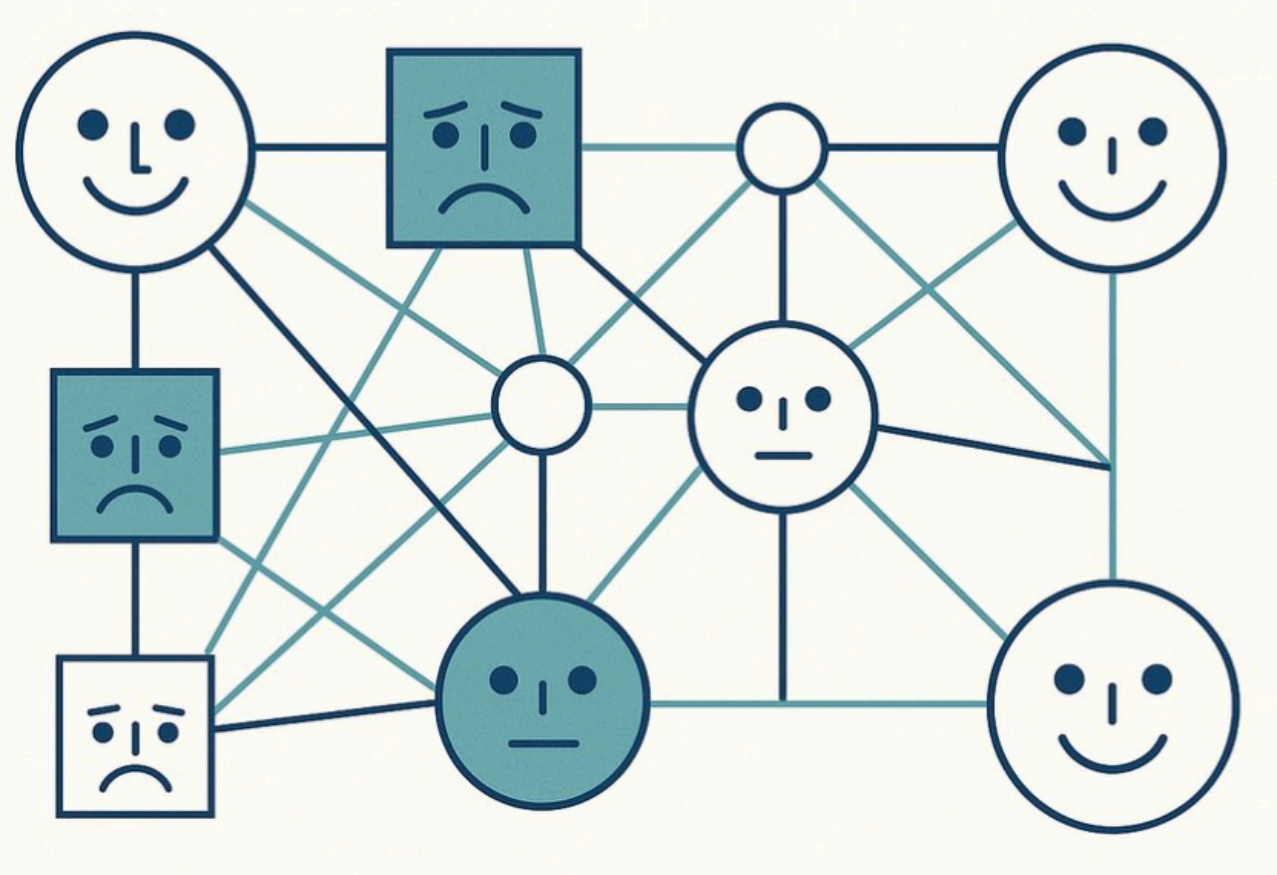
ABMHuB 2025

Wrong Face,
WRONG MOVE

The Social Dynamics of Emotion
Misperception in Agent-Based Models

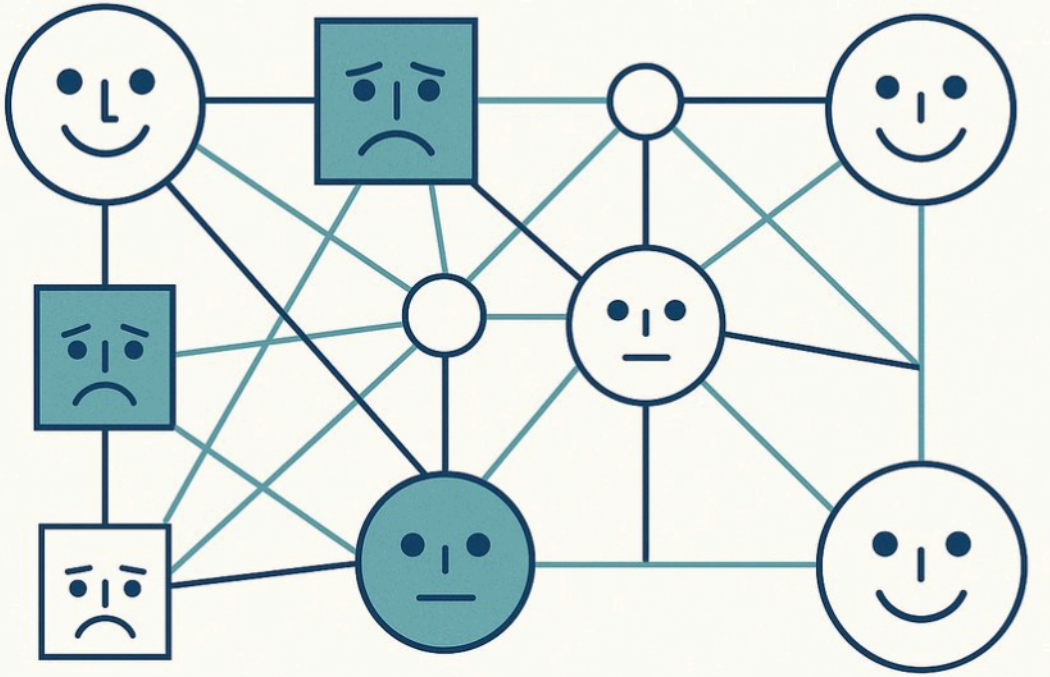


David Freire-Obregón

Universidad de Las Palmas de Gran Canaria, Spain

david.freire@ulpgc.es

Motivation and Background



Social Intelligence Foundation

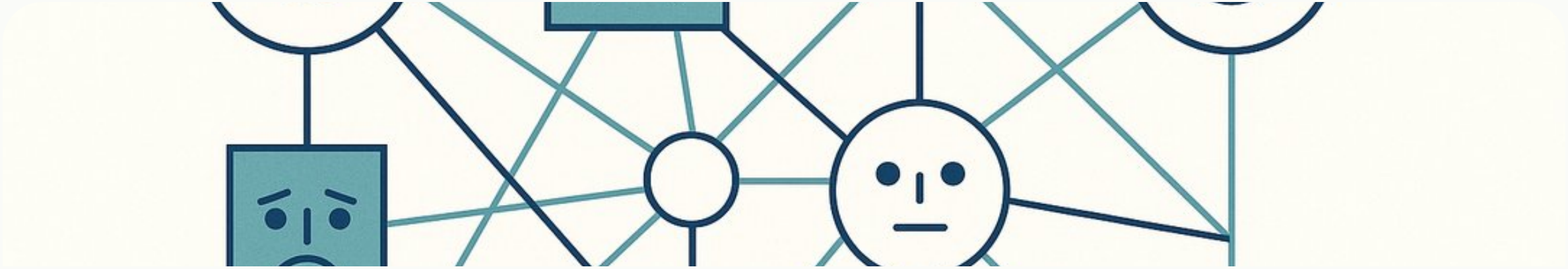
The ability to detect and respond to others' emotional states is fundamental to human social behavior and underpins cooperative societies.

Emergent Social Patterns

Accurate emotion perception shapes trust development, spatial organization, and collective emotional stability in social groups.

AI System Implications

Emotion AI deployed in social technologies may carry hidden risks if recognition accuracy is compromised by bias or poor generalization.



Research Objectives and Hypothesis

Impact of Classifier Accuracy

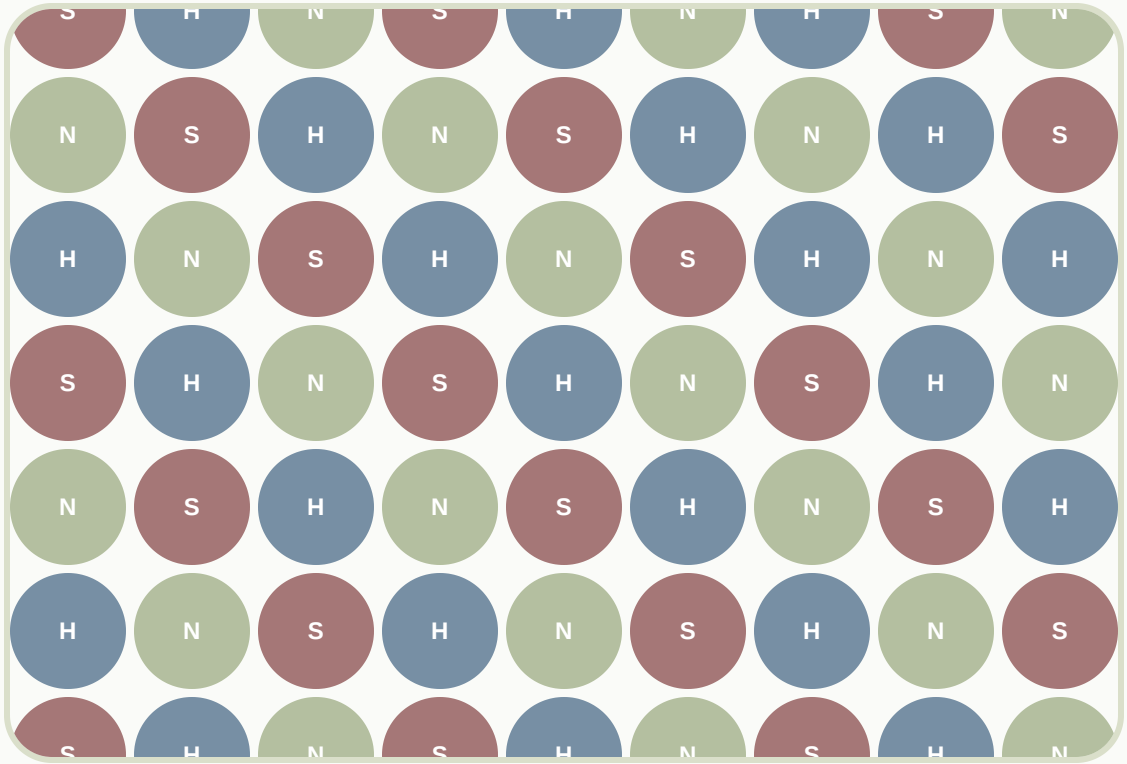
Investigate how the accuracy of emotion perception classifiers affects emergent social dynamics in agent-based models, including trust formation and emotional contagion patterns.

Misperception Consequences

Test the hypothesis that systematic emotion misperception disrupts social stability and cohesion even in emotionally neutral environments, leading to trust erosion and fragmentation.

AI System Implications

Explore how these findings translate to real-world emotion AI systems, highlighting risks of biased or inaccurate recognition for social technologies and human-machine interaction.



Model Overview

