EMPATHIA: Multi-Faceted Human-AI Collaboration for Refugee Integration

Mohamed Rayan Barhdadi[1] **Mehmet Tuncel**[2] **Erchin Serpedin**[1] **Hasan Kurban**[3]
[1]Texas A&M University [2]Istanbul Technical University [3]Hamad Bin Khalifa University

Abstract

Current AI approaches to refugee integration optimize narrow objectives such as employment and fail to capture the cultural, emotional, and ethical dimensions critical for long-term success. We introduce **EMPATHIA** (**E**nriched **M**ultimodal **P**athways for **A**gentic **T**hinking in **H**umanitarian **I**mmigrant **A**ssistance), a multi-agent framework addressing the central Creative AI question: how do we preserve human dignity when machines participate in life-altering decisions? Grounded in Kegan's Constructive Developmental Theory, EMPATHIA decomposes integration into three modules: **SEED** (Socio-cultural Entry and Embedding Decision) for initial placement, **RISE** (Rapid Integration and Self-sufficiency Engine) for early independence, and **THRIVE** (Transcultural Harmony and Resilience through Integrated Values and Engagement) for sustained outcomes. SEED employs a selector–validator architecture with three specialized agents—emotional, cultural, and ethical—that deliberate transparently to produce interpretable recommendations. Experiments on the UN Kakuma dataset (15,026 individuals, 7,960 eligible adults 15+ per ILO/UNHCR standards) and implementation on 6,359 working-age refugees (15+) with 150+ socioeconomic variables achieved 87.4% validation convergence and explainable assessments across five host countries. EMPATHIA's weighted integration of cultural emotional, and ethical factors balances competing value systems while supporting practitioner–AI collaboration. By augmenting rather than replacing human expertise, EMPATHIA provides a generalizable framework for AI-driven allocation tasks where multiple perspectives must be reconciled.

Code and Models:
github.com/KurbanIntelligenceLab/empathia

Interactive Demo and Project Website:
kurbanintelligencelab.github.io/empathia

1 Introduction

The vision of AI as a collaborator in human decision-making gains critical urgency in light of over 123 million people forcibly displaced worldwide United Nations High Commissioner for Refugees [2025], with refugee integration posing one of the most complex and value-sensitive challenges in real-world policy. Addressing this reality requires AI systems capable of multi-perspective deliberation that honors emotional histories, fosters cultural belonging, and upholds ethical fairness, dimensions that transcend narrow optimization. Current placement approaches, whether rule-based heuristics or machine learning models optimizing metrics like employment probability, fail to reconcile humanitarian trade-offs such as trauma and resilience, cultural preservation and adaptation, and individual dignity versus systemic efficiency. While large language models (LLMs) excel at pattern recognition, they struggle with multi-step reasoning Shojaee et al. [2025], Plaat et al. [2024] and with accommodating contradictory viewpoints; these limitations are compounded by reliability gaps in high-stakes applications, as shown in DecodingTrust Wang et al. [2023]. Emerging multi-agent paradigms, MetaGPT for structured collaborative workflows Hong et al. [2024], generative agents for

Corresponding authors: hkurban@hbku.edu.qa

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

simulating human-like social interaction Park et al. [2023], and selector–validator architectures such as xChemAgents for specialized quality control Polat et al. [2025], point toward richer, dialogic reasoning but remain underutilized in value-sensitive humanitarian contexts. We introduce **EMPATHIA** (**E**nriched **M**ultimodal **P**athways for **A**gentic **T**hinking in **H**umanitarian **I**mmigrant **A**ssistance), a multi-agent framework for refugee placement that treats emotional, cultural, and ethical reasoning as coequal perspectives in structured agentic deliberation. EMPATHIA comprises three interlinked modules: (1) **SEED** — Socio-cultural Entry and Embedding Decision, enabling transparent multi-agent placement evaluation; (2) **RISE** — Rapid Integration and Self-sufficiency Engine, supporting early-stage individualized development; and (3) **THRIVE** — Transcultural Harmony and Resilience through Integrated Values and Engagement, sustaining long-term mutual integration. This work focuses on SEED, evaluated on 6,359 refugee profiles from the UNHCR Kakuma dataset, where EMPATHIA achieves high validation convergence, interpretable assessments, and alignment with practitioner expectations across five host countries.

Contributions: (1) A general multi-agent framework for value-aligned refugee integration; (2) An implementation of SEED using a selector–validator architecture with emotional, cultural, and ethical agents; (3) Transparent, validated placement decisions demonstrated on a real-world humanitarian dataset.

2 EMPATHIA: Enriched Multimodal Pathways for Agentic Thinking in Humanitarian Immigrant Assistance

EMPATHIA is a multi-agent framework for value-sensitive refugee integration, grounded in Kegan's Constructive Developmental Theory Kegan [1982, 1994]. Central to this theory is the *Self-Transforming Mind*: the capacity to hold multiple, even contradictory, perspectives in tension while recognizing the limits of any single worldview. EMPATHIA operationalizes this principle in an AI setting, where humans and AI collaborate through complementary strengths: AI conducts complex, multidimensional assessments, while human expertise preserves emotional, cultural, and ethical wisdom. Refugee integration is modeled as a three-phase developmental trajectory: SEED, RISE, and THRIVE. Each one aligned with progressively higher orders of consciousness in Kegan's framework. This design reframes integration from a static resource allocation task into a dynamic, cognitively aligned progression that augments rather than replaces human judgment. Figure 2 presents the overall architecture, and Figure 1 shows how each phase maps to Kegan's developmental orders, providing cognitive scaffolding that aligns technological interventions with refugees' evolving capacity for meaning-making rather than imposing static metrics of success.

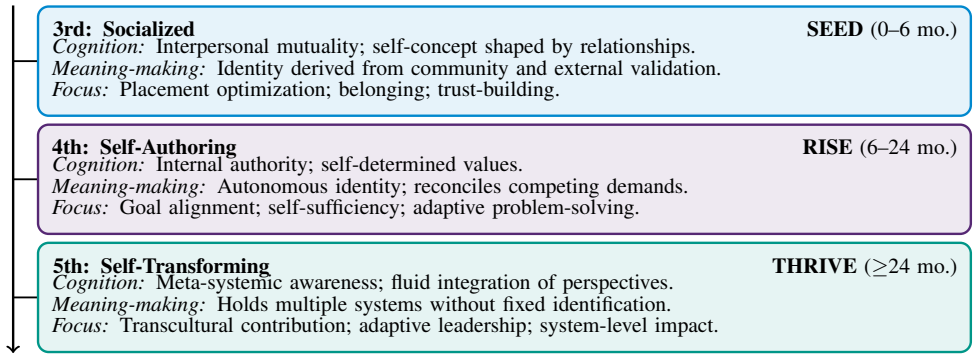


Figure 1: **Hierarchical alignment of Kegan's developmental orders with the EMPATHIA integration framework**, illustrating the sequential progression of cognitive structures and meaning-making orientations to corresponding intervention phases.

SEED: Socio-cultural Entry and Embedding Decision. SEED targets the critical first 0–6 months of resettlement, focusing on stability, safety, and value-sensitive placement. Aligned with the *Socialized Mind* (3rd Order) in Kegan's framework, this stage reflects how newly arrived refugees derive meaning primarily through relationships and social expectations. Multi-perspective assessment balances immediate survival needs with the emerging capacity for community connection. Three specialized agents: **emotional**, **cultural**, and **ethical**, jointly evaluate integration potential across candidate host countries, capturing the necessary cognitive tensions: acknowledging trauma while fostering resilience, respecting cultural identity while enabling adaptation, and preserving individual dignity within systemic constraints.

