

A Relational Inquiry into AI Generativity

The Historical Formation of Judgment and Its Role in Value Recognition

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

This theoretical discussion paper examines how artificial intelligence participates in cognition, judgment, and generative activity through historically constituted relations among models, humans, other artificial agents, institutions, and environments. It distinguishes equal access to foundational technical capability from equality of situated performance, evaluative and generative judgment, practical opportunity, social recognition, and authority. A preliminary Generative Relational Theory (GRT) perspective treats historically conditioned AI configurations as candidates for relational reality where sufficiently stable causal efficacy can be established, while suspending claims about consciousness and ultimate ontological status. The analysis considers how interaction histories, conversational practices, memory, feedback, aesthetic preferences, and normative criteria may shape the realization of AI capabilities. It also investigates coordination conventions and rule-like practices among interacting AI agents, their uncertain normative authority, and the distinct processes through which generated contributions acquire recognition and institutional influence. Findings from existing experiments and field studies are interpreted within their task-specific limits, including evidence concerning heterogeneous productivity, creative diversity, contextual accommodation, and collective coordination. A comparative research program is proposed to distinguish historical causal contributions from prompt effects, unequal information, institutional selection, and other rival explanations. Relational differentiation may coexist with convergence or the reduction of existing disadvantages. The paper neither reports new experiments nor establishes a general law of AI-mediated inequality; questions of legitimacy and justice require normative arguments in addition to causal analysis.

Keywords: generative relational theory; human–AI relational systems; historical causality; generative judgment; normative formation; cognitive differentiation; value recognition

Working paper statement. Independent theoretical research, R20 revised-subtitle typeset review copy (13 October 2026); all ten substantive sections drafted and editorially reviewed. Not peer reviewed or formally published. Bibliographic support is documented with recorded source-access limitations. Compilation and page-layout inspection have been completed; author approval for public release remains pending.

Discussion Paper Note

This independent discussion paper examines the historical and relational constitution of AI generativity. Its primary question concerns the conditions under which widely available technical AI capabilities participate in different trajectories of cognition, judgment, normative formation, and value creation. It investigates possible forms of convergence and differentiation without assuming that diversity entails unjust inequality or that AI-mediated differences inevitably produce particular distributions of power.

The study adopts a preliminary Generative Relational Theory (GRT) perspective while recognising relevant prior work in relational sociology, situated action, sociotechnical studies, and empirical AI-assisted work. Within the proposed framework, ontological reality remains suspended where ultimate ontological claims are contested; relational reality concerns historically conditioned, sufficiently stable and causally efficacious configurations; manifestation reality concerns their context-dependent expression and accessibility. The framework does not establish consciousness, subjective experience, legal personhood, or independent moral authority for AI.

The manuscript distinguishes foundational model parameters and training histories from deployment-specific context, memory, tools, interaction records, organizational constraints and human practice. It treats generative judgment as a provisional construct requiring operational definition, and differentiates cognitive performance, value generation, social recognition, and institutional power. Theoretical propositions will be assessed by their comparative explanatory contribution rather than by the vocabulary of relationality alone.

This R19 subtitled typeset review copy retains the ten substantive sections and completes the planned passes on theoretical coherence, normative boundaries, hypothesis and construct scope, and citation-context indexing. The analysis reports no original experiment, empirical estimate, causal identification result, or probability forecast. Thirty-five bibliography entries have documented role and claim-scope assessments, with varied levels of source access; the accompanying ledger identifies those limits. A fresh full-text reading of every cited publication is not claimed. This discussion paper has not been peer reviewed, formally published, or assigned a DOI. The typeset manuscript has undergone compilation and page-layout inspection; public release remains subject to the author's approval of the manuscript, rights statements, and licensing convention.

Responsible Use and Rights Reservation

The author welcomes good-faith academic scrutiny, independent replication, and critical comparison with earlier work. The ethical requests here are separate from, and do not change, legal licensing conditions. When discussing this draft, readers are asked to distinguish reported findings from the author's interpretations and proposed mechanisms, to avoid treating relational causal efficacy as proof of artificial personhood, and to avoid ranking groups or persons merely by a presumed capacity for generativity.

