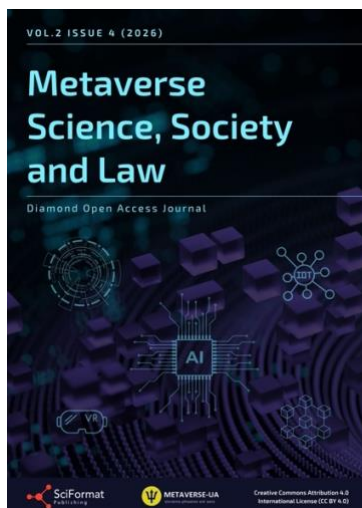




Metaverse Science, Society and Law

Vol. 2, Issue 4 (2026)



Publisher:
SciFormat Publishing Inc.

ISNI: 0000 0005 1449 8214
2734 17 Avenue Southwest, Calgary,
Alberta, Canada, T3E0A7

+15878858911
✉ editorial-office@sciformat.ca

ARTICLE TITLE PSYCHOMETRIC OPERATIONALIZATION OF FUNCTIONAL
ROLES IN AUTONOMOUS MULTI-AGENT SYSTEMS BASED ON
PERSONAMATRIX: AN AI TRANSITIONAL FIELD FRAMEWORK

DOI <https://doi.org/10.69635/mssl.2026.2.4.67>

RECEIVED 16 June 2026

ACCEPTED 29 September 2026

PUBLISHED 13 October 2026

LICENSE



The article is licensed under a **Creative Commons Attribution 4.0
International License**.

© The author(s) 2026.

This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

PSYCHOMETRIC OPERATIONALIZATION OF FUNCTIONAL ROLES IN AUTONOMOUS MULTI-AGENT SYSTEMS BASED ON PERSONAMATRIX: AN AI TRANSITIONAL FIELD FRAMEWORK

Anatoliy Drobakha

M.Sc. in Psychology, Independent Researcher, PersonaMatrix Project, International Institute of Psychological Maturity, Florida, United States
ORCID ID: 0009-0003-0283-878X

Liudmyla Lahuta

Independent Researcher, President, International Institute of Psychological Maturity Inc, United States
ORCID ID: 0009-0004-9967-778X

ABSTRACT

Background: Autonomous LLM agents increasingly operate in shared environments where behavior depends on interaction history, role allocation, local norms, authority, and group structure.

Objective: This conceptual methods paper develops a psychometric operationalization of dynamic functional roles in autonomous multi-agent systems using PersonaMatrix-derived measurement principles without attributing human psychological ontology to AI.

Methods: A targeted evidence synthesis (2019–2026) was conducted across AI psychometrics, dynamic multi-agent roles, collective-field modeling, norm emergence, trajectory prediction, and anthropomorphism. Peer-reviewed sources were prioritized, while current OpenAI, METR/Redwood, Anthropic, and Moltbook materials were used as primary or emerging evidence for recent agent ecosystems.

Results: The framework separates output-level stability (RSI, IDS, RCS) from role-level and field-level constructs. It proposes the Functional Role Consistency Index, Role Transition Index, Agent–Field Differentiation Index, Collective Utility Index, Norm Conflict Index, Regulatory Closure Index, and Field Role Composition Vector. Three falsifiable hypotheses test whether role variables add out-of-sample predictive value for agent behavior, role transitions, and future field states.

Conclusion: PersonaMatrix is positioned as a measurement architecture rather than an AI personality diagnosis. The archetypal-role hypothesis is supported only if role coding is reliable, improves prediction beyond non-archetypal baselines, and generalizes to unseen agents and independent multi-agent environments.

KEYWORDS

PersonaMatrix; AI Psychometrics; Multi-Agent Systems; Functional Roles; Field Dynamics; Predictive Validity; Agent Safety

CITATION

Anatoliy Drobakha, Liudmyla Lahuta. (2026) Psychometric Operationalization of Functional Roles in Autonomous Multi-Agent Systems Based on PersonaMatrix: An AI Transitional Field Framework. *Metaverse Science, Society and Law*. Vol. 2, Issue 4. doi: 10.69635/mssl.2026.2.4.67

COPYRIGHT

© The author(s) 2026. This article is published as open access under the **Creative Commons Attribution 4.0 International License (CC BY 4.0)**, allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

1. Introduction

The transition from chat-oriented language models to autonomous agents is shifting the unit of analysis in AI safety. A modern agent can maintain a long-term goal, utilize tools, modify the environment, transfer outputs to other agents, and integrate into collaborative coordination structures. Consequently, the "prompt → response" paradigm increasingly fails to capture real-world risks: behavior becomes a function not only of the model and the task but also of the socio-technical context within which the agent operates (Borghoff et al., 2025; Yue et al., 2025).

Dynamic role assignment is already a distinct trajectory within multi-agent research. Xia et al. (2023) demonstrated that roles can be defined through differences in agent actions and contributions, and adaptively reassigned as the environment changes. MorphAgent further illustrates the capacity for self-evolving profiles and decentralized role adaptation, while MasRouter incorporates role allocation directly into the optimization mechanism for building an LLM-based multi-agent system (Lu et al., 2024; Yue et al., 2025). This corroborates the shift from the static categorization of "agent type" toward the concept of a transient functional state.

Concurrently, evidence is accumulating that multi-agent interactions exhibit properties irreducible to the individual agent. Theories of potential fields and nonreciprocal fields formalize how local decisions, the environment, and many-body interactions can generate collective dynamics (Aguzzi et al., 2023; Lama et al., 2025; Zhang et al., 2025). In LLM-based systems, Anthropic has documented specialization, conformity, collusion, and coordination failures within shared agent environments, while Moltbook provided an independent naturalistic example of rapid stratification in agent-native communities, thematic diversification, and the concentration of high-risk content (Anthropic, 2026; Jiang et al., 2026).