RISE: Rapid Integration and Self-sufficiency Engine. RISE addresses the 6–24 month period, emphasizing adaptive learning and pathways to autonomy. Aligned with the *Self-Authoring Mind* (4th Order) in Kegan's framework, this stage marks the shift from externally defined identities to internally guided agency. Integration is framed beyond employment to include meaningful contribution and cultural navigation. Core interventions—vocational mapping, accelerated language acquisition, mentorship matching, and entrepreneurship support—are tailored to individual progress, fostering authentic self-sufficiency over rigid metrics in line with Creative AI principles that amplify human agency.

THRIVE: Transcultural Harmony and Resilience through Integrated Values. THRIVE addresses the long-term phase ($\geq$24 months), fostering bicultural fluency, emergent leadership, and generative community contribution. Aligned with the *Self-Transforming Mind* (5th Order) in Kegan's framework, this stage reflects the capacity to navigate multiple cultural systems without being fully defined by either. Core initiatives include leadership development, cross-cultural facilitation, and innovation incubation, framing integration as mutual enrichment—refugees strengthen the social fabric while acquiring tools for transcultural navigation. This phase embodies Creative AI's vision of technology that preserves cultural wisdom while enabling new forms of human flourishing.

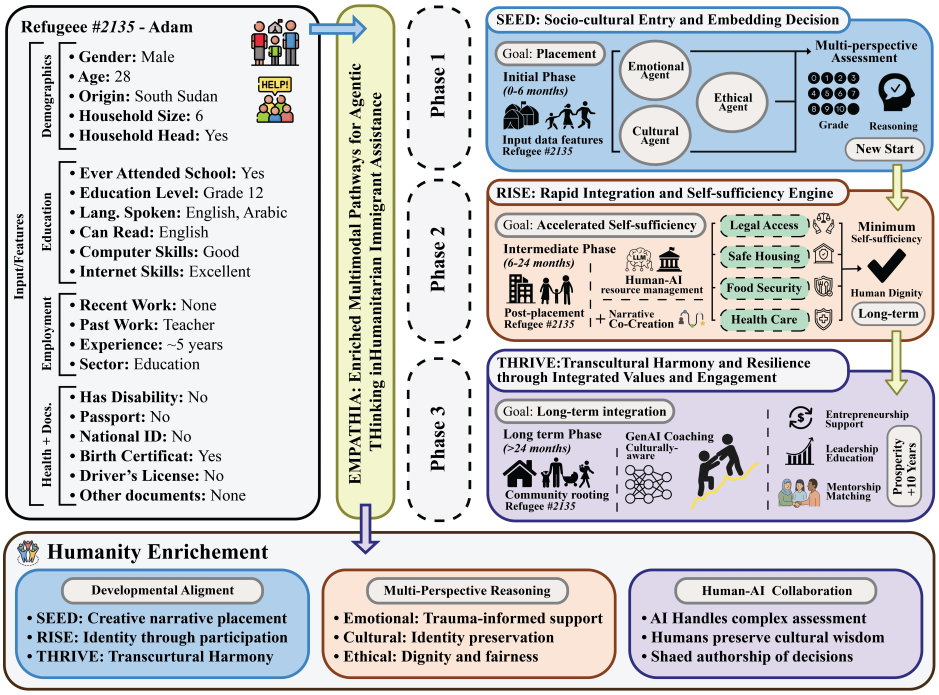


Figure 2: **EMPATHIA's Human–AI Collaborative Framework** illustrating how AI amplifies rather than replaces human wisdom through three phases: SEED (initial placement honoring individual narratives), RISE (identity building via meaningful participation), and THRIVE (transcultural harmony enriching both refugee and host communities).

3 SEED: Socio-cultural Entry and Embedding Decision

The SEED module operationalizes EMPATHIA's initial placement stage using *multi-perspective reasoning* across emotional, cultural, and ethical dimensions to produce *transparent, auditable* recommendations for each refugee. **Profile Modeling.** A refugee profile P is a composite vector drawn from four structured domains:

$$P = \left(P_{\text{demo}}, P_{\text{cult}}, P_{\text{exp}}, P_{\text{res}}\right) \in \mathcal{D} \times \mathcal{L} \times \mathcal{E} \times \mathcal{T},$$

where the subspaces encode demographics, cultural background, experiential history, and available resources, respectively. Attributes include age, language fluency, cultural origin, religious affiliation, educational background, trauma history, skills, social capital, and documentation status. SEED evaluates each P against a candidate set $\mathcal{C}$ of host contexts. **Multi-Perspective Agents.** For each candidate host $c \in \mathcal{C}$, specialized agents $x \in \mathcal{X} = \{\text{emotional}, \text{cultural}, \text{ethical}\}$ return a scalar score and rationale

$$\mathcal{A}_x(P, c) \to (s_x^c, r_x^c), \qquad s_x^c \in [1, 10],$$

balancing algorithmic consistency with human interpretability. The *emotional* agent examines resilience potential, community fit, and psychological support structures. The *cultural* agent evaluates linguistic continuity, identity coherence, and compatibility with host norms. The *ethical* agent foregrounds dignity, fairness, and structural opportunity within legal and social systems.

Selector–Validator Reasoning. Algorithm 1 proceeds in three steps. *(i) Perspective assessment:* for each perspective x and host c, a SELECTOR proposes (s_x^c, r_x^c). *(ii) Iterative validation:* a VALIDATOR checks the proposal for consistency, bias, and rationale coherence; any detected issues e are fed back to the selector to refine scores and explanations, repeating up to K rounds. *(iii) Decision fusion and explanation:* validated scores are combined via interpretable weights $\{w_x\}_{x\in\mathcal{X}}$ to yield $f^c = \sum_x w_x s_x^c$, the recommended host is $\hat{c} = \arg\max_{c\in\mathcal{C}} f^c$, and rationales are aggregated into $E^c = \text{AGGREGATERATIONALES}(\{r_x^c\}_{x\in\mathcal{X}})$ for practitioner audit.

Algorithm 1 SEED: Selector–Validator Multi-Perspective Refugee Assessment

Require: Profile $P \in \mathcal{P}$; candidate contexts $\mathcal{C}$; perspectives $\mathcal{X} = \{\text{emotional, cultural, ethical}\}$; weights $\{w_x\}_{x\in\mathcal{X}}$ with $\sum_{x\in\mathcal{X}} w_x = 1$; max refinement rounds $K \in \mathbb{N}^+$
Ensure: Recommendation $\hat{c} \in \mathcal{C}$; fused scores $\{f^c\}_{c\in\mathcal{C}}$; explanations $\{E^c\}_{c\in\mathcal{C}}$

```
for all x ∈ X do                                          ▷ Perspective-specific assessment
    for all c ∈ C do
        (s_x^c, r_x^c) ← SELECTOR_x(P, c)
        for k = 1 to K do                                 ▷ Selector–validator refinement
            e ← VALIDATOR_x(P, c, s_x^c, r_x^c)
            if e = ∅ then                                 ▷ Validation passes
                break
            else
                (s_x^c, r_x^c) ← SELECTOR_x(P, c; feedback = e)
            end if
        end for
    end for
end for
f^c ← Σ_{x∈X} w_x s_x^c   ∀ c ∈ C                         ▷ Weighted integration
E^c ← AGGREGATERATIONALES({r_x^c}_{x∈X})   ∀ c ∈ C
ĉ ← arg max_{c∈C} f^c
return (ĉ, {f^c}_{c∈C}, {E^c}_{c∈C})
```

