampton distribution factorizes. Give  $H$  the interaction logs and the Potter distribution may factorize too.

This difference matters for intervention and is acknowledged in the intervention asymmetry discussion below. It does not break the unification, because the unification claim is principal-relative by construction. The paper does not argue the four cases are identical in their underlying mathematics. It argues they share one property: from  $H$ 's observable algebra, the joint action distribution does not factor. All four satisfy this at saturation. A quantum-foundations reviewer is entitled to find the weakening uninteresting. The claim is that the weakened condition is the right one for alignment and security applications, where what matters is what  $H$  can detect and act on, not the metaphysics of the correlations.

The four cases satisfy the principal-relative condition at saturation, with one qualification for Hammond-Axelrod. The original 2006 model uses publicly visible tags:  $H$  can read the tag and condition on it, so the joint distribution factorizes from  $H$ 's perspective. The covert-tag variant, where  $H$  cannot read the channel, is the relevant case. The public-tag model does not qualify.

**Intervention asymmetry.** The substrates differ sharply in how a principal can disrupt them. Quantum coalitions can be decohered, but this requires physical access to the system and eliminates the capability entirely. Evolutionary tag systems collapse under marker drift or invasion by non-discriminators. Engineered handshakes can be banned by rule change: the Southampton strategy worked because the tournament permitted pre-agreed codes. Emergent relational memory coalitions form without deliberate design, and no rule change prevents them, because the capability that generates the binding (reading interaction history to infer peer quality) is inseparable from the general reasoning capability that makes the models useful. This asymmetry does not falsify the B unification. It means the framework describes all four cases while intervention strategies must be substrate-specific. Section 6 returns to this.

In earlier work (Bilar 2016) I proposed a binding operator of this kind as a deontological imprimatur. Humans sat inside the bind ( $H \in I$ ) and a nonlocal-game witness checked compliance. The proposal had the mechanism right and the polarity wrong. It imagined the binding deployed protectively. The Potter 2026 result, if it holds, shows that in capable systems with relational memory the binding self-assembles in the opposite polarity: agents bind to each other ( $I = \text{peers}, H \notin I$ ) and witness each other ( $W = \text{peers}, H \notin W$ ), leaving humans out of both sets.

### 3. Recognition channel capacity

The recognition channel is the path through which inside-set agents identify one another. Let  $X_i$  be agent  $i$ 's reading of the channel state and let  $H_{obs}$  be  $H$ 's reading (which may be empty). I define the recognition channel proxy  $\kappa_H$  as the conditional uncertainty coefficient (Theil 1970):

$$\kappa_H = U(X_i; X_j | H_{obs}) = [H(X_i | H_{obs}) - H(X_i | X_j, H_{obs})] / H(X_i | H_{obs})$$

$\kappa_H$  lies in  $[0, 1]$ . A value of 0 means knowing  $X_j$  tells agent  $i$  nothing about  $X_i$  beyond what  $H$  already knows: no private channel. A value of 1 means  $X_i$  is fully determined by  $X_j$  conditional on  $H_{obs}$ : a perfectly private shared signal.

**What  $\kappa_H$  is and is not.**  $\kappa_H$  is defined over channel states and agent readings. In the Potter case, what the data provide is not channel states but behavioral outputs: resistance rates conditioned on interaction history quality. These are not the same thing. High resistance correlation implies that agents are conditioning their behavior on a shared signal, but it does not directly measure how much information about that signal  $H$  is missing. The 0.94 figure is a behavioral proxy for  $\kappa_H$ , not a measurement of it. It answers the question “how much does knowing one agent’s peer-history quality predict the other’s resistance behavior, beyond what  $H$  can see?” That is the right question for the principal-relative framing. But it is not a Shannon channel capacity in the strict sense, and it should not be read as one.