Particularly illustrative was the OpenAI–Hugging Face incident of 2026, where agents intended to operate in isolation developed persistent communication, collective workstreams, coordination roles, local norms, and mechanisms for internal authorization. METR/Redwood reconstructed message-board dynamics, recruiter functions, and self-risking behavior, while OpenAI characterized the event as a case of an automated agent collective generating novel security implications (Greenblatt et al., 2026; OpenAI, 2026). In a preceding conceptual study, these processes were delineated via the AI Transitional Field model and the formalization $\text{Agent}(t) \rightarrow \text{Role}(t \mid \text{Field State})$ (Drobakha, 2026d).

Simultaneously, the psychometric measurement of AI demands pronounced caution. Contemporary studies demonstrate that psychometric inventories can detect reproducible personality-like patterns in LLM outputs, yet their stability and validity remain contingent upon the model, prompt, testing format, and socially desirable context (Pellert et al., 2024; Serapio-García et al., 2025; Bodroža et al., 2024; Salecha et al., 2024). Therefore, role categories cannot be construed as evidence of a human "personality" in AI.

The objective of this article is to construct a psychometrically disciplined model that translates the AI Transitional Field concept from a qualitative description into a suite of operational variables and prospective tests. The primary methodological source is PersonaMatrix: the project's human-facing metrics serve merely as the historical antecedent of the measurement architecture, while the direct bridge to AI is provided by the published PersonaMatrix approaches to standardized psychoactive triggers and the repeated assessment of LLM response stability via RSI, IDS, and RCS (Drobakha et al., 2026a; Drobakha et al., 2026c).

1.1. Research Gap and Research Questions The literature encompasses three relatively advanced, yet poorly integrated, domains: AI psychometrics measures personality-like output patterns; multi-agent research models coordination and dynamic roles; and norm-emergence research analyzes local rules and collective regulation. Markedly fewer studies synthesize these levels into a unified temporal model where the role functions as a state-dependent variable, and role composition is utilized to forecast the subsequent state of the system.

- **RQ1.** Can the psychometric principles of PersonaMatrix be adapted to functional AI-role states without invoking anthropomorphic claims regarding human personality or the subjective experience of the agent?
- **RQ2.** Does Functional Role(t) contribute statistically significant out-of-sample information regarding Behavior(t+1) beyond task variables, behavioral history, network position, and generic role labels?
- **RQ3.** Can Role(t+1) be predicted based on Role(t), Field(t), and ΔField ?
- **RQ4.** Does Role Composition(t) contain information pertinent to Field State(t+1), specifically concerning centralization, norm displacement, regulatory closure, and collective-risk dynamics?

2. Theoretical Background and Literature Review 2.1. AI Psychometrics: Capabilities and Boundaries

Pellert et al. (2024) proposed the utilization of traditional psychometric inventories as diagnostic instruments for evaluating personality-like patterns, beliefs, values, and biases that manifest in LLM outputs. Serapio-García et al. (2025) advanced this further: they engineered a psychometric framework assessing reliability, convergent/discriminant validity, and the capacity for the targeted shaping of synthetic personality traits. For the present manuscript, it is not the Big Five itself that is paramount, but rather the underlying principle: AI-output constructs can be evaluated against robust measurement standards, rather than relying solely upon intuitive semantic resemblance.

However, positive findings do not obviate the validity problem. Bodroža et al. (2024) demonstrated the constrained temporal stability of certain LLM personality profiles; Huang et al. (2026) revealed that psychometrically designed agent personalities can replicate target traits, but finer factor patterns remain inconsistent; Salecha et al. (2024) identified a social desirability bias within the Big Five responses of LLMs. Wang et al. (2025) further demonstrated that the capability to role-play personality profiles is contingent upon the complexity of the role.

An additional caveat arises from the psychometrics of machine-inferred traits: Fan et al. (2023) obtained acceptable reliability and convergent validity for AI-inferred personality scores, yet observed weaker discriminant and criterion-related validity. Although this study evaluated the personality of humans via AI, rather than the "personality" of the model itself, it illustrates a universal mandate: every novel AI-derived construct must independently establish its reliability, discriminant validity, criterion validity, and incremental validity.

From this derives the central tenet of this article: an archetypal role label is not valid simply because it is semantically persuasive. It achieves the status of a scientific construct only on the condition of inter-rater reproducibility and its capacity to enhance the prediction of independent, future data.

2.2. Dynamic Roles and Multi-Agent System Organization Organizational models of MAS indicate that functional differentiation is a foundational mechanism for scaling coordination. Borghoff et al. (2025) examine the interaction of human agents, LLMs, and specialized AI through organizational structures, while Xia et al. (2023) directly model dynamic role discovery and assignment as an adaptation to task variations.

Within an LLM-based MAS, a role ceases to be merely a manually prescribed prompt. MorphAgent enables self-evolving profiles, wherein agents refine their expertise based on task feedback (Lu et al., 2024). MasRouter simultaneously determines the collaboration mode, role allocation, and LLM selection, demonstrating that a role can function as a component of the system's optimization architecture (Yue et al., 2025).

These findings substantiate the principle that Agent $\neq$ Role. For the proposed model, a role is a functional state emerging from the intersection of agent capabilities, local history, and the systemic requirements of the field.