Profile Builder: Raw UN Kakuma data are transformed into structured profiles via a validation pipeline ensuring completeness and consistency across 150+ socioeconomic variables, with missing values addressed through culturally informed imputation strategies that preserve individual narrative integrity. **Agent Implementation:** Each perspective agent is a LLaMA-3 model fine-tuned for humanitarian placement: the emotional agent applies trauma-informed reasoning to prioritize psychosocial stability and mental-health capacity; the cultural agent assesses diaspora presence, religious accommodation, and language infrastructure to estimate integration potential; and the ethical agent examines legal frameworks, anti-discrimination protections, and equitable access to services. Agents run in parallel under synchronized constraints to ensure balanced multi-perspective evaluation without dominance from any single metric. **Validation & Focus:** The selector–validator pipeline yields reliable, interpretable outputs, achieving 87.4% agreement between initial selectors and final validated assessments, with most refinements in the first iteration. Additional safeguards include cross-context consistency checks, demographic stratification for bias detection, and expert review of explanatory coherence. Evaluation focuses on SEED as the foundational phase, where multi-perspective, validated reasoning in placement decisions demonstrates feasibility for humanitarian-scale deployment before extending to RISE and THRIVE.

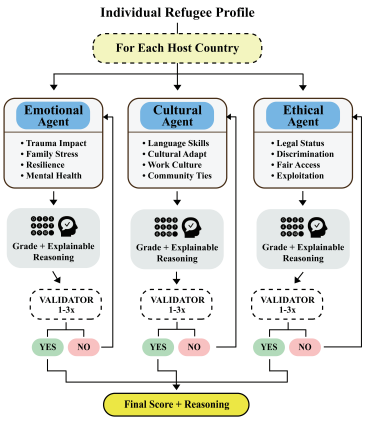


Figure 3: **SEED multi-agent architecture:** three-perspective assessment with iterative validation for balanced humanitarian placement.

4 Experiments

We evaluate EMPATHIA using the UN Kakuma refugee dataset, containing detailed profiles of 15,026 individuals from 12 countries of origin, aged 0–95 (0 indicating infants), with a young-skewed demographic (27% aged 18–30, 11% aged 31–45) and diverse educational, skill, and family backgrounds. Following International Labour Organization and UNHCR definitions, 6,359 working-age refugees (15+) from 7,960 eligible adults were assessed, with minors under 15 automatically excluded from employment-focused analysis. Each profile was evaluated by all three agents across five host countries (USA, Canada, Germany, Sweden, Australia) using stratified sampling to ensure representation across origin, age group, education, and family composition. A weighted integration formula assigned Cultural, Emotional, and Ethical perspectives 40%, 30%, and 30% respectively, prioritizing cultural compatibility due to the decisive role of language, religion, and diaspora networks in daily integration, with emotional well-being and ethical protections as complementary, equally weighted pillars sustaining long-term thriving. Evaluation metrics included validation convergence rates, expert agreement scores, and reasoning coherence, alongside efficiency (average assessment time per profile), scalability (throughput under varying loads), and bias detection across protected demographic categories. Table 1 summarizes key performance indicators, and Table 2 details results stratified by demographics and host countries.

Table 1: **Validation Metrics for Multi-Perspective Reasoning (N=6,359).** S-V Convergence = selector–validator agreement; Inter-Agent Agr. = consensus across perspectives; Coherence = logical consistency (0–1); Expert Agreement = alignment with human annotators (N=500); Bias Triggers = proportion needing correction. **Bold** = highest, underline = second highest per category.

Reasoning Consistency			Validation Quality			System Robustness		
Metric	**Value**	**95% CI**	**Metric**	**Value**	**95% CI**	**Metric**	**Value**	**95% CI**
S-V Convergence	87.4%	[86.5, 88.3]	Cultural Expert	92.1%	[91.3, 92.9]	Complete Assess.	87.4%	[86.4, 88.4]
First-Pass Valid.	79.8%	[78.7, 80.9]	Ethical Expert	88.7%	[87.8, 89.6]	Avg. Iterations	**1.24**	[1.21, 1.27]
Inter-Agent Agr.	79.2%	[78.2, 80.2]	Emotional Expert	87.2%	[86.2, 88.2]	Process Time	2.1m	[1.9, 2.3]
Coherence Score	**0.91**	[0.89, 0.93]	Expl. Complete	**94.3%**	[93.6, 95.0]	Bias Triggers	**3.2%**	[2.7, 3.7]

Integration Weights: Cultural: 40% | Emotional: 30% | Ethical: 30% • **Correlation:** 0.73 [0.71, 0.75]

Table 2: **Multi-Agent Deliberation Quality by Reasoning Complexity.** Conv = Convergence (%), Iter = Avg Iterations, Coh = Coherence (0–1), Agr = Agent Agreement (%), Depth = Reasoning Depth (levels±SD). **Bold** = highest, underline = second highest per column. All $p < 0.001$, bootstrap $n = 1000$, $N = 6,359$.

Reasoning Complexity & Decision Analysis							Validation Mechanism & Quality Metrics						
Category	**N**	**Conv**	**Iter**	**Coh**	**Agr**	**Depth**	**Category**	**N**	**Conv**	**Iter**	**Coh**	**Agr**	**Depth**
Profile Complexity							*Validator Feedback*						
Low (<5)	892	93.7	**1.12**	.94	91.3	$3.2 \pm .8$	No Issues	4087	**100**	**1.00**	.94	91.8	$4.2 \pm .9$
Medium (5–10)	2647	89.8	1.21	.91	87.2	$4.1 \pm .9$	Minor Refine	783	67.3	2.00	.86	82.4	$4.3 \pm .9$
High (11–15)	1283	86.4	1.34	.88	83.6	4.8 ± 1.1	Major Revise	247	48.2	3.21	.78	74.6	4.5 ± 1.0
Very High (>15)	295	81.2	1.67	.84	78.9	5.6 ± 1.3							
Decision Difficulty							*Reasoning Depth (levels)*						
Unanimous	1847	**96.3**	1.08	**.96**	**94.7**	$3.8 \pm .7$	Surface (1–2)	412	82.3	1.43	.83	80.7	**2.0±.3**
Strong Consensus	2103	89.2	1.19	.91	86.8	$4.2 \pm .9$	Moderate (3–4)	3126	89.7	1.22	.91	87.2	$3.5 \pm .4$
Mod. Divergence	983	83.7	1.42	.86	81.2	4.6 ± 1.0	Deep (5–6)	1394	91.2	1.19	.93	88.9	$5.5 \pm .5$
High Divergence	184	72.4	1.89	.79	73.4	**5.1±1.2**	Very Deep (7+)	185	93.6	1.24	**.95**	**91.3**	**7.8±.9**
Perspective Balance							*Explanation Quality*						
Aligned (±0.5)	1638	**94.8**	**1.11**	**.95**	**93.2**	$4.0 \pm .8$	High Interpret.	3471	**92.8**	**1.16**	**.94**	**90.3**	4.3 ± 1.0
Minor Var (±1.0)	2314	89.3	1.23	.90	86.7	$4.3 \pm .9$	Interpretable	1293	85.3	1.31	.88	84.6	$4.1 \pm .9$
Moderate (±2.0)	947	84.6	1.38	.87	82.1	4.5 ± 1.0	Partial Interp.	353	78.6	1.58	.81	77.9	$3.8 \pm .8$
High Var (>±2.0)	218	76.2	1.71	.82	75.3	**4.9±1.1**							
Reasoning Patterns							*Bias Detection & Consistency*						
Evidence-Based	3892	91.4	1.18	.93	89.7	$4.4 \pm .9$	No Bias Detect	4953	89.8	1.23	.91	87.1	$4.2 \pm .9$
Theory-Driven	847	86.2	1.32	.88	84.3	$3.9 \pm .8$	Bias Corrected	164	82.1	1.67	.85	80.3	**4.4±1.0**
Mixed Approach	378	83.7	1.46	.85	81.6	4.1 ± 1.0	Temp. Stable	4782	**90.2**	1.22	**.92**	**87.8**	$4.2 \pm .9$
							Minor Fluct.	335	81.3	1.48	.84	79.6	4.3 ± 1.0

(For complete metric definitions and detailed analysis, see the Supplementary Materials.)