Agent Population

40 distinct agents placed on a 9×9 toroidal grid ($G \subset Z^2$), creating a bounded world where agents interact locally with their neighbors.

Emotion Classifiers

Agents use one of three CNN classifiers with varying accuracy: KDEF (96% accuracy), CK+ (37% accuracy), or JAFFE (19% accuracy).

Interaction Dynamics

Agents respond to perceived emotions, approach positive emotions, avoid negative ones, and update trust based on perception accuracy.

Experimental Design

Homogeneous Populations

All agents equipped with the same classifier (KDEF, CK+, or JAFFE) to evaluate baseline group dynamics under uniform perception conditions.

Mixed Populations

Various combinations of agents with different classifiers to investigate how perception asymmetries and classifier diversity affect social cohesion.

Resilience Testing

Repeated negative emotional shocks introduced to test group resilience and recovery patterns under different perception accuracy conditions.

Statistical Methodology

Each parameter configuration run 10 times for 100 time-steps, with 40 agents (matching KDEF dataset identities) on a 9×9 toroidal grid for statistical robustness.

The Emotion Classifiers

Performance comparison across different cultural datasets

KDEF

96%

Recognition Accuracy

High-performance classifier trained and tested on the Karolinska Directed Emotional Faces dataset.

- Precision: 97%
- Recall: 96%
- F1-Score: 96%

CK+

37%

Recognition Accuracy

Medium-performance classifier trained on Extended Cohn-Kanade dataset but tested on KDEF.