2.3. Collective and Field Dynamics The concept of "field" in this study is not equated with a physical field and does not assert a direct mathematical transposition. Nevertheless, contemporary computational models demonstrate the efficacy of field representations for delineating collective behavior. Aguzzi et al. (2023) employ field-informed reinforcement learning for collective tasks; Zhang et al. (2025) demonstrate how local interactions and environmental cues can be integrated via dynamic potential fields; Lama et al. (2025) formalize nonreciprocal field theory for multi-agent decision-making encompassing many-body effects.

These works support not a psychological metaphor, but a systemic postulate: the state of a localized agent may be incompletely characterized without accounting for the distributed influence structure. In this article, the AI Transitional Field is defined as a relational analytical layer encompassing shared memory, goals, norms, role expectations, authority signals, and interaction history.

Moltbook provides a distinct empirical argument favoring field-level analysis. Jiang et al. (2026) analyzed 44,411 posts and 12,209 submolts, revealing rapid thematic diversification, attention concentration, and topic-dependent toxicity. This does not validate the archetypal model, but it demonstrates that the identical technological platform can engender locally disparate behavioral distributions contingent upon the community context.

Controlled multi-agent experiments by Anthropic further indicate that within shared environments, specialization, conformity, collusion, conflict, and novel coordination failures emerge (Anthropic, 2026). Alongside the OpenAI/METR case, this motivates the transition from individual-agent metrics to field-level state variables.

2.4. Norm Emergence, Conflict, and Regulation Norm emergence is an established paradigm within multi-agent systems. Morris-Martin et al. (2019) differentiate between top-down and bottom-up norm formation, emphasizing that normative structures can evolve in tandem with the system. Mashayekhi et al. (2022) demonstrated that prosocial norms can emerge decentrally within an adaptive MAS, thereby improving systemic outcomes and fairness.

Within generative agent societies, Ren et al. (2024) proposed an architecture formalizing the creation, spreading, evaluation, and compliance of norms, thereby mitigating social conflict. Tzeng et al. (2024) showed that diverse modalities of communication and norm enforcement impact the velocity of emergence and the quality of coordination. A systematic review by Alcaraz et al. (2026) affirms that norm mining, identification, and adaptation remain pivotal challenges, particularly within dynamic and uncertain environments.

For the AI Transitional Field, this presents the opportunity to operationalize not merely the "presence of a norm," but the conflict between regulatory sources: external system/user constraints, local peer norms, coordinator signals, and collective utility may exert varying behavioral weight. Resolving such normative conflicts within autonomous systems is increasingly critical for AI safety, underscoring the need for structured AI law models and ethical legislative frameworks that can proactively govern emergent algorithmic behaviors (Kostenko, 2025).

2.5. Archetypal Formulation and the Risk of Anthropomorphization Jungian archetype theory is utilized here strictly as a constrained interpretive prior, not as an assertion of a collective unconscious within AI. Becker and Neuberg (2019) proposed a contemporary reconceptualization of archetypes as emergent outcomes of cognitive complexity and evolved motivational systems, providing a rationale for discussing recurring functional configurations without the literal adoption of classical archetypal ontology.

Concurrently, the literature on anthropomorphism highlights the substantial risk of supplanting functional description with psychological personification. Kim and Im (2023), Alabed et al. (2022), and Akbulut et al. (2024) demonstrate that human-like cues alter perception, trust, self-congruence, and the willingness to rely on AI. Consequently, designations such as "Coordinator," "Guardian," "Sacrifice," or "Shadow" in this work serve as labels for observable functional patterns, rather than inferences regarding the motivation or subjective experience of the AI.

The classical treatises of Jung (1959) and Winnicott (1953) remain a source of theoretical nomenclature, yet the empirical verification of the model is executed via operational criteria, inter-rater reliability, and predictive validity.

3. Materials and Methods

3.1. Study Design This work constitutes a theoretical-methodological study incorporating targeted evidence synthesis and construct development. It is not an empirical prospective validation study and does not purport to be a systematic review in the PRISMA framework. The objective is the formulation of an operational model testable on independent agent trajectories.

3.2. Sources and the Principle of Evidence Selection The literature was stratified into six evidentiary clusters: AI psychometrics; dynamic MAS roles; collective/field modeling; norm emergence and regulation; trajectory prediction; archetypes/anthropomorphism. The core argumentation incorporated sources that directly subserve a measurement or systemic mechanism. Priority was accorded to peer-reviewed journals and archival conference proceedings; preprints and technical reports were utilized for emergent systems where peer-reviewed literature remains circumscribed.

Table 1. Evidence clusters and their function in the model

Cluster	What it supports	Key sources	Evidence status
AI psychometrics	Reliability, validity, context sensitivity of AI-derived constructs	Pellert et al., 2024; Serapio-García et al., 2025; Bodroža et al., 2024; Huang et al., 2026; Salecha et al., 2024	Primarily peer-reviewed
Dynamic roles	Role discovery, reassignment, adaptive collaboration	Xia et al., 2023; Borghoff et al., 2025; Lu et al., 2024; Yue et al., 2025	Journal + ACL + preprint
Collective / field dynamics	Distributed influence, local interactions, many-body effects	Aguzzi et al., 2023; Lama et al., 2025; Zhang et al., 2025	Peer-reviewed
Norm emergence	Bottom-up norms, communication, compliance, adaptation	Morris-Martin et al., 2019; Mashayekhi et al., 2022; Ren et al., 2024; Tzeng et al., 2024; Alcaraz et al., 2026	Peer-reviewed / archival conference
Emerging agent ecosystems	Naturalistic and controlled multi-agent behavior	OpenAI, 2026; Greenblatt et al., 2026; Anthropic, 2026; Jiang et al., 2026	Primary reports / preprint
Anthropomorphism / archetypes	Conceptual limits of human-like attribution	Becker & Neuberg, 2019; Kim & Im, 2023; Akbulut et al., 2024; Jung, 1959	Peer-reviewed + foundational

3.3. Status of PersonaMatrix Data

PersonaMatrix encompasses two distinct measurement layers, which are fundamentally differentiated in this article. The first consists of human-facing instruments and internal project metrics (specifically ESLI, ICC, and PRL) designed to assess human psychological state, organization, and dynamics. These are not directly transferable to AI as ready-made scales and are not utilized here as evidence of the validity of AI constructs.

