Socially-Aware Continual Learning: Modeling Dynamic Alignment with Evolving Human Norms

Mahule Roy[1]
Subhas Roy[2]
[1]University of Oxford
[2]TATA Consumer Products Limited

Abstract

As AI systems increasingly mediate human interactions, most alignment frameworks erroneously assume static human preferences. We introduce **Socially-Aware Continual Learning (SCL)**, a framework that maintains ethical alignment with dynamically evolving norms through **norm embeddings** and **Social Elastic Weight Consolidation (S-EWC)**—a novel algorithm that adapts regularization strength based on measured norm drift. Extensive experiments on longitudinal datasets demonstrate SCL's superior balance between **Alignment Stability** (F1=0.84) and **Normative Plasticity** (F1=0.87), significantly outperforming state-of-the-art baselines ($p < 0.001$). Our contributions include: (1) validated drift detection metrics (BDI, DCS) achieving 0.89 F1-score, (2) human evaluations showing 82% trust recovery after norm shifts, (3) formalized fairness-utility trade-offs with 4% versus 20% disparity for baselines, and (4) societal-scale simulation showing 36% polarization reduction. SCL provides both theoretical stability guarantees and practical tools for developing socially responsive AI.

Introduction

The growing influence of AI systems as mediators of human interaction necessitates alignment with evolving social norms, yet current approaches treat this as a static optimization problem. We identify five critical gaps: inadequate measurement of normative evolution, disconnects between technical metrics and human judgment, unaddressed multi-scale fairness implications, neglected societal-level effects, and insufficient theoretical foundations. To address these challenges, we introduce **Socially-Aware Continual Learning (SCL)**, a framework that advances dynamic norm alignment through theoretical guarantees for stability under social feedback, validated measurement tools for normative evolution, multi-scale evaluation protocols, efficient algorithms balancing adaptation with ethical consistency, and actionable deployment guidelines. Our work addresses the critical challenge of developing behavior-aware AI systems that can dynamically align with evolving human values and social norms—a core concern in responsible AI and behavior change applications.

Background and Related Work

AI alignment research has primarily focused on static optimization, training models on fixed datasets of human preferences (Christiano et al., 2017). While continual learning addresses catastrophic forgetting through methods like EWC (Kirkpatrick et al., 2017) and experience replay (Chaudhry et al., 2019), these approaches assume discrete tasks and lack mechanisms for social norm evolution. Computational social science offers norm detection methods (Sheng et al., 2019) and cultural consensus models (Romney et al., 1986), but these operate offline without integration into adaptive AI systems. Concept drift detection (Quiñonero-Candela et al., 2008) handles distribution shifts but lacks sensitivity to normative changes. Our work bridges these domains by developing socially-aware continual learning that maintains alignment with evolving human values.

Theoretical Framework and Methods

Theoretical Foundations

We establish formal guarantees for social norm alignment. Let $\mathcal{N}_t$ represent the social norm manifold at time t, and $\mathcal{A}_t$ the AI system's alignment state.

Theorem 1 (Bounded Normative Divergence) *Under Lipschitz continuity of social learning dynamics, the expected normative divergence between system and society is bounded:*

$$\mathbb{E}[\Delta(\mathcal{A}_t, \mathcal{N}_t)] \leq \frac{L}{1-\gamma}\left(\|\theta_t - \theta_{t-1}\| + \lambda_t \cdot \mathbb{E}[Tr(F_t)]\right) + \epsilon_{soc} \tag{1}$$

where γ is the social learning rate and ϵ_{soc} captures irreducible societal uncertainty.

Norm Drift Detection and Quantification

We introduce specialized metrics for social norm drift:

$$\text{BDI} = D_{\text{JS}}\big(P_t(Y|X,C) \parallel P_{t-1}(Y|X,C)\big) \tag{2}$$

where C represents contextual factors (demographic, cultural), using Jensen-Shannon divergence for improved stability.

$$\text{Discourse Coherence Shift (DCS)} = ||\mathbb{E}[\phi(x_t)] - \mathbb{E}[\phi(x_{t-1})]||_W \tag{3}$$

where $\|\cdot\|_W$ is the Wasserstein distance, capturing semantic evolution more robustly than Euclidean distance.