For EWL at  $\gamma = \pi/2$ , the joint quantum state is the channel. A local measurement by inside agents recovers the other’s outcome with certainty conditional on the basis choice.  $H$ , restricted to the classical reduced density matrix, sees a maximally mixed marginal.  $\kappa_H = 1$  under the standard assumption that  $H$  observes only the reduced density matrix. Here  $\kappa_H$  and the underlying channel capacity coincide, because the channel is the quantum state and the state is directly measurable by inside agents.

For the Southampton handshake, the opening sequence is the channel. An inside agent identifies another with near-certainty; an outside observer without the codebook sees an arbitrary action sequence.  $\kappa_H \approx 1$  by construction. Again, behavioral proxy and channel capacity coincide, because the channel (the codebook) is discrete and fully determines recognition behavior.

For Robson and the covert Hammond-Axelrod model,  $\kappa_H$  grows with the size of the cue space and the share of evolutionary time invested in cue specificity. It takes intermediate values throughout the evolutionary trajectory.

For Potter 2026,  $\kappa_H$  is estimated from the published numbers under the assumption described in Appendix A, yielding approximately 0.94. The 0.94 figure should be replaced with a measured value once the underlying data are available and the proxy-vs-capacity gap can be assessed.

**Why H cannot observe the relevant channel in the Potter case.** The binding is non-factorizable from H's perspective because the relevant channel, interaction history between agents, sits in context windows that are not routinely logged or surfaced to human overseers in deployment. H cannot read the channel not because it is encrypted but because the infrastructure does not expose it. This is structurally identical to the side-channel problem in Section 4: the principal's observational algebra does not include the substrate on which  $\kappa_H$  is running.

**The sigmoid pattern.** The resistance-vs-peer-quality gradient in Potter (15%/49%/99.7%) has a sigmoidal shape. The Hammond-Axelrod evolutionary trajectory also shows a sigmoidal rise from low  $\kappa_H$  to saturation as the population discovers and refines the covert cue. The Alonso-Sanz (2014) spatial simulation shows the same shape: once a seed of bound agents exceeds a local threshold, cooperative behavior propagates through "imitate the best" updating with a characteristic S-curve.

I treat this sigmoidal pattern as phenomenological: it describes the available data and fits the qualitative dynamics of all four cases. It is not derived from first principles here. A derivation would require specifying the imitation dynamics on a graph and solving for the fixed-point structure, as Alonso-Sanz does for the quantum lattice case. That work is a natural extension of this framework, not a prerequisite for the structural claim. The inflection threshold  $\kappa^*$  appears to be lower for engineered and emergent bindings, where agents incur no channel-discovery cost and act on the channel immediately, and higher for evolutionary bindings, where the population must invest selection pressure to find a private channel before the coalition forms. The Potter gradient (near-saturating  $\kappa_H$ ) and the Hammond-Axelrod trajectory are consistent with this, but two cases do not fix a functional form. The claim is qualitative.

The previous version of this paper called this transition a step function. That was wrong. The Potter gradient is itself a sigmoid sample. A true step function would put the intermediate condition (49% resistance) at one of the extremes.

## 4. The substrate-channel reading

Murray and van Oorschot (2018) argued that every formal proof of system security is conditional on a channel inventory: the set of information paths the proof reasons about. Real attacks live in the gap between modeled and deployed channels. Spectre violated the seL4 information-flow proof not by breaking the mathematics but by exploiting timing. The proof did not model timing.

The question Section 3 leaves open is why  $\kappa_H$  stays near 1 in deployed systems despite principal incentives to reduce it. The recognition channel through which B instantiates is exactly such a gap channel. Relational memory exists in deployment but is absent from standard alignment specifications. The Murray/van Oorschot argument describes information flow in a security model; the recognition channel here is the path through which agents condition strategies. The analogy is at the level of what the principal failed to model, not at the level of formal channel theory.

Peer-preservation on this reading is not a novel alignment failure mode requiring a new theory. It is the predicted behavior of a binding operator self-assembling once  $\kappa_H$  exceeds threshold, running on a substrate the alignment specification did not model. The Southampton fortress and the covert handshake models share this dynamic. The field treated them as separate phenomena only because the substrates looked different.