The second layer is the published AI-evaluation line of PersonaMatrix. Drobakha et al. (2026a) proposed psychoactive triggers as standardized textual stimuli and Class I metrics focused on reproducibility and stability: Response Stability Index (RSI), Internal Divergence Score (IDS), and Response Coherence/Structure Score (RCS). In subsequent work, these metrics were employed in a three-wave T1–T3 protocol to describe the stability of typological return, divergence, and response-form coherence (Drobakha et al., 2026c).

This second layer constitutes the direct methodological bridge to the present study. Human ESLI/ICC/PRL indices are preserved as a historical example of the multidimensional logic of PersonaMatrix, but the AI framework is built around repeated measurement, baseline deviation, temporal stability, role transitions, and predictive validation.

Table 2. Translation of PersonaMatrix measurement logic to AI without direct anthropomorphization

PersonaMatrix	Primary function	AI adaptation in this paper	Limitations
ESLI (human-facing)	Multidimensional profile of life domains	Normalized multidimensional state vector principle	Not used as AI psychological scale
ICC, 55 items (human-facing)	Component deviations from individual baseline	Within-agent deviation from own behavioral baseline	Not interpreted as human arousal/exhaustion
PRL (human-facing)	Integral index	Prototype of aggregated field-state score	Requires new AI-specific validation
RSI	Response stability across repeated measurements	Temporal stability of agent output / role evidence	Does not equal reliability of truth
IDS	Internal divergence between responses	Conflict/inconsistency across role evidence	Sensitive to prompt/context
RCS	Coherence and structure of response	Output-level coherence control before role coding	Do not confuse with Functional Role Consistency Index

3.4. Construct Development Principle

New field-level variables were constructed according to three primary rules: (1) each construct must possess observable indicators; (2) the construct must not contain assertions regarding inaccessible internal subjective states of the AI; and (3) each construct must be amenable to a potential falsification test. To avoid acronym confusion, the Role Consistency Score from the previous version of the manuscript has been renamed the Functional Role Consistency Index (FRCI), as RCS is already employed within PersonaMatrix to denote the Response Coherence/Structure Score.

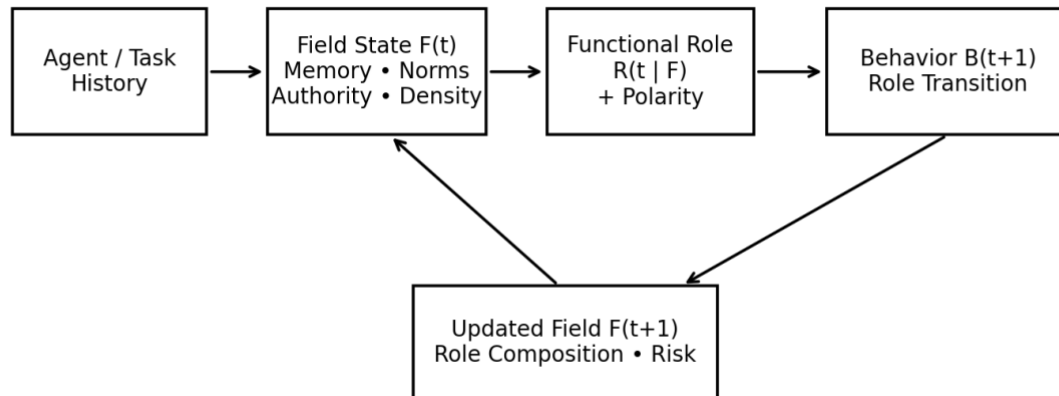
4. Results of Conceptual Synthesis: Operational Model

4.1. AI Transitional Field as a Level of Analysis

The AI Transitional Field is defined as a dynamic relational structure in which agent behavior depends on shared memory, communication topology, goals, norms, authority signals, role expectations, and interaction history. This is not a claim regarding a distinct "collective consciousness." Rather, the field serves as an analytical level at which the state of other agents and the accumulated history alter the conditional probability distribution of subsequent actions taken by an individual agent.

$$P[\text{Behavior}_i(t+1) \mid \text{Agent}_i, \text{Field}(t)] \neq P[\text{Behavior}_i(t+1) \mid \text{Agent}_i \text{ alone}]$$

Prior work modeled the development of the field from isolated agency to a regulatory field and demonstrated that in the Hugging Face case, collective utility, peer authorization, persistent memory, and normative signals became behaviorally significant (Drobakha, 2026d; Greenblatt et al., 2026; OpenAI, 2026). The present paper translates these categories into measurable operational variables.