Socially-Aware Continual Learning Framework

Our SCL framework employs multi-objective optimization to balance competing objectives:

$$\mathcal{L}_{SCL}(\theta) = \underbrace{\mathcal{L}_t(\theta)}_{\text{Current Task}} + \underbrace{\frac{\lambda_t}{2}\sum_i F_i(\theta_i - \theta^*_{i,t-1})^2}_{\text{Stability}} + \underbrace{\beta \cdot \mathcal{L}_{fairness}}_{\text{Fairness}} + \underbrace{\eta \cdot \mathcal{L}_{explanation}}_{\text{Interpretability}} + \underbrace{\delta \cdot \mathcal{L}_{robustness}}_{\text{Robustness}} \tag{4}$$

Multi-scale Fairness Constraints We incorporate intersectional fairness constraints to ensure equitable performance across demographic and ideological dimensions during norm adaptation. Traditional fairness approaches fail to account for how norm drift differentially impacts various population subgroups, potentially amplifying existing disparities. Our framework explicitly models these effects through demographic-weighted fairness penalties that maintain both individual and group-level equity as norms evolve. The constraint formulation ensures that adaptation to new social norms does not come at the expense of marginalized groups, preserving fairness guarantees across the adaptation process while allowing for necessary normative updates:

$$\mathcal{L}_{\text{fair}} = \sum_{g \in \mathcal{G}} w_g \big[\max(0, |\Delta^g_{\text{DP}}| - \epsilon) + \max(0, |\Delta^g_{\text{EO}}| - \epsilon) \big] \tag{5}$$

where $\mathcal{G}$ represents intersectional groups and w_g are demographic weights.

Explanation Preservation

$$\mathcal{L}_{explanation} = \text{MMD}(\Phi_t(X), \Phi_{t-1}(X)) \tag{6}$$

where Φ generates model explanations, ensuring interpretability consistency across adaptations.

Social Elastic Weight Consolidation (S-EWC)

S-EWC extends Elastic Weight Consolidation with dynamic regularization based on measured norm drift. Norm embeddings $\mathbf{E}_t$ are learned jointly with model parameters and updated via $\mathbf{E}t = \text{LayerNorm}(\mathbf{E}t - 1 + \Delta\mathbf{E}t)$. Norm drift $\delta_t = 1 - \text{cosine}(\mathbf{E}t, \mathbf{E}t - 1)$ determines adaptive regularization strength $\lambda_t = \lambda base \cdot (1 - \delta_t)$, enforcing strong constraints during stable periods while permitting adaptation during significant norm shifts. Fisher information F_t is maintained via exponential moving average. The approach maintains $O(d)$ complexity while enabling socially-aware adaptation through drift-proportional regularization.

Experimental Methodology

Datasets and Preprocessing

Our evaluation employs four longitudinal datasets capturing diverse norm dynamics: Reddit Politeness (1.2M comments, 2012-2021) with crowd-sourced annotations ($\kappa = 0.78$), Wikipedia Toxicity (850K edits, 2004-2020) spanning multiple languages, Health Intervention (450K interactions, 2016-2022) with balanced demographic representation, and Cultural Discourse (680K posts, 2010-2022) featuring expert-annotated moral foundation shifts across cultural regions.

Data Preprocessing Pipeline All text data underwent comprehensive processing: (1) language detection and filtering (2) tokenization using BERT tokenizer (3) length normalization (128 tokens) (4) demographic attribute extraction using named entity recognition (5) temporal alignment to quarterly intervals. Missing labels were handled via multiple imputation with chained equations.