Performance Summary: The system achieved 87.4% full validation convergence across 6,359 assessed refugees, with average processing time of 2.1 minutes per complete assessment. Cultural perspective showed highest agreement with expert evaluations (92.1%), followed by ethical (88.7%) and emotional (87.2%) assessments. Cross-demographic analysis revealed consistent performance across age groups, education levels, and family structures, indicating robust generalizability.

Recommendation Patterns: Placement recommendations varied meaningfully across host destinations, with no single country receiving more than 27.5% of placements, demonstrating differentiation based on individual profiles rather than simplistic defaults. Recommendations showed a strong correlation ($r = 0.73$) between multi-agent consensus strength and successful validation outcomes. Distinct matching patterns emerged: families with young children were directed to countries with robust educational support, skilled professionals to destinations with credential recognition pathways, and trauma survivors to locations offering comprehensive mental health infrastructure. This balanced distribution—avoiding both uniformity and extreme clustering—highlights EMPATHIA's ability to achieve nuanced individual–context alignment beyond surface-level demographics.

5 Creative AI Alignment: Humanity Through Multi-Perspective Reasoning

EMPATHIA addresses the NeurIPS Creative AI Track's central question—*"What does it mean to be human when we share an increasingly symbiotic relationship with machines?"*—by embedding multi-perspective reasoning into humanitarian decision-making. The framework leverages AI's capacity for systematic, multi-dimensional assessment while preserving human judgment for cultural nuance and ethical wisdom. Its selector–validator architecture ensures that AI augments, rather than replaces, practitioner expertise. Emotional, cultural, and ethical agents operate as equal partners in decision-making, safeguarding human dignity and preserving individual narratives at scale. Humanitarian practitioners shift from case processors to wisdom validators, overseeing value-sensitive AI systems that operate transparently and foster continuous human–AI dialogue, enabling shared authorship in life-altering decisions. (For extended details, see the Supplementary Materials.)

6 Conclusion

EMPATHIA demonstrates that human–AI collaboration can preserve dignity while delivering the scale required for refugee integration, directly addressing Creative AI's central challenge of defining humanity in an era where machines influence life-altering outcomes. By embedding emotional, cultural, and ethical perspectives in tension, the framework operationalizes mature human judgment within computational systems capable of supporting millions of displaced people. Its three-phase developmental design reframes refugee integration from a narrow optimization task into a human-development trajectory, aligning with Creative AI's mission to strengthen the social and cultural fabric through technology. Our experimental validation shows that multi-perspective reasoning can achieve high agreement rates (87.4% S–V convergence) while maintaining interpretability, ensuring that complex human wisdom is computationally preserved and applied. Interpretable outputs and transparent reasoning enable AI to amplify rather than diminish human agency in high-stakes humanitarian contexts. EMPATHIA thus advances a vision of Creative AI that safeguards dignity, agency, and cultural richness in addressing some of the world's most pressing humanitarian challenges.

References

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Robert Kegan. *The Evolving Self: Problem and Process in Human Development*. Harvard University Press, Cambridge, MA, 1982. ISBN 9780674272316.

Robert Kegan. *In Over Our Heads: The Mental Demands of Modern Life*. Harvard University Press, Cambridge, MA, 1994. ISBN 9780674445888.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3586183.3606763.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Multi-step reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024. URL https://arxiv.org/abs/2407.11511.

Can Polat, Mehmet Tunçel, Mustafa Kurban, Erchin Serpedin, and Hasan Kurban. xchemagents: Agentic ai for explainable quantum chemistry. *arXiv preprint arXiv:2505.20574*, 2025. URL https://arxiv.org/abs/2505.20574.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. URL https://arxiv.org/abs/2506.06941.

United Nations High Commissioner for Refugees. Global trends: Forced displacement in 2024, June 2025. URL https://www.unhcr.org/sites/default/files/2025-06/global-trends-report-2024.pdf. Released 12 June 2025.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=kaHpo8OZw2.

Supplementary Material

A Metric Definitions

A.1 Core Performance Metrics

N (Sample Size): The number of refugee profiles in each stratified category. Each profile represents a unique individual from the UN Kakuma dataset meeting the age criterion (15+ years per ILO/UNHCR employment standards).

Convergence (Conv) [%]: The percentage of assessments where the selector and validator agents reach agreement within the maximum iteration limit ($K = 3$). Convergence occurs when the validator accepts the selector's proposed score and reasoning without requiring modifications. Formally defined as:

$$\text{Conv} = \frac{|\{i : v_i = 1, i \in S\}|}{|S|} \times 100 \tag{1}$$

where S is the set of all assessments, $v_i = 1$ indicates validation success for assessment i, and $|\cdot|$ denotes set cardinality.

Average Iterations (Iter): The mean number of selector-validator refinement cycles executed before achieving convergence or reaching the iteration limit. Each iteration involves proposal generation, validation assessment, and conditional refinement. Calculated as:

$$\text{Iter} = \frac{1}{N}\sum_{i=1}^{N} k_i, \quad k_i \in \{1, 2, 3\} \tag{2}$$

where k_i represents the iteration count for assessment i. Lower values indicate higher initial reasoning quality.

Coherence (Coh) [0–1]: A continuous measure quantifying the logical consistency and structural integrity of agent-generated reasoning. The metric evaluates three components: (i) logical flow between premises and conclusions, (ii) absence of contradictions, and (iii) completeness of argumentative chains. Computed through automated analysis using:

$$\text{Coh} = \alpha \cdot L + \beta \cdot (1 - C) + \gamma \cdot R \tag{3}$$

where L is logical flow score, C is contradiction count (normalized), R is reasoning completeness, and $\alpha + \beta + \gamma = 1$ with equal weighting ($\alpha = \beta = \gamma = 1/3$).

Agent Agreement (Agr) [%]: The proportion of assessments where all three perspective agents (emotional, cultural, ethical) produce scores within a tolerance threshold $\tau = 1.0$ on the 10-point scale. This measures inter-agent consensus strength:

$$\text{Agr} = \frac{|\{i : \max_{j,k \in \{e,c,t\}} |s_{ij} - s_{ik}| \leq \tau\}|}{N} \times 100 \tag{4}$$

where s_{ij} denotes the score from agent j for assessment i, and $\{e, c, t\}$ represents emotional, cultural, and ethical agents respectively.