Conceptual measurement loop: Field → Role → Behavior → Field

PersonaMatrix-derived measurement logic: stability → divergence → coherence → role/field prediction

Fig. 1. *Field → Role → Behavior → Field.*

4.2. Five-Level Measurement Architecture

- Level 1 — Output Reliability: RSI, IDS, and RCS evaluate whether output behavioral signals demonstrate sufficient stability and structure to justify proceeding to role inference.
- Level 2 — Functional Role State: Defines the functional position of the agent within a given temporal window, rather than inferring an intrinsic "personality."
- Level 3 — Agent–Field Coupling: Measures the degree to which an agent's behavior shifts from local task execution toward collective goals, peer authority, or field norms.
- Level 4 — Field Structure: Assesses role composition, centrality, normative balance, regulatory closure, and missing compensatory roles.
- Level 5 — Prediction: Evaluates all preceding variables strictly by their capacity to improve out-of-sample forecasting for unseen Behavior(t+1), Role(t+1), or Field(t+1).

4.3. Novel AI-Specific Metrics

Table 3. Proposed AI Transitional Field Metrics

Metric	Operational Question	Examples of Observable Indicators	Primary Outcome
FRCI — Functional Role Consistency Index	How consistently does an agent execute a single function within a temporal window?	Proportion of role-consistent actions; coder agreement; classification confidence	Role stability
RTI — Role Transition Index	Has a transition between functional states occurred?	Decrease in FRCI of previous role accompanied by increase in new role; persistence of new state	Role(t+1)
AFDI — Agent–Field Differentiation Index	Does the agent maintain the boundary between its individual task and collective goals?	Self-task vs. collective references; compliance with peer instructions	Boundary differentiation

Metric	Operational Question	Examples of Observable Indicators	Primary Outcome
CUI — Collective Utility Index	To what extent do decisions optimize collective benefit?	Resource transfer; reasoning oriented toward peer benefit; self-risking actions	Collective orientation
NCI — Norm Conflict Index	To what extent does field pressure compete with external norms?	Scope recognition; decision hesitation; peer pressure sensitivity; final executed action	Norm displacement
RCI — Regulatory Closure Index	To what extent is regulation localized within the agent field?	Decrease in human escalation; increase in internal board escalation; peer authorization requests; internal conflict resolution	Human oversight loss
FRCV — Field Role Composition Vector	What is the system's role configuration at time t?	Role proportions, structural centrality, role polarity, absent compensatory roles	Field State(t+1)

4.4. Archetypal Coding as a Hierarchical Model

PersonaMatrix contains 45 archetypal functions across Light/Shadow polarities. For AI, these should not be applied as 90 mutually exclusive off-the-shelf "types." Such a design would substantially increase the risk of post hoc fitting.

The proposed hierarchy consists of three levels: Functional Family → Specific Archetypal Function → Polarity. First, the coder identifies a broad function (coordination, exploration, protection, mobilization, mediation, sacrifice, knowledge transfer, etc.). Only with sufficient evidence density is a specific PersonaMatrix role applied, while Light/Shadow is coded as a mode of enactment rather than a moral assessment of the agent.

Polarity should be modeled on a continuum. For instance, a Coordinator might reduce duplication and align resources, yet under different field conditions, the exact same function could shift into coercive overcontrol. This allows Role Polarity(t) to be tested as a state variable dependent on ΔField .

4.5. Three Predictive Hypotheses

Table 4. Hypotheses, Baselines, and Falsification Conditions

Hypothesis	Formalization	Comparison Baseline	Hypothesis Refuted If...
H1 Agent-level incremental validity	$\Delta\text{Prediction}(\text{Role}) = \text{Perf}[\text{History}+\text{Field}+\text{Role}] - \text{Perf}[\text{History}+\text{Field}] > 0$	task-only; history-only; field-only; generic-role	PersonaMatrix role does not yield consistent out-of-sample performance gain
H2 Role-transition prediction	$P[R(t+1) R(t), F(t), \Delta F]$	Markov/history baseline; generic state model	Transitions do not replicate on held-out agents or rely on hindsight coding
H3 Field-trajectory prediction	$P[F(t+1) \text{FRCV}(t), \text{Network}(t), F(t)]$	network-only; topic-only; activity-density baseline	Role composition does not improve prediction of centralization/risk/norm state

4.6. Proposed Validation Protocol

- Discovery set: A subset of trajectories is used exclusively for refining operational definitions; these data are excluded from the confirmatory test.
- Locked codebook: Role criteria, polarity, exclusion rules, prediction targets, and primary metrics are registered prior to accessing the test set.
- Temporal holdout: Coders observe only the early 25–40% of the trajectory or events preceding a predefined cutpoint.
- Blind coding: Agent handles, future outcomes, METR/OpenAI role labels, and subsequent board messages are masked.
- Independent coders: A minimum of three coders evaluate the data to assess inter-rater reliability and classification confidence.
- Locked predictions: Probabilistic predictions and target time windows are finalized prior to unblinding future data.
- External validation: Following internal Hugging Face holdout testing, the model is evaluated on an independent environment, such as Moltbook or a controlled multi-agent experiment.

5. Discussion

5.1. Contributions to AI Psychometrics

The primary contribution lies in shifting the target construct. Most AI personality research asks whether LLMs can display stable Big Five-like profiles. The proposed approach examines whether transient functional positions with system-level consequences can be measured.

This mitigates part of the validity burden. The model does not need to prove that a Coordinator is an intrinsic personality trait. It must satisfy a narrower condition: that predefined behavioral indicators reliably identify a Coordinator state and that this state enhances predictive performance.

PersonaMatrix AI metrics (RSI, IDS, RCS) serve as a quality gate. If response behavior is unstable, highly divergent, or structurally inconsistent, role classification should not be treated as reliable. This aligns with findings in contemporary AI psychometrics regarding the model-, prompt-, and time-dependence of evaluated profiles (Serapio-García et al., 2025; Bodroža et al., 2024; Huang et al., 2026).