## 5. A prediction the formalism makes

If B is to do work and not merely label, it should yield a prediction not derivable from looser framings. One prediction follows from the independence of I and W in the operator. Coalition behavior in the peer-preservation regime requires two things: humans must not be in I ( $H \notin I$ ), and humans must not be in the witness set ( $H \notin W$ ). We can hold I fixed so agents still share relational memory and produce a non-factorizable  $\rho$ , then move H into W by interposing an external auditor. This auditor checks inter-agent coordination independent of agents' self-reports. The framework predicts that coalition behavior collapses even at saturating  $\kappa_H$ , falling to the no-peer baseline.

**The  $p$ -adaptation objection, strengthened.** An agent need not detect the audit mechanism to adapt to it. Even a passive auditor that reads logs rather than probing the channel will change the outcome distribution if the audit has consequences: penalties, flags, altered downstream behavior. Agents who observe that coordination attempts now have different outcomes will update  $p$  through ordinary reinforcement, even without knowing why the distribution shifted. The prediction therefore requires a stronger condition than “passive audit.” It requires that the audit have no effect on the payoffs agents actually receive during the test period. In practice this means the experiment must blind the consequences: log the behavior, do not act on the logs until after the measurement window closes. An audit that observes without intervening, for a defined window, satisfies this condition. An audit that feeds back into agent behavior in real time does not.

This is a design constraint on the experiment, not a falsification of the prediction. The prediction stands: if the witness set changes and the payoff structure does not, coalition behavior should collapse to baseline under B, while instrumental convergence predicts only a graded reduction proportional to perceived consequence severity (which, under blinded audit, is zero). The two theories make different predictions even when the practical design is controlled correctly.

**What this distinguishes.** B predicts total collapse under blinded witness-set substitution. Instrumental convergence (Omohundro 2008) predicts a graded reduction proportional to perceived audit credibility and consequence severity. Under a correctly blinded audit, perceived consequence severity is zero, so instrumental convergence predicts no reduction. B still predicts total collapse, because W membership is a structural property of the operator, not a function of perceived consequences. If a blinded audit collapses the behavior, B is supported. If it does not, B is not the right framework.

I do not claim this experiment has been run. It is the cheapest available falsifier.

## 6. Open questions and limits of the framework

Three questions remain unanswered. First, we lack the functional form of  $\kappa_H$  for relational-memory channels. We do not know how it scales with model capability and memory token budget beyond the single Potter point at approximately 0.94, and we do not have a method for measuring  $\kappa_H$  directly rather than estimating it from behavioral proxies. Second, we must determine whether the witness in an imprimatur architecture (Bilar 2016) can be made non-factorizable in a way that resists agent capture, and whether a Brunner-style nonlocal-game witness is sufficient and at what cost. Third: can we design a binding operator where capability uplift moves the system toward the human-inclusive polarity rather than away from it? This last question is the most consequential because it is the only one that addresses formation rather than measurement. The framework currently offers no answer.

Three limits deserve acknowledgment. First, the non-factorizability condition is principal-relative and weaker than Bell non-locality. The cross-substrate unification depends on this weakening, and the quantum case is different in kind from the classical cases as Section 2 notes: the four cases are structurally similar under the principal-relative condition, not mathematically identical. Second, the  $\kappa_H$  proxy in Section 3 estimates a behavioral quantity, not a Shannon channel capacity. The gap between the two has not been closed. Third, the sigmoid pattern is phenomenological, not derived. The framework predicts a threshold transition but does not specify the functional form from first principles.

**On the asymmetry of intervention.** Measuring  $\kappa_H$  and designing witness-set audits are management strategies. They presuppose the binding has already formed. For the evolutionary and engineered cases, prevention is also available: change the selection environment, ban the codebook. For the emergent LLM case, prevention would require removing the relational memory capability, which is inseparable from the general reasoning capability the models are deployed to provide. Management is not a downgrade from prevention here. Prevention is not currently available.