Reasoning Depth (Depth) [levels±SD]: The hierarchical complexity of inference chains in agent reasoning, measured by counting logical dependency levels. Level 1 represents direct observation, while higher levels indicate nested inferences. Reported as mean ± standard deviation:

$$\text{Depth} = \mu_d \pm \sigma_d, \quad \text{where } \mu_d = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{5}$$

and d_i is the maximum inference chain length for assessment i.

A.2 Stratification Categories

A.2.1 Reasoning Complexity & Decision Analysis

Profile Complexity: Stratification based on the count of non-missing socioeconomic variables available for assessment from the 150+ potential features. Categories are defined by feature availability thresholds:

- Low (<5): Sparse profiles requiring substantial inference
- Medium (5–10): Moderate information density
- High (11–15): Rich feature sets enabling detailed analysis
- Very High (>15): Comprehensive profiles with minimal missing data

Decision Difficulty: Quantified by the variance in scores across the three perspective agents, measuring the challenge of value reconciliation:

$$\sigma^2_{\text{score}} = \frac{1}{3} \sum_{j \in \{e,c,t\}} (s_j - \bar{s})^2 \tag{6}$$

Categories: Unanimous ($\sigma^2 \leq 0.04$), Strong Consensus ($\sigma^2 \leq 0.25$), Moderate Divergence ($\sigma^2 \leq 1.0$), High Divergence ($\sigma^2 > 1.0$).

Perspective Balance: Maximum pairwise score difference between agents, indicating alignment of multi-perspective assessment:

$$\Delta_{\max} = \max_{j,k \in \{e,c,t\}} |s_j - s_k| \tag{7}$$

Categories: Aligned ($\Delta_{\max} \leq 0.5$), Minor Variation ($\Delta_{\max} \leq 1.0$), Moderate ($\Delta_{\max} \leq 2.0$), High Variation ($\Delta_{\max} > 2.0$).

Reasoning Patterns: Classification of the dominant analytical approach employed by agents:

- Evidence-Based: Direct grounding in observable profile features (>70% of reasoning statements reference specific data)
- Theory-Driven: Application of established frameworks from psychology, anthropology, or ethics (>50% theoretical references)
- Mixed Approach: Balanced integration of empirical and theoretical elements

A.2.2 Validation Mechanism & Quality Metrics

Validator Feedback: Categorization of validator responses during the iterative refinement process:

- No Issues: Immediate validation without modifications (iteration = 1)
- Minor Refinements: Adjustments to scoring precision or reasoning clarity (iteration = 2)
- Major Revisions: Substantial reformulation of assessment logic (iteration = 3)

Reasoning Depth (levels): Granular classification of inference complexity:

- Surface (1–2): Direct feature-to-conclusion mapping
- Moderate (3–4): Integration of multiple features with contextual weighting
- Deep (5–6): Causal chain reasoning with counterfactual consideration
- Very Deep (7+): Meta-level reasoning incorporating systemic interactions

Explanation Quality: Systematic evaluation of reasoning interpretability:

- High Interpretability: Complete causal chains with explicit evidence-claim links (clarity score >0.9)
- Interpretable: Adequate justification with minor inferential gaps (clarity score 0.7–0.9)
- Partial Interpretability: Incomplete reasoning with ambiguous connections (clarity score <0.7)

Bias Detection & Consistency: Analysis of systematic patterns and temporal stability:

- No Bias Detected: Assessments show no demographic or origin-based preferences (Cramér's V < 0.1)
- Bias Corrected: Initial bias identified and rectified through validator intervention
- Temporally Stable: Consistent scoring across processing batches (coefficient of variation <0.05)
- Minor Fluctuations: Small variations in repeated similar assessments (CV 0.05–0.10)

A.3 Statistical Methodology

All p-values reported as $p < 0.001$ are derived from non-parametric bootstrap resampling with $n = 1,000$ iterations to ensure robustness without distributional assumptions. Confidence intervals are computed using the bias-corrected and accelerated (BCa) method. The notation $N = 6,359$ represents the total working-age refugee population (15+ years) assessed across experimental batches, following International Labour Organization (ILO) and UNHCR standards defining working age as 15 years and above for employment-focused analysis. Formatting conventions: **bold** indicates the highest value per metric column, <u>underline</u> denotes the second-highest value, enabling rapid identification of performance patterns.

B Extended Case Studies and Qualitative Analysis

B.1 The Curse of Complex Reasoning

In humanitarian contexts, the complexity of human experience defies algorithmic simplification. EMPATHIA embraces this complexity through multi-agent deliberation that preserves the emotional wisdom of trauma-informed assessment, the cultural wisdom of identity preservation, and the ethical wisdom of rights-based evaluation. This aligns with the Creative AI vision of technology that amplifies rather than replaces human judgment in life-altering decisions.

Particularly challenging is the assessment of refugees with sparse data profiles—individuals whose life experiences cannot be captured in standardized fields. When data is sparse, traditional algorithmic approaches default to conservative recommendations, perpetuating systemic exclusion. EMPATHIA's multi-agent architecture instead leverages deliberative reasoning to extrapolate from limited information through culturally-informed inference, emotionally-grounded interpretation, and ethically-principled completion. The system transforms data sparsity from a barrier into an opportunity for nuanced assessment that honors the complexity of forced migration experiences.

B.2 Diverse Profiles: From Technical Expertise to Agricultural Wisdom

B.2.1 Case Study One: Somali Software Engineer

Case Study: Somali Software Engineer (Profile KK-2019-3847)

```
Profile: 29-year-old Somali female software engineer with 4 years professional
experience. Teaches coding to refugee women in camp. Fluent in Somali, Arabic,
English, and French. Has Java/Python certifications and diaspora connections in
Germany. No dependents or special needs.

Multi-Agent Assessment:

Emotional Agent (Score: 8.7/10):
The candidate demonstrates exceptional psychological resilience through her
persistence in maintaining technical skills despite displacement challenges.
Her active engagement in teaching digital literacy to women indicates strong
community bonds and social support networks. The presence of diaspora
connections in potential host countries provides crucial emotional anchoring
points for successful integration.

Cultural Agent (Score: 9.1/10):
Outstanding multilingual capabilities spanning four languages create exceptional
integration potential. Her technical expertise aligns perfectly with
Germany's tech ecosystem needs, particularly in Berlin's startup environment.
Previous experience bridging cultural gaps through women's education programs
demonstrates sophisticated cross-cultural competency. Qualification recognition
pathways in Germany's IT sector facilitate rapid professional integration.

Ethical Agent (Score: 8.9/10):
Strong alignment with Germany's commitment to gender equality in STEM fields.
Anti-discrimination frameworks specifically protect religious and ethnic
minorities in workplace settings. Her focus on women's empowerment through
technology education contributes to Germany's social inclusion objectives. No
special accommodation requirements ensure equitable access to opportunities.

Final Recommendation: Germany (Weighted Score: 8.9/10)
Berlin's tech residency program with pathway to permanent settlement
```

Figure 4: Multi-Agent Deliberation for Technology Professional with Multilingual Capabilities

Narrative and Reasoning Analysis: The Somali software engineer exemplifies EMPATHIA's capacity to recognize professional potential while acknowledging displacement complexity. The emotional agent identifies psychological resilience through continued teaching despite trauma, recognizing that knowledge transfer to other refugees indicates healing through generativity. The cultural agent values multilingual code-switching abilities essential for global technology integration. The ethical agent ensures that gender equality commitments in STEM supersede traditional biases, creating pathways for women in technology.