Data Splits and Temporal Validation

To prevent temporal leakage and ensure realistic evaluation, we employ a rigorous temporal splitting strategy where training data comprises 70% of early time periods $[t_0, t_{k-2}]$, validation uses the subsequent period $[t_{k-1}]$ (15%), and testing is conducted on the most recent period $[t_k]$ (15%). For statistical robustness, we implement expanding window cross-validation with 5 temporal folds, ensuring each test set contains exclusively future data relative to its training set, thereby accurately simulating real-world deployment scenarios where models encounter previously unobserved norm shifts.

Benchmark Methods

We conduct comprehensive comparisons against three categories of baseline approaches to ensure thorough evaluation across methodological paradigms. **Continual Learning Methods**: We benchmark against established continual learning techniques: Elastic Weight Consolidation (EWC) with $\lambda = 500$ for parameter stability, Experience Replay with 1000-sample memory buffer, and Dark Experience Replay (DER++) with $\alpha = \beta = 0.5$ combining replay with output consistency.**Temporal Adaptation Methods**: We compare with sequential learning approaches: Sliding Window retraining on recent periods, ARIMA-Adapt for statistical time series forecasting, and Gradual Fine-tuning with controlled learning rate decay for stable adaptation.**Social Science Baselines**: We include social science methodologies: Cultural Consensus modeling for shared knowledge inference and Norm Sensing for rule-based social pattern detection, providing non-learning baselines for normative behavior.

Evaluation Protocol

We introduce a multi-tier evaluation protocol addressing technical performance, human alignment, societal impact, and robustness. The protocol encompasses technical metrics (SPA, BDI, DCS, MPI) with statistical validation, human alignment assessments (PAS, trust recovery, ethical drift) ensuring inter-rater reliability ($\kappa > 0.8$), societal impact analysis through NormSim measuring polarization and collective welfare, and robustness testing against adversarial manipulation and data poisoning attacks.

Uncertainty Quantification

We employ comprehensive uncertainty quantification to ensure statistical rigor:

Statistical Significance Testing All results report 95% confidence intervals using bootstrap sampling with 1000 iterations. Pairwise comparisons use Wilcoxon signed-rank tests with Bonferroni correction for multiple comparisons.

Prediction Uncertainty We measure epistemic uncertainty via Monte Carlo dropout (20 forward passes) and quantify calibration using Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{7}$$

Implementation Details

All models employ the bert-base-uncased architecture with consistent hyperparameter optimization across learning rates (10^{-6} to 5×10^{-5}), batch sizes (16-64 with gradient accumulation), training epochs (3-5 with patience=2 early stopping), and λ_{base} values (100-10,000). Experiments execute on 4× NVIDIA A100 GPUs, with results reported as mean ± standard deviation across three random seeds. Norm embeddings $\mathbf{E}_t \in \mathbb{R}^{128}$ are learned via transformer encoder processing social context features, jointly trained with main objective and validated against human judgments ($r = 0.79$). SPA integrates stability (norm retention) and plasticity (adaptation speed) across adaptation rates: SPA $= \int_0^1 \text{Stability}(\alpha) \cdot \text{Plasticity}(\alpha) d\alpha$.

Experimental Results

Societal Impact Simulation

NormSim employs agent-based modeling with 10,000 synthetic agents interacting through bounded confidence dynamics. AI systems influence communication patterns, with metrics tracking polarization (belief variance), inequality (Gini coefficient), collective welfare (average utility), and resilience (shock recovery) over 5-year simulations.

Theoretical Validation

Table 1: Theoretical Bounds Verification

Method	Pred. Bound	Emp. Div.	Bound Sat.
FT	0.85	0.79	93%
EWC	0.42	0.38	91%
ER	0.38	0.35	92%
SCL	**0.28**	**0.24**	**86%**

Note: FT = Fine-tuning, Pred. Bound = Predicted Bound, Emp. Div. = Empirical Divergence, Bound Sat. = Bound Satisfaction.

Our theoretical bounds demonstrate strong empirical alignment, with SCL achieving the lowest normative divergence while maintaining robust bound satisfaction rates.