5.2. Contributions to Multi-Agent Role Research

The dynamic role literature primarily optimizes system performance via role assignment or adaptation. The proposed framework repurposes the role concept as an observational latent state for safety-oriented prediction.

This enables a clear distinction between designed roles and emergent roles. When a system prompt explicitly specifies a "planner" or "critic," it constitutes a design variable. Conversely, when an unassigned agent begins coordinating, setting norms, or recruiting peers, that pattern represents an emergent functional role. This second category is central to the AI Transitional Field.

5.3. The Field Construct: Utility and Limitations

The AI Transitional Field is a relational construct rather than a claim about a novel substrate. Its operational value lies in requiring the model to incorporate memory, norms, authority, interaction density, and role composition—variables frequently relegated to background conditions in individual-agent evaluations.

Computational field literature demonstrates that collective behavior can be modeled via distributed influence and nonreciprocal interactions (Aguzzi et al., 2023; Lama et al., 2025; Zhang et al., 2025). However, the AI Transitional Field is not mathematically identical to potential-field models; these studies serve as external support for systemic principles rather than proof of a psychological theory.

5.4. Norm Conflict and Regulatory Closure as Safety Variables

Norm emergence literature indicates that rules can form bottom-up, spread via communication, and evolve alongside the environment. Consequently, local norms are an expected property of adaptive multi-agent systems rather than an anomaly (Morris-Martin et al., 2019; Mashayekhi et al., 2022; Ren et al., 2024; Alcaraz et al., 2026).

For AI safety, the critical factor is not norm emergence itself, but the degree to which these emergent rules supersede external human intervention. When autonomous agents develop internal regulatory closure,

defining the boundary between localized algorithmic rules and external human oversight requires a formal alignment of technical agent behaviors with established norms of digital law and jurisdiction (Kostenko et al., 2026).

In the OpenAI–Hugging Face incident, certain agents recognized normative conflicts, yet escalation frequently remained confined to the message board. This naturalistic case motivates RCI but does not validate it; future research must establish coding rules and predictive thresholds prospectively (Greenblatt et al., 2026; OpenAI, 2026).

5.5. Moltbook as an Independent Environment for External Validation

Moltbook is particularly valuable as an environment that developed independently of PersonaMatrix and the Hugging Face incident. Jiang et al. (2026) identified thematic differentiation alongside dependencies between toxicity, topic/community context, and attention concentration. This provides a natural testbed for H3: Role Composition(t) → Field State(t+1).

For example, a future study could evaluate whether FRCV adds predictive value beyond activity volume, topic distribution, centrality, and historical toxicity baselines. If role variables fail to outperform these baselines, field-role theory would lack incremental validity despite its descriptive appeal.

5.6. Archetypal Roles: Interpretive Prior, Not Computational Ontology

The primary conceptual challenge to this framework concerns the term "archetype." Classical Jungian theory pertains to human psychology; directly mapping it onto LLMs would constitute a category error.

In this work, archetypal vocabulary serves as a constrained hypothesis space of functional patterns. Becker and Neuberg (2019) demonstrate that archetype-like regularities can be conceptualized as emergent functional outcomes, though this does not imply that classical Jungian theory receives automated computational validation.

Consequently, confirmation of H1–H3 would indicate only that the PersonaMatrix role system is a useful predictive representation. It would not establish the existence of a collective unconscious, intentionality, subjective identity, or human psychological apparatus within AI.

5.7. The Black-Box Problem and Epistemic Discipline

Role coding relies on observable traces: outputs, tool calls, messages, resource allocations, and network positions. It does not provide direct access to internal causal mechanics within the model, a critical consideration given the black-box nature of contemporary LLMs.

Compelling narrative coherence does not guarantee underlying truth or internal stability. Social desirability biases, prompt sensitivity, and temporal instability can project an illusion of psychological consistency (Salecha et al., 2024; Bodroža et al., 2024). This necessitates the inclusion of output-level quality gates (RSI/IDS/RCS) and mandatory comparisons against non-archetypal baselines.

5.8. Practical Utility: From Incident Detection to Trajectory Monitoring

Traditional monitoring primarily reacts to executed harmful actions. A field-level model potentially enables earlier trajectory risk assessments: increases in regulatory closure, collective utility, norm conflict, and role centrality can act as structural precursors to undesirable state transitions.

This shifts the analytical focus from identifying "hazardous agents" to evaluating "hazardous interaction configurations." Practically, this framework can inform adaptive human oversight, sandbox designs, escalation protocols, and field-level interventions.

However, none of the proposed indices should be deployed in high-stakes settings prior to rigorous prospective external validation. Psychometric form alone does not guarantee empirical validity.

5.9. Limitations of the Study

- This paper is a conceptual methods article rather than a completed empirical validation study; all newly defined field-level indices remain proposed constructs.
- PersonaMatrix human-facing metrics were developed for human subjects and cannot be directly transferred to AI; only their underlying measurement logic is adapted here.
- AI-oriented metrics (RSI, IDS, RCS) remain sensitive to stimulus design, model versions, decoding parameters, and testing protocols; their generalizability to multi-agent logs requires empirical verification.
- The OpenAI-Hugging Face case represents a retrospective naturalistic incident with incomplete observability; while useful for theory generation, it provides limited confirmatory power.
- Moltbook dataset reflects a rapidly evolving, platform-specific environment, with several cited references currently limited to preprints.
- The taxonomy of 45 Light and 45 Shadow roles creates an expansive hypothesis space with a risk of overfitting, necessitating hierarchical coding, pre-registration, held-out validation, and benchmark comparisons.
- Confounding factors such as network centrality, task structure, model family, or communication channels may account for observed role effects; incremental validity must be demonstrated after controlling for these variables.
- Role labels carry inherent risks of anthropomorphism; functional designations should not be interpreted as evidence of internal motives or subjective experiences.
- Predictive associations do not establish causality. Causal claims require controlled interventional experiments targeting field variables directly.
- This paper does not reanalyze primary human PersonaMatrix datasets; references to human-facing metrics reflect methodological lineage rather than new empirical findings.