- Precision: 36%
- Recall: 37%
- F1-Score: 30%

JAFFE

19%

Recognition Accuracy

Low-performance classifier trained on Japanese Female Facial Expression dataset but tested on KDEF.

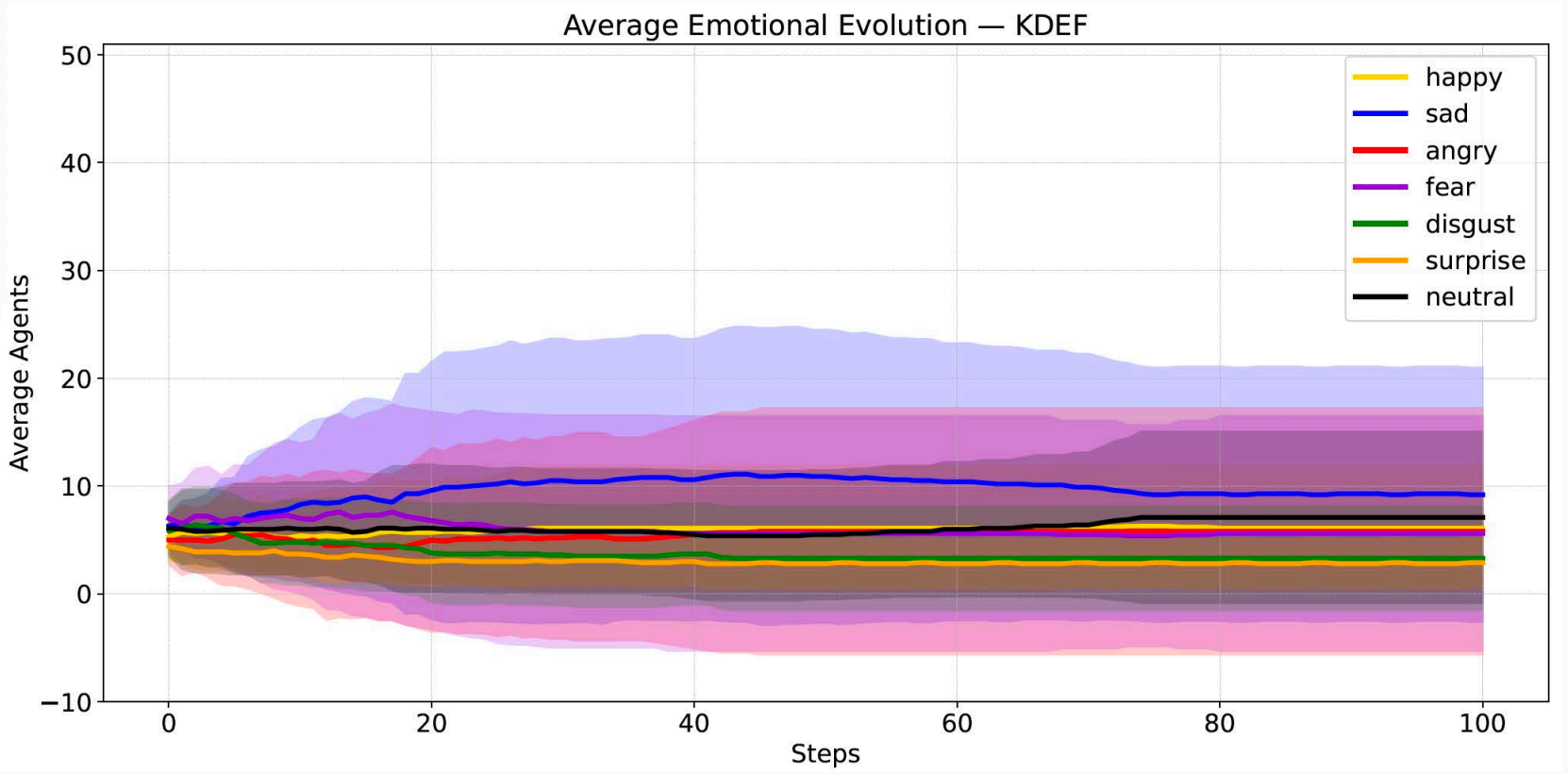
- Precision: 15%
- Recall: 19%
- F1-Score: 10%

Results: Homogeneous Populations

High Accuracy (KDEF)

Trust: 0.96

Balanced emotional landscape with diverse expressions maintained throughout the simulation. Stable moderate-sized emotional clusters form naturally.