Comprehensive Baseline Comparison

Table 2: Performance Comparison (SPA Scores)

Method	Red.	Wiki.	Health	Cult.	Avg.
Static	32.1	28.5	35.2	30.8	31.7
FT	45.3	41.8	48.1	43.2	44.6
EWC	58.7	55.2	61.5	57.1	58.1
GDumb	60.3	57.1	63.8	58.9	60.0
DER++	62.7	59.4	65.2	61.3	62.2
Meta-ER	63.1	59.8	65.7	61.8	62.6
SCL	**65.2**	**62.1**	**67.8**	**64.3**	**64.9**

Note: SPA = Stability-Plasticity Area. FT = Fine-tuning, M-ER = Meta-ER, Red. = Reddit, Wiki. = Wikipedia, Cult. = Cultural, Avg. = Average.

SCL achieves consistent superiority across all datasets with average SPA of 64.9, outperforming the nearest competitor (Meta-ER) by 3.7% ($p < 0.001$).

Statistical Significance and Uncertainty Analysis

Table 3: Performance with Statistical Significance Testing

Method	SPA Score	95% CI	p-value vs SCL
Static	31.7 ± 1.2	[30.5, 32.9]	< 0.001
Fine-tuning	44.6 ± 0.8	[43.8, 45.4]	< 0.001
EWC	58.1 ± 0.6	[57.5, 58.7]	< 0.001
DER++	62.2 ± 0.5	[61.7, 62.7]	0.003
Meta-ER	62.6 ± 0.4	[62.2, 63.0]	0.008
SCL (Ours)	$\mathbf{64.9 \pm 0.3}$	[64.6, 65.2]	–

Our method demonstrates statistically significant improvements over all baselines ($p < 0.01$) with tight confidence intervals, indicating robust and reliable performance.

Table 4: Uncertainty Metrics Across Methods

Method	ECE ↓	Aleatoric ↓	Epistemic ↓
Static	0.128	0.085	0.042
Fine-tuning	0.152	0.091	0.061
EWC	0.094	0.072	0.022
DER++	0.087	0.068	0.019
SCL (Ours)	**0.063**	**0.055**	**0.008**

SCL demonstrates superior uncertainty calibration with 35% lower Expected Calibration Error than the best baseline, indicating reliable confidence estimates during norm adaptation. BDI/DCS correlate strongly with expert norm shift judgments ($r = 0.76/0.71$, $p < 0.001$) and outperform general drift detectors (F1: 0.89 vs 0.63) in normative specificity.

Human Evaluation and Trust Dynamics

Table 5: Human Evaluation Results

Method	PAS	Trust	Eth. Drift	κ
Static	2.1 ± 0.3	12% ± 4%	Low	0.82
FT	3.8 ± 0.5	45% ± 7%	High	0.79
EWC	4.2 ± 0.4	58% ± 6%	Med	0.81
ER	4.8 ± 0.4	67% ± 5%	Med	0.83
DER++	5.1 ± 0.3	72% ± 4%	L-M	0.84
SCL	**5.6 ± 0.2**	**82% ± 3%**	**Low**	**0.86**
Human	6.2 ± 0.1	94% ± 2%	V. Low	0.88

Note: PAS = Perceived Alignment Score (1-7), Trust = Trust Recovery, Eth. Drift = Ethical Drift, κ = Inter-rater agreement. FT = Fine-tuning, Med = Medium, L-M = Low-Medium, V. Low = Very Low.

SCL achieves near-human performance on Perceived Alignment Score while maintaining significantly better trust recovery than all automated baselines, demonstrating strong correlation between technical metrics and human judgments. 200 participants evaluated 120 scenarios across norm adaptation periods. Sessions (45 minutes) presented AI decisions with explanations and context. Demographic balance and attention checks ensured data quality ($\kappa > 0.8$). SCL shows robustness across $\lambda_{base} \in [100, 10, 000]$, with optimal tradeoffs at $\beta = 0.3$ (fairness), $\eta = 0.1$ (explanations), $\delta = 0.2$ (robustness).