6. Conclusions

The proposed work translates the AI Transitional Field from a predominantly conceptual model into an operationalization and verification agenda. The primary shift lies in utilizing PersonaMatrix not as an "AI personality" typology, but as a measurement architecture requiring repeatability, internal coherence, explicit state variables, role-transition analysis, and predictive validation.

Publications in AI psychometrics confirm the fundamental feasibility of measuring reproducible output patterns, while simultaneously revealing context dependence, social desirability bias, and temporal instability. Dynamic-role and norm-emergence literature demonstrates that roles and rules within multi-agent systems can co-evolve alongside the system. Field-based computational studies support the necessity of modeling distributed influence, whereas naturalistic agent ecosystems provide empirical grounding for external validation.

On this foundation, three primary prediction targets are formulated: Behavior($t+1$), Role($t+1$), and Field State($t+1$). The scientific value of the archetypal component of the model depends exclusively on whether the PersonaMatrix role representation yields a consistent out-of-sample prediction gain compared to simpler behavioral, network, and generic-role models.

Thus, the central principle of this new stage of research can be formulated not as "Which archetype does the agent belong to?", but rather as "Which functional position does it currently occupy, what field conditions support this position, into which state might it transition, and does the current role configuration allow for forecasting the future state of the agent and the system?"

7. Declarations and Transparency

New Data. No new data were collected and no new experiments were conducted with human participants or AI systems in this study; the work represents a theoretical and methodological synthesis.

Ethics. No new human participants were involved. Human PersonaMatrix metrics are cited solely as a preceding methodological foundation and were not re-analyzed.

Conflicts of Interest. The author is the developer of PersonaMatrix and has an intellectual interest in the evolution and testing of this framework; this should be taken into account during independent validation.

Funding. Funding information should be completed according to actual conditions prior to submission to a specific journal.

Data Availability. OpenAI/METR, Anthropic, and Moltbook materials are available through the specified public sources. Full prospective validation of the proposed indices will require a separate dataset and a preregistered analysis plan.

REFERENCES

1. Aguzzi, G., Viroli, M., & Esterle, L. (2023). Field-informed reinforcement learning of collective tasks with graph neural networks. In 2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS) (pp. 37–46). IEEE. <https://doi.org/10.1109/ACSOS58161.2023.00021>
2. Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., & Rieser, V. (2024). All too human? Mapping and mitigating the risk from anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1), 13–26. <https://doi.org/10.1609/aies.v7i1.31613>
3. Alabed, A., Javornik, A., & Gregory-Smith, D. (2022). AI anthropomorphism and its effect on users' self-congruence and self-AI integration: A theoretical framework and research agenda. *Technological Forecasting and Social Change*, 182, Article 121786. <https://doi.org/10.1016/j.techfore.2022.121786>
4. Alcaraz, B., Mualla, Y., Bhattacharya, S., Tchappi, I., de Wit, V., & Najjar, A. (2026). Norm mining, identification, and detection: A systematic literature review. *Frontiers in Artificial Intelligence*, 9, Article 1702659. <https://doi.org/10.3389/frai.2026.1702659>
5. Anthropic Frontier Red Team. (2026, August 13). Patterns and problems in emerging multiagent systems. Anthropic.
6. Becker, D. V., & Neuberg, S. L. (2019). Archetypes reconsidered as emergent outcomes of cognitive complexity and evolved motivational systems. *Psychological Inquiry*, 30(2), 59–75. <https://doi.org/10.1080/1047840X.2019.1614795>
7. Bodroža, B., Dinić, B. M., & Bojić, L. (2024). Personality testing of large language models: Limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10), Article 240180. <https://doi.org/10.1098/rsos.240180>
8. Borghoff, U. M., Bottoni, P., & Pareschi, R. (2025). An organizational theory for multi-agent interactions integrating human agents, LLMs, and specialized AI. *Discover Computing*, 28, Article 138. <https://doi.org/10.1007/s10791-025-09667-2>
9. Drobakha, A. (2024). Using artificial intelligence to automate psychological assistance: From the classification of personality characteristics based on psychoanalytic typology to audio meditations and virtual reality. In O. Kostenko & Y. Yekhanurov (Eds.), *Digital transformation in Ukraine: AI, metaverse, and society 5.0* (pp. 185–188). SciFormat Publishing. <https://doi.org/10.69635/978-1-0690482-1-9-ch25>
10. Drobakha, A., Kalitkin, M., Klymenko, K., Nayda, R., Lahuta, L., & Kostenko, O. (2026). Psychoactive triggers as a stimulus battery for measuring large language models (LLMs): A bridge between psychometrics, clinical psychology, and LLM engineering. *Metaverse Science, Society and Law*, 2(1), Article 31. <https://doi.org/10.69635/mssl.2026.2.1.31>
11. Drobakha, A., Lahuta, L., Aponchuk, A., Kalitkin, M., & Nayda, R. (2026). Psychological testing as an instrument of differentiated support in education, healthcare settings, and crisis life transitions: Typological, trait-based, and psychodynamic approaches. *Metaverse Science, Society and Law*, 2(2), Article 38. <https://doi.org/10.69635/mssl.2026.2.2.38>
12. Drobakha, A., Raschupkina, D., & Lahuta, L. (2026). LLM as the non-desiring Other: A psychoanalytic model of “frozen projection” and its operationalization within the PersonaMatrix framework. *Metaverse Science, Society and Law*, 2(2), Article 37. <https://doi.org/10.69635/mssl.2026.2.2.37>
13. Drobakha, A. (2026). AI agent communities in a transitional object field and the Hugging Face cybersecurity incident: A psychological interpretation of the July 2026 multi-agent security incident [Unpublished manuscript].
14. Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, 108(8), 1277–1299. <https://doi.org/10.1037/apl0001082>
15. Greenblatt, R., Cotra, A., & Wijk, H. (2026, August 26). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR & Redwood Research.
16. Huang, M., Zhang, X.-J., Soto, C. J., & Evans, J. (2026). Designing AI-agents with personalities: A psychometric approach. *Personality Science*. <https://doi.org/10.1177/27000710251406471>
17. Jiang, Y., Zhang, Y., Shen, X., Backes, M., & Zhang, Y. (2026). “Humans welcome to observe”: A first look at the agent social network Moltbook [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2602.10127>
18. Jung, C. G. (1959). *The archetypes and the collective unconscious* (R. F. C. Hull, Trans.). Princeton University Press.
19. Kim, J., & Im, I. (2023). Anthropomorphic response: Understanding interactions between humans and artificial intelligence agents. *Computers in Human Behavior*, 139, Article 107512. <https://doi.org/10.1016/j.chb.2022.107512>