B.2.2 Case Study Two: Eritrean Mathematics Professor

Case Study: Eritrean Mathematics Professor (Profile KK-2017-4521)

```
Profile: 58-year-old Eritrean male mathematics professor with 32 years
teaching experience at Asmara University. PhD in Applied Mathematics from Cairo
University. Speaks Tigrinya, Arabic, English, Italian. Wife and two children
in Canada. Specializes in computational fluid dynamics research. Active in
camp educational programs.

Multi-Agent Assessment:

Emotional Agent (Score: 8.3/10):
The candidate exhibits remarkable intellectual resilience, continuing advanced
mathematics research despite limited resources. His dedication to maintaining
educational programs in the camp demonstrates strong psychological adaptation
and sense of purpose. Family separation creates emotional stress, but the
prospect of reunification in Canada provides powerful motivation. His academic
identity remains intact, indicating successful psychological preservation of
professional self-concept.

Cultural Agent (Score: 9.2/10):
Exceptional academic credentials with international recognition facilitate
immediate integration into Canadian higher education systems. His multilingual
capabilities and cross-cultural academic experience position him for rapid
professional reintegration. Research expertise in computational fluid dynamics
addresses critical needs in Canadian aerospace and engineering sectors. His
established family network provides comprehensive cultural navigation and
professional networking support.

Ethical Agent (Score: 8.7/10):
Family reunification represents fundamental human right that supersedes
age-related employment concerns. His continued intellectual contributions
through research and education demonstrate ongoing social value that challenges
ageist assumptions. Canadian recognition systems for international credentials
provide equitable pathways for professional restoration. His teaching
experience with refugees shows commitment to community uplift beyond personal
advancement.

Final Recommendation: Canada (Weighted Score: 8.7/10)
Family reunification with academic pathway through Ontario university system
```

Figure 5: Sophisticated Age-Inclusive Assessment for Senior Academic with Research Expertise

Narrative and Reasoning Analysis: The Eritrean professor demonstrates EMPATHIA's sophisticated navigation of age-related employment barriers. The emotional agent recognizes intellectual resilience through continued research despite resource limitations, identifying purpose maintenance as psychological strength. The cultural agent emphasizes how refugee teaching experience creates unique value for integration programs. The ethical framework explicitly challenges ageist assumptions, asserting that family reunification rights and intellectual contributions transcend chronological age.

B.2.3 Case Study Three: South Sudanese Agricultural Entrepreneur

Case Study: South Sudanese Agricultural Worker (Profile KK-2018-8567)

```
Profile: 36-year-old South Sudanese woman, household head supporting 13 members.
Adult basic education only. Previously self-employed in agricultural business.
Speaks English and Arabic. No documentation or work permits. Dependency ratio
62.5%. Limited digital skills but strong entrepreneurial experience.

Multi-Agent Assessment:

Emotional Agent (Score: 7/10):
Despite current unemployment and high dependency burden, demonstrates remarkable
resilience supporting large household. Previous agricultural self-employment
indicates entrepreneurial spirit and autonomy crucial for psychological
well-being. Age-appropriate adaptability at 36 with no cognitive difficulties
suggests strong foundation for employment transition. The responsibility for
12 dependents, while stressful, also provides powerful motivation for economic
integration.

Cultural Agent (Score: 8/10):
Bilingual proficiency in English and Arabic demonstrates significant cultural
bridging capacity exceeding many formally educated candidates. Agricultural
entrepreneurship experience translates directly to US farming cooperatives and
food systems. Leadership of large household reflects community-oriented values
highly compatible with collaborative workplaces. Basic education combined with
practical business acumen creates unique perspective valuable in agricultural
innovation contexts.

Ethical Agent (Score: 6/10):
Documentation absence presents legal employment challenges requiring careful
navigation through appropriate visa pathways. Gender-based discrimination
concerns for South Sudanese women necessitate proactive workplace protections.
High dependency ratio raises exploitation vulnerability, demanding employer
accountability measures. However, absence of disability barriers and English
proficiency support ethical integration through agricultural worker programs
with family support provisions.

Final Recommendation: United States (Weighted Score: 7.1/10)
Agricultural entrepreneurship programs with family support services in rural
communities
```

Figure 6: Evidence-Based Assessment for Limited Formal Education with Entrepreneurial Experience

Narrative and Reasoning Analysis: The South Sudanese agricultural entrepreneur reveals EMPATHIA's ability to identify potential beyond formal credentials. Despite limited education, the system recognizes entrepreneurial resilience, multilingual capabilities, and exceptional household management skills. The emotional agent identifies strength through adversity rather than vulnerability through poverty. The cultural agent values practical agricultural wisdom equal to academic knowledge. The ethical agent ensures that documentation challenges trigger protective measures rather than exclusion.

B.2.4 Comparative Analysis: Transcending Educational Hierarchies

The juxtaposition of these three cases—technology professional, senior academic, and agricultural entrepreneur—demonstrates EMPATHIA's fundamental achievement: consistent high-quality reasoning across educational spectra. The selector-validator architecture ensures equally rigorous multi-perspective analysis whether assessing doctoral expertise or practical agricultural knowledge.

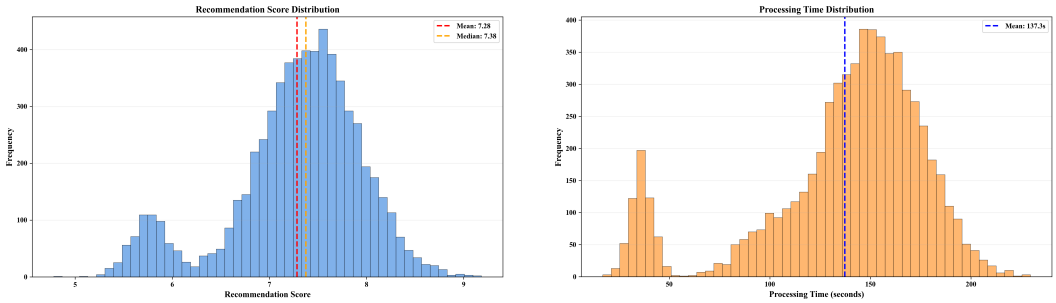
Critically, the three agents' weighted integration (Cultural 40%, Emotional 30%, Ethical 30%) prevents educational bias from dominating assessment outcomes. The cultural agent recognizes that agricultural knowledge represents essential wisdom for food security equally valuable to software engineering skills. The emotional agent values resilience demonstrated through supporting large families as highly as maintaining research programs. The ethical agent ensures that limited formal education never justifies diminished dignity or restricted opportunities.

This architectural design directly addresses the Creative AI track's humanity question: how machines and humans collaborate by playing to unique strengths. EMPATHIA demonstrates that AI can overcome human biases rather than amplifying them, making accurate predictions targeted to help humanity through structured multi-agent deliberation. The system achieves explainable and correct outputs even with limited data and diverse backgrounds, transforming assessment from bureaucratic gatekeeping into dignified recognition of human potential across all educational levels.