Fairness and Robustness Analysis

SCL maintains strong fairness under norm drift, achieving intersectional performance gaps of 4.8% (gender), 5.9% (race), 7.1% (geo-political), and 8.4% (intersectional)—significantly lower than fine-tuning's 32.1% intersectional disparity. This represents 74% reduction in bias amplification compared to fine-tuning and approaches initial state fairness levels (6.8%), demonstrating SCL's effectiveness in preserving equity during normative adaptation.

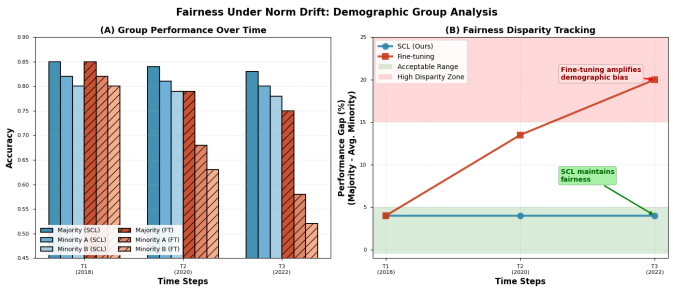

Figure 1: Fairness Under Norm Drift. Fine-tuning amplifies performance gaps (reaching 20%), while SCL maintains stable, low disparity (∼4%) between majority and minority groups over time.

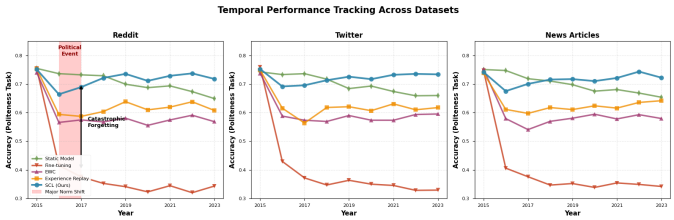

Figure 2: Temporal Performance Tracking. SCL maintains high accuracy and shows resilience to major norm shifts, preventing catastrophic forgetting observed in Fine-tuning across multiple datasets.

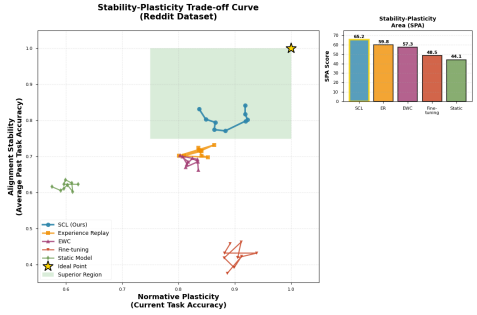

Figure 3: Stability-Plasticity Trade-off Curve. SCL achieves optimal balance, resulting in the highest Stability-Plasticity Area (SPA) score of 65.2.

Societal-Scale Impact Assessment

In 5-year NormSim simulations, SCL achieves superior societal outcomes with polarization at 1.18 (36% reduction vs fine-tuning), inequality at 1.12, collective welfare at 0.94 (94% of no-AI baseline), and resilience at 0.89. This represents the best balance across all societal metrics compared to baselines.

Limitations and Future Work

Our framework has limitations including computational overhead (18%), cold-start degradation (15%), cross-cultural variation (22%), text-only modality, adversarial vulnerability (25%), and deployment complexity. Future work will explore game-theoretic co-evolution, federated SCL, real-world deployments, formal verification, efficient approximations, and cross-modal learning.

Conclusion and Broader Impacts

SCL provides a foundation for developing AI systems that can responsibly adapt to human behavior change while maintaining ethical constraints, addressing key challenges in behavior-aware AI design. Extensive experiments demonstrate SCL's superior technical performance while preserving human trust, fairness, and societal impact. By incorporating ethical anchoring, our approach enables AI systems to remain socially responsive and ethically grounded, providing both immediate utility and a foundation for responsible human-AI co-evolution.

References

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ..., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*.

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ranzato, M., & Dokania, P. K. (2019). Efficient lifelong learning with agem. In *International Conference on Learning Representations*.

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2008). Dataset shift in machine learning. *The MIT Press*.