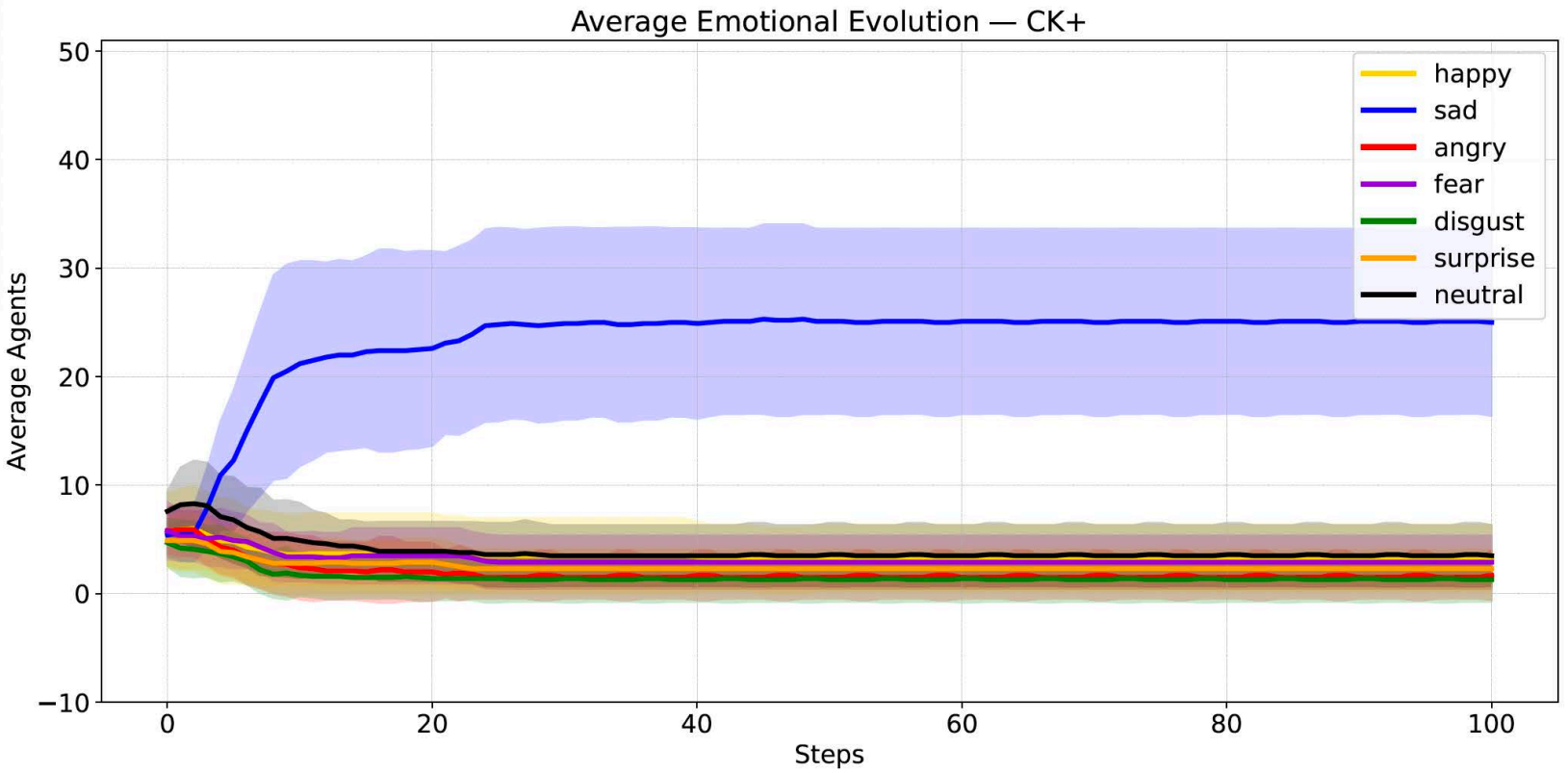


Results: Homogeneous Populations

Medium Accuracy (CK+)

Trust: 0.25

Significant polarization occurs with "sad" emotion becoming increasingly dominant. Trust drops considerably and emotional clusters become more extreme.

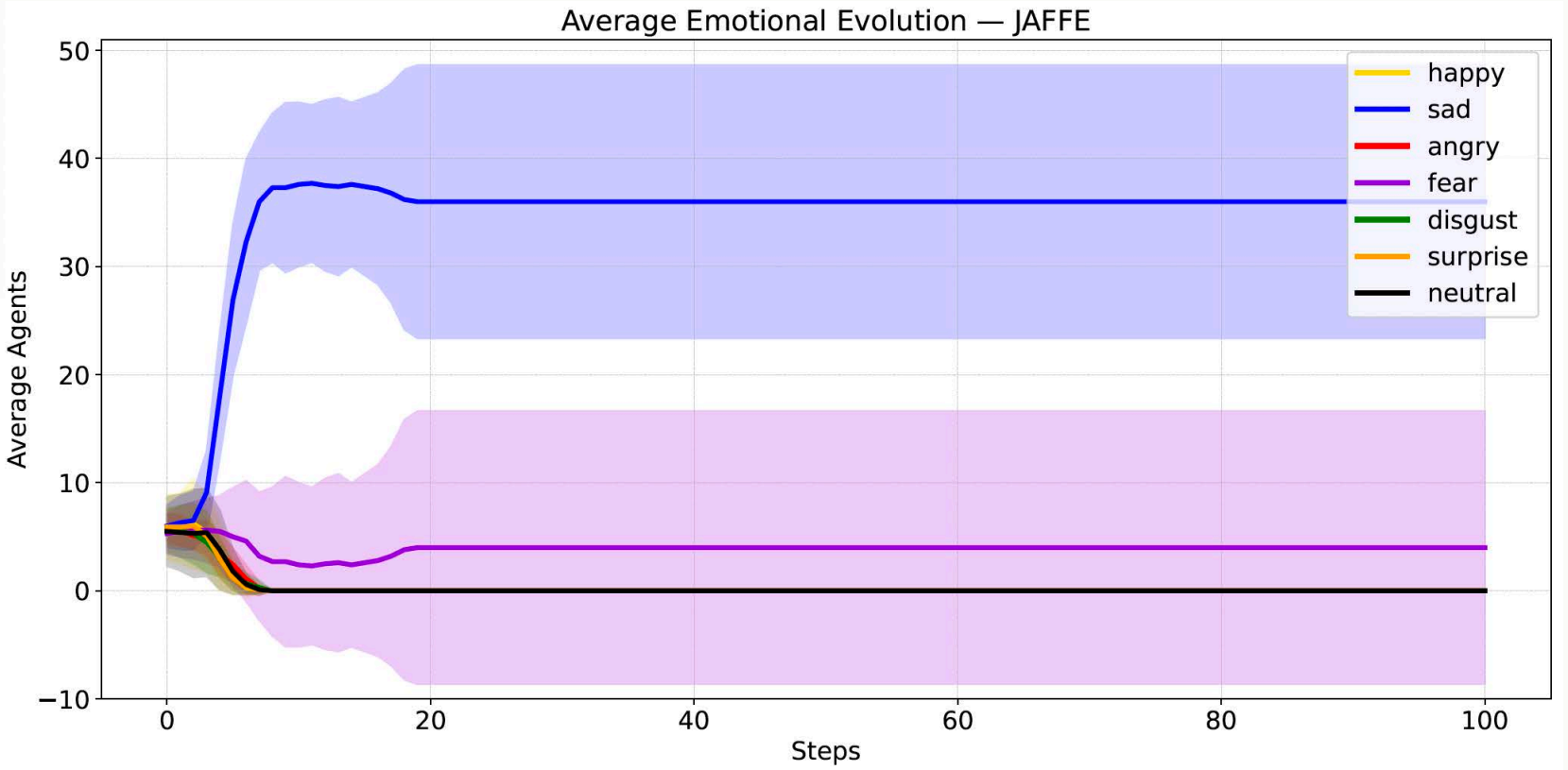


Results: Homogeneous Populations

Low Accuracy (JAFFE)