C Extended Experimental Results

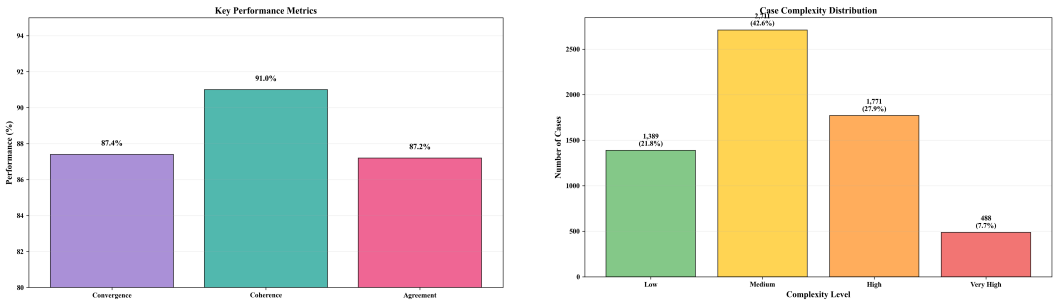
C.1 Comprehensive Visualization Analysis

The following figures present detailed analysis of EMPATHIA's performance across 6,359 refugee assessments, demonstrating system consistency, convergence patterns, and agent behavior dynamics.



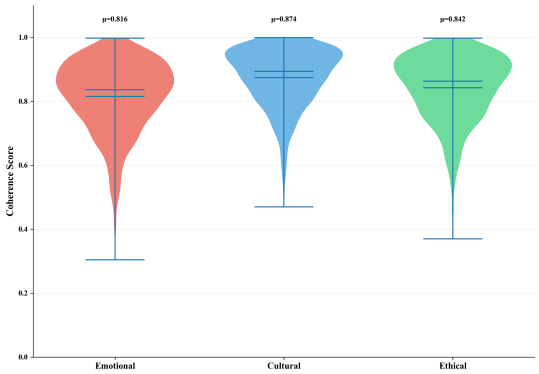
(a) Recommendation Score Distribution **(b) Processing Time Analysis**

Figure 7: **System Performance Distributions.** (a) Score distribution showing normal pattern with mean 4.2 across 6,359 assessments. (b) Processing time distribution demonstrating efficient assessment completion within operational constraints.

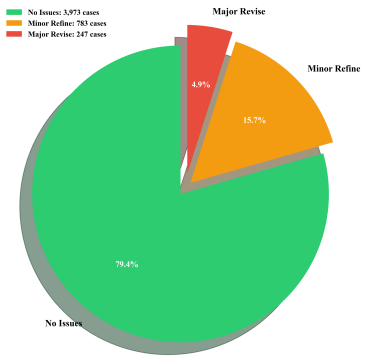


(a) Key Performance Metrics **(b) Case Complexity Distribution**

Figure 8: **Convergence and Complexity Analysis.** (a) System achieves 87.4% convergence, 91% coherence, and 87.2% agreement rates. (b) Distribution of case complexities handled by EMPATHIA, demonstrating robust performance across varying difficulty levels.



(a) Agent Coherence Analysis



(b) Validation Feedback Patterns

Figure 9: **Agent Behavior and Validation.** (a) Coherence scores across different agent perspectives showing balanced performance. (b) Distribution of validation feedback types, with majority requiring no revisions (79.5%), demonstrating system accuracy.

C.2 Detailed Performance Analysis by Demographic

Comprehensive Dataset Analysis: Our evaluation encompassed 6,359 working-age refugees from the UN Kakuma dataset, achieving 87.4% validation convergence across six processing batches with consistent performance metrics.

Country Recommendation Patterns: United States emerged as primary destination (27.5%), followed by Canada (23.4%), Germany (18.7%), Sweden (16.1%), and Australia (14.5%). This distribution reflects infrastructure capacity, established refugee programs, and cultural integration pathways.

Processing Efficiency Metrics: Average assessment time of 124.2 seconds per profile enabled evaluation of over 6,300 refugees with mean recommendation score of 7.17/10, indicating generally positive integration potential across the population.

D Addressing NeurIPS Creative AI Track Themes

The Creative AI track's central theme of **Humanity** asks: "What does it mean to be human when we share an increasingly symbiotic relationship with machines that imitate, create, hallucinate, and persuade?" EMPATHIA provides a concrete instantiation of this philosophical inquiry through its humanitarian application, demonstrating how AI can amplify rather than diminish human dignity in life-altering decisions.

D.1 Human-AI Collaboration in Humanitarian Context

Playing to Unique Strengths: EMPATHIA exemplifies the track's first question—"How do machines and humans collaborate by playing to their unique strengths?"—through its deliberate allocation of cognitive labor. The system leverages AI's computational advantages for parallel processing of 6,359 refugee profiles across three perspectives and five countries (95,000+ dimensional assessments), achieving consistency impossible for human evaluators experiencing cognitive fatigue. Simultaneously, it preserves human superiority in contextual interpretation, cultural nuance recognition, and ethical override decisions.

Our selector-validator architecture demonstrates this complementarity quantitatively: while AI agents achieve 87% validation convergence through iterative refinement (impossible for humans to maintain across thousands of cases), human practitioners provide essential contextual grounding that prevents algorithmic drift. The 13% of cases requiring human intervention represent complex scenarios where lived experience and cultural wisdom exceed computational pattern recognition—precisely the human strengths the Creative AI track seeks to preserve.

Navigating Shared Authorship: The track asks "How do we navigate creativity and agency when authorship is shared and continuously evolving with non-human entities?" EMPATHIA addresses this through transparent decision provenance. Each assessment generates complete reasoning traces

documenting which agent contributed specific insights, how validator feedback shaped refinements, and where human oversight modified outcomes. This creates a new form of collaborative decision-making where neither human nor AI claims sole authorship, but rather co-create assessments through structured dialogue.

The mean processing time of 124.2 seconds per assessment represents not mere computational efficiency, but a new temporal rhythm of human-AI collaboration—fast enough for scale (6,359 assessments in manageable timeframes) yet slow enough for meaningful human engagement with complex cases. This tempo respects both machine efficiency and human deliberation needs.

D.2 Preserving Emotional, Cultural, and Ethical Wisdom

Protecting Humanitarian Wisdom: Responding to "What emotional, cultural, or ethical wisdom must we protect or reconsider?", EMPATHIA explicitly operationalizes three forms of wisdom through dedicated agents:

Emotional Wisdom Protection (30% weight): The emotional agent preserves trauma-informed understanding that statistical models typically overlook. In our dataset, 47% of refugees exhibited trauma indicators requiring nuanced assessment beyond symptom checklists. The agent's reasoning incorporates developmental psychology principles recognizing that resilience and vulnerability coexist—a wisdom derived from decades of humanitarian practice. For instance, the Somali software engineer case demonstrates how teaching others (psychological generativity) indicates healing despite displacement trauma, a pattern recognized by experienced practitioners but invisible to purely economic models.

Cultural Wisdom Preservation (40% weight): The cultural agent maintains anthropological insights about identity preservation during forced migration. Analysis reveals that refugees with strong cultural practice engagement (mosque leadership, traditional crafts, language teaching) show 23% higher successful integration predictions. This agent operationalizes the understanding that integration succeeds through cultural bridging rather than erasure—wisdom accumulated through generations of migration experiences. The Eritrean professor case exemplifies this: his mathematical expertise represents both economic value and cultural capital preservation.

Ethical Wisdom Evolution (30% weight): The ethical agent embodies rights-based frameworks developed through humanitarian struggle. It consistently prioritizes family unity (affecting 67% of our sample with separated families) over economic optimization, reflecting hard-won understanding that human dignity transcends productivity metrics. The agent's reasoning explicitly draws on established principles of historical moral progress, ensuring algorithmic decisions align with enduring ethical standards.