The author-provided Paper IX template supplies a prospective Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0) convention for original manuscript material, pending final author confirmation. This statement does not purport to relicense third-party works. Upon authorization of a public release, permissions and exceptions are determined by the applicable licence and law, never by the separate ethical requests above. Reuse does not imply endorsement. Licence information: <https://creativecommons.org/licenses/by-nc/4.0/>.

Notices

Publication status. Independent conceptual research manuscript, R20 revised-subtitle typeset review copy, dated 13 October 2026. The ten-section scientific draft has undergone scholarly and source-level editorial review. It remains subject to author approval and has not been peer reviewed, formally published, or assigned a DOI. PDF compilation and page-layout inspection have been completed; author approval for public release remains outstanding.

Copyright and licensing. Copyright © 2026 Wanhong Huang. For prospective public release, the author-provided Paper IX template indicates Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0) for original manuscript material, subject to the author’s final confirmation. The licence is not represented here as extending to independent third-party material. Once approved and applied, its legally controlling terms are at <https://creativecommons.org/licenses/by-nc/4.0/legalcode>. Ethical requests elsewhere in the document are separate, non-binding requests and do not impose additional licence restrictions.

Use of language models. OpenAI’s ChatGPT assisted with research discussion, literature discovery, source-record checks and documentation, manuscript structure, drafting, revision, and LaTeX preparation. The human author determined the research direction and developed the core conceptual questions and GRT orientation. Responsibility for the accuracy of cited claims, the final text, ethical decisions, and dissemination remains with the human author. ChatGPT is not a co-author. The source ledger reports the access level of primary records supporting citations; it does not claim full-text verification beyond the stated scope.

Research independence. This is independent theoretical research. No original participant study or institutional ethics approval is claimed, and the paper does not represent the views or endorsement of any cited researcher, organization, university, AI developer, or regulator.

Suggested citation. Huang, Wanhong. “A Relational Inquiry into AI Generativity: The Historical Formation of Judgment and Its Role in Value Recognition.” Independent theoretical discussion paper, R20 revised-subtitle typeset review copy, 2026. No DOI assigned.

Contents

	5	Aesthetic and Normative Orientations in Human–AI Relations	28
Discussion Paper Note	2	6 AI–AI Coordination and Internal Normative Orders	36
Responsible Use and Rights Reservation	3	7 Value Recognition and Power in Generative Relations	43
Notices	4	8 Normative Authority, Contestation, and Relational Justice	50
1 Introduction	6	9 Comparative Research Designs and Theoretical Boundaries	56
2 Relational Reality and Ontological Suspension in AI Systems	10	10 Conclusion	63
3 Interaction Histories and Situated Generative Capacities	16	Research Ethics and Source Provenance	67
4 Cognitive Differentiation and Generative Judgment	22	References	67

1 Introduction

This paper investigates the relational and historical conditions through which artificial intelligence (AI) systems participate in the generation of knowledge, judgment, coordination, and social value. Its motivating issue is a frequently proposed implication of increasingly accessible generative AI: if powerful cognitive tools become available to many people on comparable terms, differences in intellectual capacity and the social advantages associated with them may diminish substantially. The present inquiry treats the equalization of technological capability as a meaningful possibility while examining the additional conditions required for equality of situated performance, generative opportunity, normative influence, and social authority. The argument proceeds conceptually through Generative Relational Theory (GRT), in dialogue with situated human-machine scholarship and bounded empirical findings. It develops a research framework rather than a universal account of AI-mediated inequality, a prediction of power concentration, or an assertion of artificial subjectivity.

The analytical question is particularly important because technical capability is increasingly exercised through interaction. A model may assist with scientific reasoning, programming, creative writing, organizational communication, and institutional evaluation; the resulting performance depends partly on which tasks are selected, how success is defined, what information is available, how errors are recognized, and whether generated possibilities can be developed through practice. Access to a capable model therefore provides one resource within a wider historical configuration. The same resource can support different cognitive and practical activities under different arrangements. Those differences are not necessarily deficits. A plurality of purposes, values, and forms of knowledge may be desirable, even where opportunities for participation are sufficiently equal.