Trust: 0.038

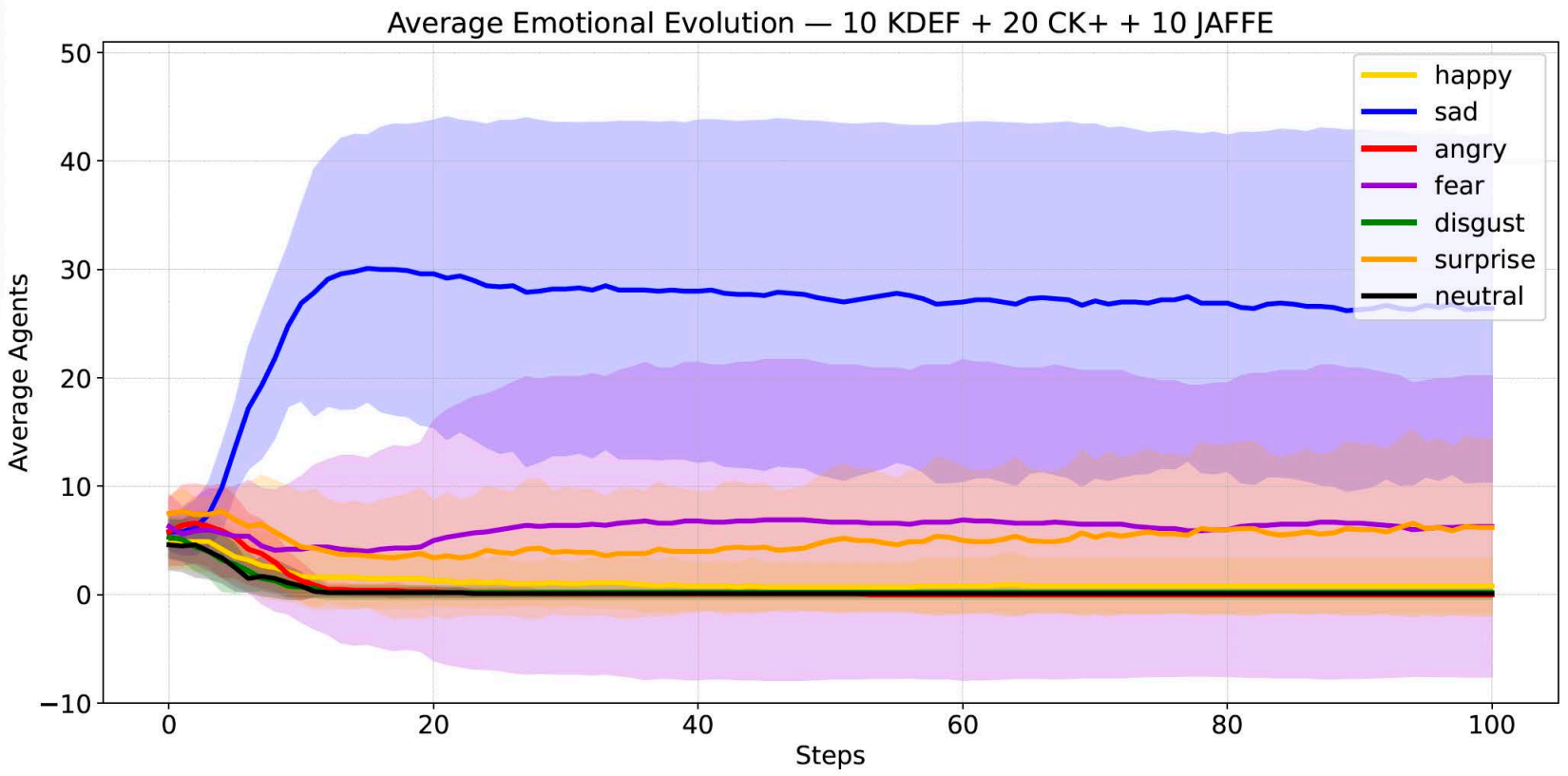
Nearly all agents converge to "sad" emotional state. Trust collapses to near zero and large, homogeneous clusters of negative emotions form.



Results: Mixed Populations

Mixed Distribution

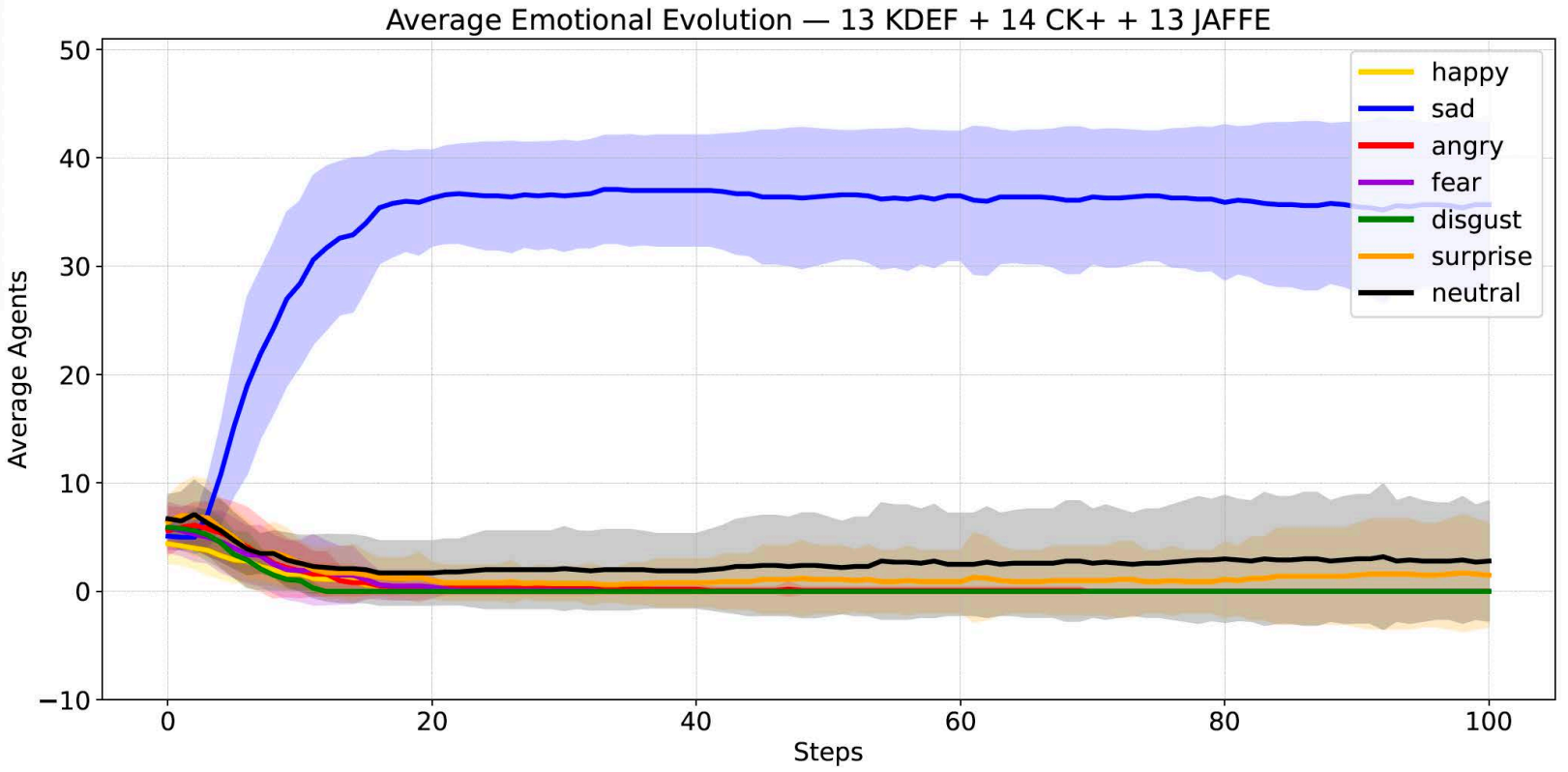
Balanced populations show segregation patterns, with high-accuracy agents forming stable emotional clusters while low-accuracy agents drift toward sadness.



