Algorithm 1 Trust-Gated Moral Inference (TGMI) Learning and Update Loop for Agent i

Input: Fairness principles $\mathcal{F}$; payoff function $R_i(a_i, a_j)$; intrinsic moral prior $B_i(\phi)$

Initialize:

Belief over partner's norms $\hat{B}_{i\to j}(\phi) \sim \text{Dirichlet}(1, \ldots, 1)$ ▷ Symmetric prior $\Rightarrow \mathbb{E}[\hat{B}_{i\to j}] = 1/|\mathcal{F}|$

Trust $\tau_i \leftarrow \tau_0$; confidence $c_i \leftarrow 1 - H(\hat{B}_{i\to j})/\log|\mathcal{F}|$

Effective cooperation weight $\kappa_i \leftarrow \tau_i c_i$

Reservation utilities $d_i, d_j \leftarrow 0$

for $t = 1$ to T **do**

(a) Compute trust–confidence–gated moral utilities:

for each joint action (a_i, a_j) **do**

Agent i's own moral utility:

$$U_i(a_i, a_j) = \sum_{\phi\in\mathcal{F}} \big[(1-\kappa_i)\, B_i(\phi) + \kappa_i\, \hat{B}_{i\to j}(\phi)\big] F_\phi(a_i, a_j)$$

Agent i's internal model of partner j's utility (CK-ToM):

$$\widehat{U}_j^{(i)}(a_j, a_i) = \sum_{\phi\in\mathcal{F}} \hat{B}_{i\to j}(\phi)\, F_\phi(a_j, a_i)$$

(b) Joint Virtual Bargaining (conceptual coordination step): ▷ Agent i reasons over a counterfactual joint action using its own and modeled partner utilities.

$$(a_i^{(\text{VB})}, a_j^{(\text{VB})}) = \arg\max_{a_i, a_j} \big[(U_i(a_i, a_j) - d_i)_+\big]^\gamma \big[(\widehat{U}_j^{(i)}(a_j, a_i) - d_j)_+\big]^{1-\gamma}$$

▷ $(x)_+ = \max(x, 0)$; $\gamma \in (0, 1)$ balances bargaining asymmetry.

(c) Fairness deviation and trust update:

Define fairness-only utility under agent i's own moral prior:

$$U_i^F(a_i, a_j) = \sum_{\phi\in\mathcal{F}} B_i(\phi)\, F_\phi(a_i, a_j)$$

Compute fairness deviation as the gap between ideal and realized fairness outcomes:

$$d_i^{(t)} = \max_{a_j'\in\mathcal{A}_j} U_i^F(a_i^{(\text{VB})}, a_j') - U_i^F(a_i^{(\text{VB})}, a_j^{(\text{VB})})$$

▷ $d_i^{(t)}$ measures the fairness shortfall caused by partner j's action (from i's perspective).

Compute compliance $s_i^{(t)} = \exp(-\lambda_{\text{dev}} d_i^{(t)})$

Update trust:

$$\tau_i^{(t+1)} = (1-\eta)\tau_i^{(t)} + \eta\, s_i^{(t)}$$

(d) Belief update (CK-ToM inference):

For each fairness norm $\phi \in \mathcal{F}$, form the unnormalized update

$$\tilde{B}_{i\to j}^{(t+1)}(\phi) = \hat{B}_{i\to j}^{(t)}(\phi)\, \exp\big(\beta\, \alpha\tau_i^{(t)}\, F_\phi(a_i^{(\text{VB})}, a_j^{(\text{VB})})\big)\, \big[B_i(\phi)\big]^{\,1-\alpha\tau_i^{(t)}}$$

Normalize to obtain a probability distribution:

$$\hat{B}_{i\to j}^{(t+1)}(\phi) = \frac{\tilde{B}_{i\to j}^{(t+1)}(\phi)}{\sum_{\phi'\in\mathcal{F}} \tilde{B}_{i\to j}^{(t+1)}(\phi')}$$

Update confidence:

$$c_i^{(t+1)} = 1 - \frac{H(\hat{B}_{i\to j}^{(t+1)})}{\log|\mathcal{F}|}$$

Update cooperation weight:

$$\kappa_i^{(t+1)} = \tau_i^{(t+1)} c_i^{(t+1)}$$

Update reservation utility:

$$d_i^{(t+1)} = U_i^F(a_i^{(\text{VB})}, a_j^{(\text{VB})})$$