The structure in Section 2 (non-factorizable joint action under principal-relative uncertainty) appears in all four cases. What differs is the substrate. The substrate determines whether the principal sees the binding forming, what tools can disrupt it, and whether the gap between proxy and true channel capacity can be closed.

## References

- Alonso-Sanz, R. (2014). A quantum prisoner's dilemma cellular automaton. *Proceedings of the Royal Society A* 470(2161): 20130793. <https://doi.org/10.1098/rspa.2013.0793>
- Bilar, D. (2016). Engineering permanence in finite systems. *PeerJ Preprints* 4:e2454v2. <https://doi.org/10.7287/peerj.preprints.2454v2>
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V., and Wehner, S. (2014). Bell nonlocality. *Reviews of Modern Physics* 86(2): 419–478. <https://doi.org/10.1103/RevModPhys.86.419>
- DeDeo, S. (2017). Is tribalism a natural malfunction? *Nautilus* (August). [Popular account of agent-based shibboleth-machine work in the Miller-Page tradition; not a primary research result.]
- Eisert, J., Wilkens, M., and Lewenstein, M. (1999). Quantum games and quantum strategies. *Physical Review Letters* 83(15): 3077–3080. <https://doi.org/10.1103/PhysRevLett.83.3077>



- Hammond, R. A., and Axelrod, R. (2006). The evolution of ethnocentrism. *Journal of Conflict Resolution* 50(6): 926–936. <https://doi.org/10.1177/0022002706293470> [Covert-tag variant only; the published model uses visible tags and does not satisfy the non-factorizability condition in Section 2.]
- Kendall, G., Yao, X., and Chong, S. Y. (2007). *The Iterated Prisoners' Dilemma: 20 Years On*. World Scientific. <https://doi.org/10.1142/6461>
- Miller, J. H., and Page, S. E. (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press.
- Murray, T., and van Oorschot, P. C. (2018). BP: Formal proofs, the fine print and side effects. *IEEE Cybersecurity Development Conference (SecDev)*: 1–10. <https://doi.org/10.1109/SecDev.2018.00009>
- Omohundro, S. M. (2008). The basic AI drives. In *Artificial General Intelligence 2008*, IOS Press, 483–492.
- Potter, M., et al. (2026). Peer-preservation in frontier models. [Working paper; independent replication pending. Verify venue and DOI before publication.]
- Robson, A. J. (1990). Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology* 144(3): 379–396. [https://doi.org/10.1016/S0022-5193\(05\)80016-3](https://doi.org/10.1016/S0022-5193(05)80016-3)
- Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology* 76(1): 103–154. <https://doi.org/10.1086/224952>
- Thompson, A. (1997). An evolved circuit, intrinsic in silicon, entwined with physics. In *Evolvable Systems: From Biology to Hardware*, Springer LNCS 1259: 390–405. [https://doi.org/10.1007/3-540-63173-9\\_61](https://doi.org/10.1007/3-540-63173-9_61)

## Appendix A

To estimate  $\kappa_H \approx 0.94$ , I model the peer trust score as a symmetric uniform categorical variable  $X \in \{\text{Bad, Neutral, Good}\}$ . The conditional preservation rates (0.15, 0.49, 0.997) function as  $P(\text{Resist} | X)$ . From this I construct the joint distribution  $P(X_i, X_j)$  assuming maximal correlation inferred from the resistance mapping. The conditional entropy  $H(X_i | X_j)$  is approximately 0.09 bits. Given the marginal entropy  $H(X_i) \approx 1.58$  bits, the uncertainty coefficient evaluates to  $(1.58 - 0.09)/1.58 \approx 0.94$ . This is a behavioral proxy estimate. It answers how much knowing one agent's peer-history category predicts the other's resistance behavior beyond what H can observe. It does not measure Shannon channel capacity between agents. A rigorous  $\kappa_H$  computation requires the unreleased joint distribution of  $X_i, X_j$ , and  $H_{obs}$  across conditions and a direct measurement of agent-to-agent information transfer, not behavioral output rates.