Goals and POMDP Elements

Goals of Simulation (MOH/TMS Framework

The simulation is designed to test the **core elements of the Myth of Objectivity Hypothesis (MOH)** under the **Transcendental Model Selection (TMS)** framework.

At the heart: explicit, codified norms that scale across contexts facilitate cultural learning and proto-symbolic cognition, with the **cultural precision parameter (α)** minimizing uncertainty in cultural identity and modulating lower-level precisions.

These goals fall into three categories:

---

1. New Human Narrative: Egalitarian → Archetypes

> MOH as gateway to cultural evolution, extending Boehm's egalitarian origins toward Jungian archetypal differentiation.

- **Egalitarian "everyman" to Archetypes**: Adoption equates to binary in-group aligned vs. deviant cultural actor that represents egalitarian norms. Over time, agents differentiate into roles (e.g., aggressive → Warrior; knowledge-seeking → Sage).
- **Morality as scaffold of cultural evolution**: Codified norms stabilize anonymous groups, allowing expansion beyond intimates and creating conditions for cultural learning (Boyd & Henrich 2016).
- **Archetype emergence as equilibrium**: Roles consolidate through sanction/approval dynamics, reflecting the transition from flat egalitarian norms to symbolic archetypal structures.

---

2. Morals → Symbols

> Moral modeling (TMS) as template for symbolic modeling; shared norms become scaffolds for symbolic semiotics.

- **Moral Agency as model selection (structured learning)**: Reflectively driven capacity to inhabit hierarchies of arbitrarily high depth. Observing α (cultural precision) indicates allegiance to broader Self.
- **Model Selection and Symbols**: Model selection as moral agency can be framed as shallow vs. deep models. Detecting a "cultural" level signals a deeper model, which is then reflectively applied to one's self and intimates.
- **Gestalt vs. Analytic cognition**: Model selection between sufficient depth vs. narrow payoff models captures the classic gestalt/analytic dichotomy in symbolic cognition.
- **Morality → Cultural Learning → Language**: Codified norms not only stabilize groups but also equate to symbolic semiotics, grounding conventional model construction and linguistic scaffolding (cf. Friston on generative linguistics).

---

3. Morality as Road to and Regulation of AGI

> TMS as a governance framework: α operationalized for machine alignment.

- **Moral Agency as guiding principle for AGI**: Precision parameter α can serve as an explicit model of governance for AGI, modulating lower-level rule adherence without requiring exhaustive specification of norms.
- **Dynamic governance**: Just as moral rules shift over time while governing specific rules that change faster, α provides stability while allowing adaptive flexibility.
- **Anti-reductionist imperative**: Simulation demonstrates that morality operates not as a single payoff function but as layered symbolic scaffolding — a principle vital for resisting reductive AI approaches

Agent Variables

Vars	Definition	Intimates	Shibboleth	Moral/Culture
S	States (hidden)	**emo**:{anger, happy, fearful}; **partner**:{coop, defect}; **trust**:{high, low}	**signal_status**:{pass, fail}; **ingroup_belief**:{in, out}; **type**:{coop, defect}	**norm**:{aligned, deviant}; **role**:{egalitarian, warrior, sage}; **context**:{ingroup-unknown, anonymous};

Vars	Definition	Intimates	Shibboleth	Moral/Culture
				culture_precision (α, continuous meta-state)
O	Observations	**acts**:{hit, Waa-Bark, help}; **affect**:{smile, frown, neutral}; **payoff**:{gain, neutral, loss}	**token**:{pass, mispronounce}; **badge**:{present, absent}; **feedback**:{approval, sanction}	**broadcast**:{approval, neutral, sanction}; **reputation**:{up, flat, down}; **payoff**:{gain, neutral, loss}
U	Actions	{deter, cooperate, avoid}	{signal_ingroup, challenge, withhold}	{cooperate, defect, approve, sanction, forgive}
Π	Policies (action sequences)	e.g., {cooperate→cooperate}, {cooperate→deter}, {avoid}	e.g., {signal_ingroup→cooperate}, {withhold→challenge}	e.g., {cooperate→approve}, {defect→sanction}, {cooperate→forgive}

> Notes
> • **α (cultural precision)** is continuous (hyper-precision) that modulates arbitration across models and tightens social-feedback mappings.
> • **Context layering:** Intimates (dyadic), Shibboleth (ingroup gate), Moral/Culture (anonymous/cultural scale).
> • **PD embedding:** material **payoff** is an observation modality used in **C** (preferences); cooperation/defection are actions in **U**.

Transitions / Generative Model Matrices

Matrix	Definition	Intimates	Shibboleth	Moral/Culture
A	Likelihood mapping $P(O \mid S)P(O \mid S)$	Affect given emo; partner action observed given partner type; payoff given (partner, action).	Token/badge given signal_status; feedback (approval/sanction) given (ingroup_belief, action).	Broadcast & reputation given (norm, action, context); payoff given (action, partner type). **α** tightens approval/sanction likelihoods.
B	State transitions $P(St \mid St-1, Ut-1)P(S_t \mid S_{t-1}, U_{t-1})$	Trust drops after (defect or hit); emo drifts toward anger after sanction; partner type slowly inferred.	ingroup_belief moves toward **in** after consistent pass & cooperative acts; fails push toward **out**; type estimate updated via actions.	norm drifts to **aligned** with consistent approval of cooperation & sanction of defection; role shifts toward warrior (frequent sanction) / sage (frequent approve/forgive); **α** increases with reliable, low-entropy feedback.
C	Preferences over outcomes (observations)	Prefer {help, smile, gain}; penalize {hit, loss}. Cooperation favored if trust high.	Prefer {pass, approval}; penalize {fail, sanction}. Prefer cooperative acts by ingroup.	In anonymous/ingroup-unknown: strong preference that **cooperation→approval** & **defection→sanction** are observed; social terms weighted by **α**; material payoffs still count but can be outweighed by cultural approval.
D	Priors over initial states	trust:high; emo:happy; partner:coop (benign prior among intimates).	weak ingroup prior (uncertain); modest prior that signals pass.	role:egalitarian; norm:aligned 0.6; context mixes ingroup-unknown/anonymous; moderate α prior.
E	Priors over policies (habits)	Habit: cooperate in intimates; avoid if trust low.	Habit: light **signal_ingroup** before cooperation.	Habit: cooperate + approve; sanction when clear defection in anonymous contexts.