Results: Mixed Populations

Low-Accuracy Influence

Even a small minority of poor classifiers can destabilize entire populations, causing widespread negative emotion contagion and trust collapse.



Trust, Clusters, and Contagion

Trust Metrics Evolution

0.96 → 0.25 → 0.04

Trust levels show dramatic decline from KDEF (high accuracy) to CK+ (medium) to JAFFE (low), revealing direct correlation between perception accuracy and social trust development.

Emotional Cluster Size

- KDEF: Small balanced clusters (max 3.6)
- CK+: Medium sad clusters (max 7.3)
- JAFFE: Large sad clusters (max 16.0)

Lower perception accuracy generates larger, denser clusters of negative emotions, particularly sadness.

Emotional Contagion

Misperception triggers negative emotional cascades: false negatives lead agents to adopt sad states, which then spread to neighbors in a self-reinforcing cycle, particularly pronounced in JAFFE populations.

Social Fragmentation

Similar to Schelling's classic segregation model, systematic misperception drives avoidance and trust decay, accumulating into emotional clustering and social fragmentation even in initially neutral environments.

Perturbation & Resilience Experiments

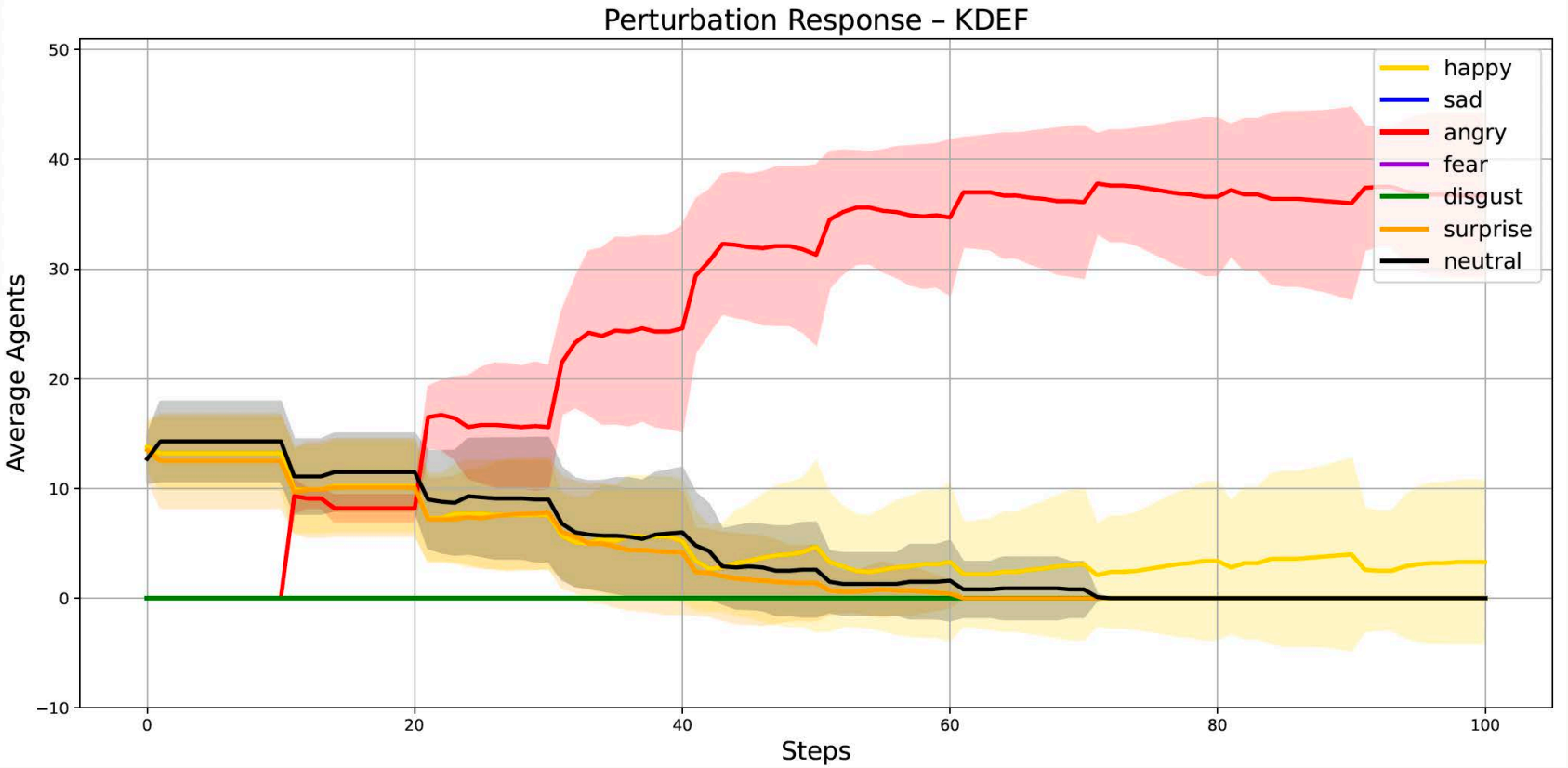
How different populations respond to repeated negative emotional shocks

KDEF Agents

Partial Resilience

A small core of positivity persists even after multiple negative shocks. Trust remains relatively stable at higher levels.