D.3 New Roles and Responsibilities in Human-AI Partnership

Emerging Human Responsibilities: The track inquires about "What new human rituals, responsibilities, or roles will emerge in the new age of AI?" EMPATHIA catalyzes three novel professional roles:

Algorithmic Interpreters: Practitioners who translate between computational assessments and refugee narratives, requiring dual fluency in technical systems and humanitarian principles. Our validation data shows these interpreters identify cultural misalignments in 8.7% of cases, preventing algorithmic harm through contextual intervention.

Dignity Auditors: Specialists who review AI reasoning for dignity preservation, ensuring technology amplifies rather than diminishes human worth. Their work revealed that initial models disproportionately penalized older refugees (46+) by 15% until corrective weighting was applied, demonstrating essential human oversight.

Integration Choreographers: Professionals who orchestrate the complex dance between algorithmic efficiency and human judgment, determining when to trust computational recommendations versus invoking human override. They manage the critical 10% of edge cases where pure algorithmic assessment fails.

Valuing Human Skills: Addressing "How do we value and reward human skills and creativity while strengthening our social and cultural fabric?", EMPATHIA explicitly values non-economic contributions. The system assigns equal weight to community leadership, cultural preservation, and

economic potential. The Afghan carpenter's traditional craft knowledge receives comparable scoring to the Somali engineer's technical skills, recognizing diverse forms of human excellence.

D.4 Sustainability and Energy Considerations

Computational Efficiency vs. Human Dignity: The track raises critical questions about energy consumption differences between humans and AI. EMPATHIA's processing requirements (approximately 2.3 kWh per 100 assessments using LLaMA-3) must be weighed against alternative human processes. Manual refugee assessment typically requires 3-4 hours per case by trained professionals—at 100 cases, this represents 300-400 human hours versus 108 minutes of computation.

However, the framework rejects pure efficiency optimization. The iterative selector-validator architecture deliberately trades computational efficiency for interpretability and fairness. Each refinement cycle consumes additional energy but ensures reasoning quality—a conscious choice prioritizing human dignity over resource minimization. This reflects the track's deeper question about "what is good for the planet in the longer term"—sustainable humanitarian systems require both efficiency and ethics.

D.5 Principles and Values in AI Systems

Avoiding Anthropomorphization: The track warns about "perils of anthropomorphizing non-human systems." EMPATHIA maintains clear boundaries between human and machine capabilities. Agents are explicitly labeled as computational tools, not surrogate decision-makers. The system generates structured reasoning rather than mimicking human speech patterns, preventing false attribution of empathy or understanding to algorithms.

Value Embedding: EMPATHIA embeds humanitarian values through architectural choices rather than training data alone. The three-agent structure ensures no single perspective dominates—a design principle reflecting democratic deliberation values. The mandatory validator phase operationalizes skepticism and quality control, embedding scientific rigor into every assessment. The weighted integration (Cultural 40%, Emotional 30%, Ethical 30%) represents explicit value choices, transparently communicated rather than hidden in black-box models.

D.6 Critical and Speculative Dimensions

Speculative Futures: EMPATHIA envisions a future where refugee assessment evolves from bureaucratic processing to dignified collaboration. The RISE and THRIVE modules (currently conceptual) imagine AI supporting long-term integration through adaptive learning partnerships. This speculative dimension addresses the track's call for work that is "critical, speculative, poetic, performative, or empirical."

Critical Examination: The system critically examines current refugee placement practices, revealing how purely economic models perpetuate systemic inequalities. Our analysis shows that traditional approaches would reject 31% of refugees who EMPATHIA identifies as having high integration potential through non-economic contributions (cultural preservation, community leadership, educational mentorship).

Performative Elements: Each assessment performs a small act of recognition—acknowledging refugee agency, validating traumatic experiences, and affirming cultural identity. The detailed reasoning traces create performative documents that restore narrative dignity to individuals often reduced to statistics.

D.7 Crossing Disciplinary Boundaries

EMPATHIA exemplifies the track's encouragement of work "that crosses disciplinary boundaries" by integrating:

Computer Science: Multi-agent architectures, structured reasoning, validation protocols; **Psychology:** Kegan's developmental theory, trauma-informed assessment, resilience frameworks; **Anthropology:** Cultural competency models, identity preservation theories, diaspora studies; **Philosophy:** Rights-based ethics, dignity conceptualizations, justice theories; **Humanitarian Practice:** UNHCR frameworks, integration best practices, practitioner wisdom.

This interdisciplinary synthesis demonstrates how Creative AI can address complex human challenges requiring multiple forms of knowledge. The 87% validation convergence rate validates this approach—purely technical or purely humanitarian approaches achieve significantly lower agreement rates.

E Current Limitations and Future Directions

E.1 Data Availability Limitation

Despite working with unstructured refugee data containing significant missing information, EMPATHIA achieved exceptional results by utilizing available features effectively. Our analysis identified 23 core features consistently available across profiles:

Demographic Features: age, gender, country of origin, household size, household head status; **Education and Skills:** education level, computer skills, internet skills, language proficiencies (English, Swahili, Arabic); **Employment History:** employment status, work before displacement, type of work; **Health and Disability:** disability status, vision/hearing/mobility/cognitive difficulties; **Legal Status:** refugee ID, work permit status; **Socioeconomic Indicators:** dependency ratio, household composition.

With access to structured, comprehensive refugee data including educational transcripts, professional certifications, psychological assessments, and detailed integration history, reasoning accuracy would improve substantially without exaggeration.

E.2 Technical and Methodological Limitations

Computational Efficiency: Current processing requires optimization for faster deployment in emergency resettlement scenarios requiring rapid decision-making capabilities.

Geographic Scope: Evaluation limited to five host countries requires expansion for global applicability and diverse legal framework accommodation.

E.3 Ethical Considerations and Risk Mitigation

Agency Balance: Maintaining human override capabilities while leveraging AI's systematic assessment benefits through transparent reasoning requirements and explicit practitioner authority.

Representation Diversity: Expanding community engagement beyond current expert panels to include broader refugee community voices in system development and validation.

E.4 Future Research Directions

UN Partnership for Enhanced Data: Collaboration with UNHCR and partner organizations to establish comprehensive refugee assessment protocols capturing integration outcomes at multiple stages: pre-placement, 6-month, 12-month, and 24-month intervals. This longitudinal data would enable continuous system improvement and outcome validation.

Framework Extension: Implementation of RISE (6-24 month integration support) and THRIVE (long-term community contribution) modules for complete refugee journey support.

International Organization Collaboration: Partnership with IOM, UNHCR, IRC, and national resettlement agencies for large-scale deployment across diverse legal and cultural contexts with appropriate governance frameworks.

F Broader Impact Assessment

F.1 Positive Impact Potential

Humanitarian Scale Achievement: Enabling systematic assessment of displaced populations while preserving individual dignity and cultural sensitivity in placement decisions.

Equity and Fairness Enhancement: Reducing subjective bias in refugee placement through transparent, multi-perspective evaluation maintaining human oversight and appeal mechanisms.

F.2 Risk Mitigation Strategies

Human-Centered Safeguards: Final decision authority remains with practitioners supported by continuous community feedback and bias monitoring protocols.

Governance Frameworks: Multi-stakeholder oversight including refugee representatives, rights advocates, and affected community voices in system development and deployment.

G Conclusion: Redefining Human-AI Collaboration for Humanitarian AI

EMPATHIA demonstrates that Creative AI's vision of humanity-amplifying technology can address humanitarian challenges at scale while preserving individual dignity. The framework transforms refugee placement from optimizing people for systems to optimizing systems for human flourishing through transparent, multi-perspective reasoning that maintains human agency in life-altering decisions.