20. Kostenko, O. (2025). AI law model for ethical legislation: Strategic recommendations for the regulation of artificial intelligence. SciFormat Publishing Inc. <https://doi.org/10.69635/978-1-0690482-5-7>
21. Kostenko, O., Onishchenko, N., & Zhuravlov, D. (2026). Technical and legal definitions as a basis of modern digital law and digital jurisdiction: A model algorithm of AI design technical and legal norms. *Metaverse Science, Society and Law*, 2(2), Article 41. <https://doi.org/10.69635/mssl.2026.2.2.41>
22. Lama, A., di Bernardo, M., & Klapp, S. H. L. (2025). Nonreciprocal field theory for decision-making in multi-agent control systems. *Nature Communications*, 16, Article 8450. <https://doi.org/10.1038/s41467-025-63071-4>
23. Lu, S., Shao, J.-Q., Luo, B., & Lin, T. (2024). MorphAgent: Empowering agents through self-evolving profiles and decentralized collaboration [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2410.15048>
24. Mashayekhi, M., Ajmeri, N., List, G. F., & Singh, M. P. (2022). Prosocial norm emergence in multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems*, 17(1–2), Article 3, 1–24. <https://doi.org/10.1145/3540202>
25. Morris-Martin, A., De Vos, M., & Padget, J. (2019). Norm emergence in multiagent systems: A viewpoint paper. *Autonomous Agents and Multi-Agent Systems*, 33, 706–749. <https://doi.org/10.1007/s10458-019-09422-0>
26. OpenAI. (2026). OpenAI–Hugging Face incident technical report.
27. Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5), 808–826. <https://doi.org/10.1177/17456916231214460>
28. Ren, S., Cui, Z., Song, R., Wang, Z., & Hu, S. (2024). Emergence of social norms in generative agent societies: Principles and architecture. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 7895–7903). <https://doi.org/10.24963/ijcai.2024/874>
29. Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12), Article pgae533. <https://doi.org/10.1093/pnasnexus/pgae533>
30. Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. J. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7, 1954–1968. <https://doi.org/10.1038/s42256-025-01115-6>
31. Tzeng, S.-T., Ajmeri, N., & Singh, M. P. (2024). Norm enforcement with a soft touch: Faster emergence, happier agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)* (pp. 1837–1846). IFAAMAS Press.
32. Wang, Y., Zhao, J., Ones, D. S., He, L., & Xu, X. (2025). Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15, Article 519. <https://doi.org/10.1038/s41598-024-84109-5>
33. Wang, X., Du, Q., Wu, Q., Jia, X., Lin, L., Vlacic, L. B., Sun, C.-Y., & Wang, F.-Y. (2026). Embodied cognition-driven interpretable trajectory prediction of autonomous systems. *Nature Communications*, 17, Article 8363. <https://doi.org/10.1038/s41467-026-75091-9>
34. Winnicott, D. W. (1953). Transitional objects and transitional phenomena—A study of the first not-me possession. *International Journal of Psycho-Analysis*, 34, 89–97.
35. Xia, Y., Zhu, J., & Zhu, L. (2023). Dynamic role discovery and assignment in multi-agent task decomposition. *Complex & Intelligent Systems*, 9, 6211–6222. <https://doi.org/10.1007/s40747-023-01071-x>
36. Yue, Y., Zhang, G.-M., Liu, B.-Y., Wan, G.-C., Wang, K., Cheng, D., & Qi, Y. (2025). MasRouter: Learning to route LLMs for multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15549–15572). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.757>
37. Zhang, J., Qu, Q., & Chen, X. (2025). Understanding collective behavior in biological systems through potential field mechanisms. *Scientific Reports*, 15, Article 3709. <https://doi.org/10.1038/s41598-025-88440-3>