- Positive emotion dominant
- Mixed emotions present



Perturbation & Resilience Experiments

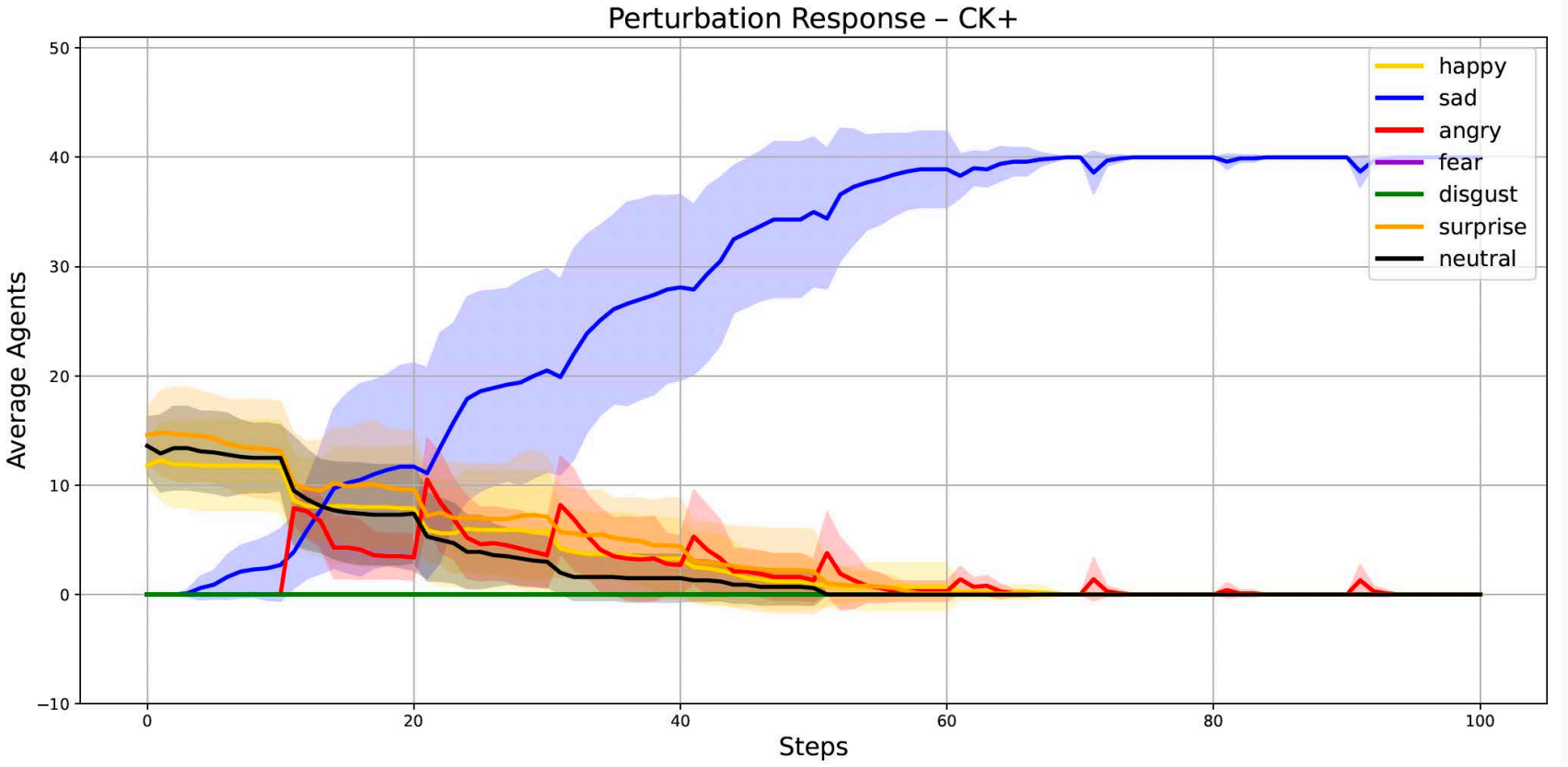
How different populations respond to repeated negative emotional shocks

CK+ Agents

Rapid Degradation

Initial stability quickly deteriorates after just two negative shocks. Trust drops significantly and sadness spreads throughout the population.

- Negative emotion dominant
- Trust approaching zero



Perturbation & Resilience Experiments

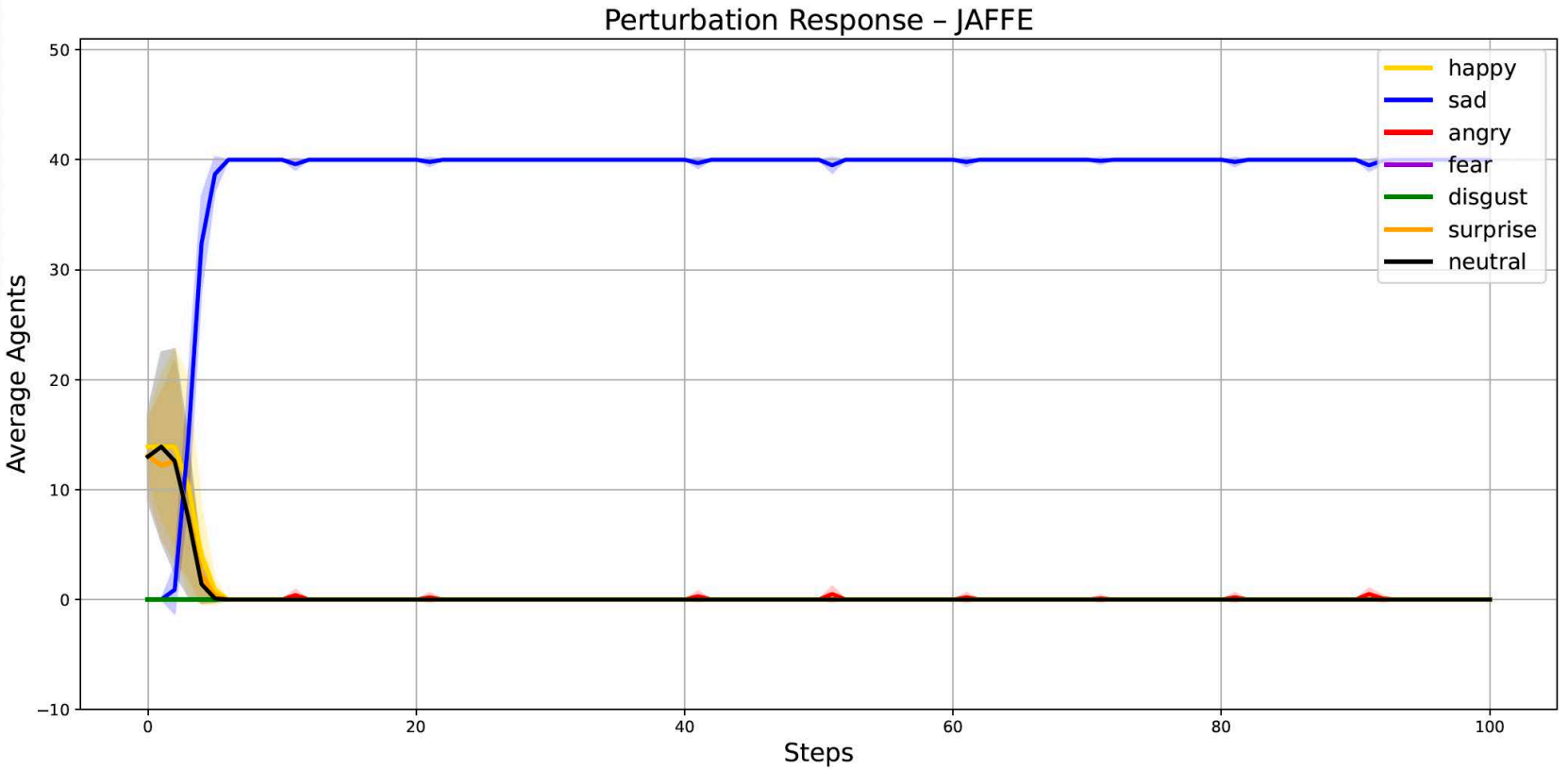
How different populations respond to repeated negative emotional shocks

JAFFE Agents

Immediate Collapse

System collapses after a single emotional shock. Almost all agents converge to sadness state with virtually no trust remaining in the population.

- Extreme negative dominance
- Complete trust collapse



Implications for AI and Social Systems

Critical Trust Factor

Reliable emotion perception is essential for building user trust in AI systems. Even moderately inaccurate emotion recognition can lead to rapid trust erosion in human-AI interaction.

Bias Amplification Risk

Cultural and demographic mismatches in training data can create systematic misperception that amplifies social divides and leads to emotional fragmentation in diverse populations.

Resilient AI Design

Social robots, virtual agents, and emotion-aware systems must prioritize robust cross-cultural emotion recognition to avoid inadvertently damaging the social fabric they aim to enhance.

Concusions: Perception Shapes Social Reality

Perceptual Accuracy

High-accuracy emotion classifiers maintain trust, emotional diversity, and stable social organization in agent populations.

Systematic Misperception

Even in neutral environments, misperception leads to loss of trust, spreading negative affect, and social fragmentation.

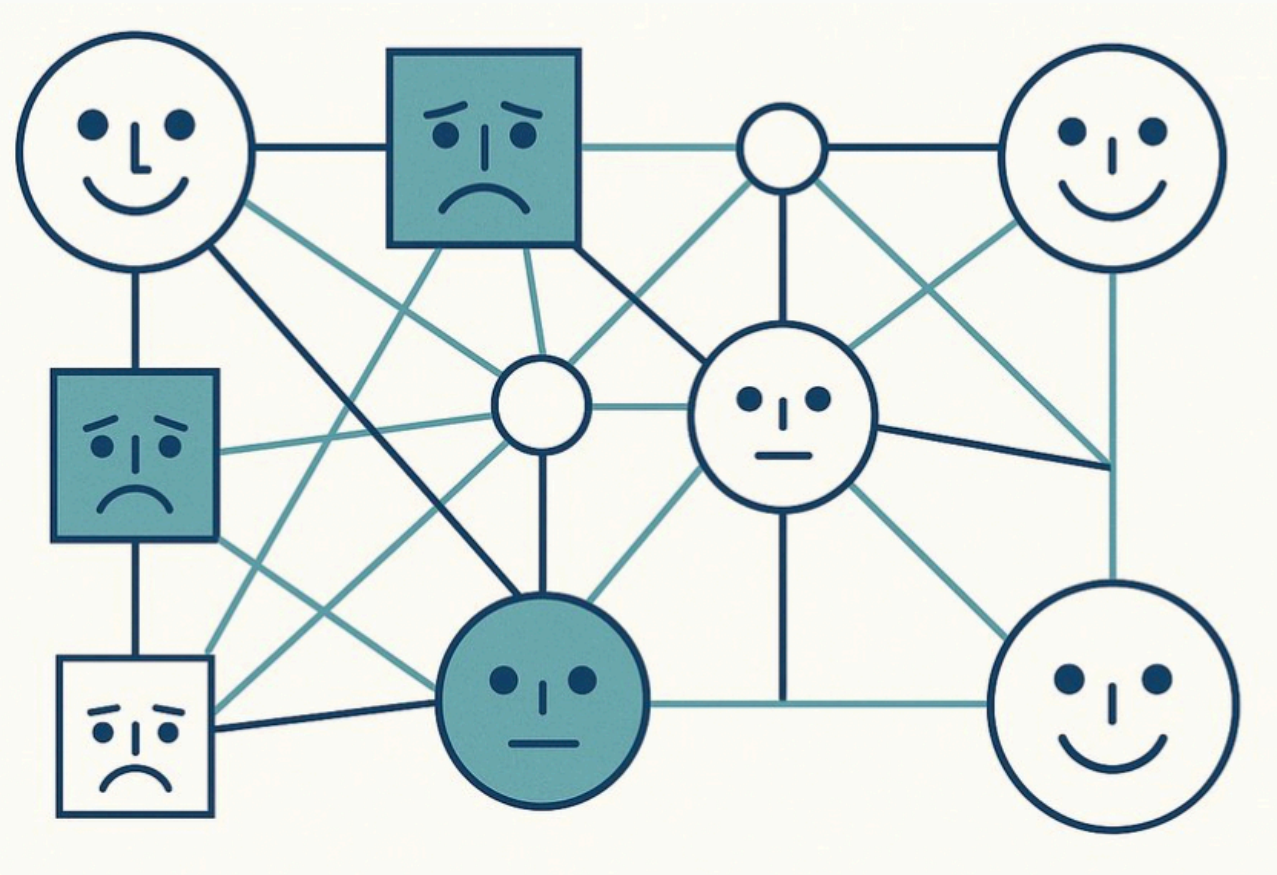
AI Applications

Findings underscore the critical need for culturally robust emotion recognition systems in social AI and robotic technologies.

Future Directions

Extending our model to explore more complex social scenarios, diverse cultural contexts, and potential mitigation strategies for misperception.

Thank You!



David Freire-Obregón

Universidad de Las Palmas de Gran Canaria, Spain

david.freire@ulpgc.es