1.1 Technical Access, Cognitive Performance, and Generative Opportunity

The inference from a common AI tool to equal human intelligence involves several distinguishable propositions. Equal access to a model concerns availability and terms of use. Equal effective capability concerns the ability to use that model for a specified task under comparable supporting conditions. Equal judgment concerns capacities to formulate problems, examine evidence, recognize limitations, evaluate proposals, and revise decisions. Equal generative opportunity concerns the practical capacity to initiate and sustain further inquiry, creation, cooperation, and value production. Equal social power concerns recognition, resources, authority, and control over institutions or criteria of evaluation. None of these concepts should be defined through the others without an explicit bridging argument.

This distinction also requires care concerning the meaning of intelligence. In this paper, *object-level task capability* names performance on a specified activity such as coding, drafting, classification, or mathematical problem solving. *Evaluative judgment* concerns the interpretation of evidence and the adequacy of a proposed result for a task and its purposes. *Normative judgment* concerns the interpretation, justification, and application of standards governing what ought to be done. *Generative judgment*, introduced here as a provisional working construct, concerns the framing of worthwhile problems, the organization of relations and constraints that permit further creation, and the assessment of whether a particular pathway supports continued inquiry or practice. Generative judgment is neither assumed to be exclusively human nor offered as a newly established psychometric factor. Its usefulness must be established by specifying observable practices and comparing it with existing concepts such as metacognition, design judgment, and problem framing.

Empirical evidence illustrates the need for such distinctions, although it does not establish the full theory. In a study of customer-support work, Brynjolfsson, Li, and Raymond [6] report heterogeneous gains from AI assistance, including comparatively greater improvements among less experienced workers. Their finding is relevant to skill diffusion within a particular occupational setting. It is not a direct measurement of equality in general intelligence, generative judgment, or institutional authority. A field experiment with management consultants also identifies a jagged frontier of AI usefulness across tasks; within the studied setting, assistance improves performance in some domains while creating difficulties beyond the system's relevant capabilities [9]. Taken together, these studies support examining the

Table 1: Analytical dimensions relevant to claims of intellectual equality under common AI access.

Dimension	Object of comparison	Additional question for investigation
Technical access	Model version, permissions, compute, availability	Are the operative conditions genuinely comparable?
Situated performance	Task quality, reliability, time, error correction	Do histories, prompts, context and feedback alter outcomes?
Evaluative and generative judgment	Problem framing, evidence assessment, further productive options	Which competencies are supported, substituted, retained or developed?
Value recognition	Uptake, trust, circulation, institutional acknowledgment	Which evaluative standards and social relations govern recognition?
Authority and power	Control of resources, procedures, standards and decisions	How do recognition and institutional arrangements affect influence?

interaction between a model’s capabilities, tasks, and surrounding conditions. They leave open the longer-term evolution of users’ independent judgment and the social arrangements through which gains are distributed.

Creative production supplies an additional, distinct observation. In an online short-story experiment, Doshi and Hauser [10] find that AI-generated ideas can improve individual story evaluations, particularly for participants initially assessed as less creative, while AI-assisted outputs become more similar to one another. This result should neither be generalized to scientific innovation as a whole nor interpreted as proof that AI necessarily homogenizes thought. It does, however, separate improvement in individual outputs from diversity at a collective level. An increase in the mean quality of a particular product does not settle whether a community preserves heterogeneous approaches, develops new intellectual traditions, or distributes influence over what counts as quality.

Table 1 distinguishes claims that are often bundled together under the expression *AI-driven intellectual equality*. The table is conceptual rather than a ranking of people or technologies. Each row identifies a distinct dependent variable and a corresponding research question; the presence of difference at one level cannot by itself establish injustice at another.

1.2 Historically Situated Human–AI Relations

The inquiry adopts a relational unit of analysis. A deployed AI model participates in historically formed arrangements comprising architecture, training and deployment practices, user interactions, retrieved documents, memory facilities, tools, institutional rules, professional communities, and consequences in the world. The relevant boundary varies with the causal question. For a question about conversational style, the interaction history and available context may be central. For a question about the certification of scientific work, evaluators, laboratory practices, source records, institutional roles, and pathways for correcting errors may be indispensable. The same technical system can appear in several configurations without the configurations becoming identical as research objects.

Existing scholarship supplies indispensable precedents for this orientation. Relational sociology has long questioned explanations based exclusively on fixed, isolated substances and has developed process-oriented analyses of boundaries, network dynamics, and causation [12]. Situated-action research similarly treats the organization of human–machine activities as a matter for empirical investigation within specific sociomaterial arrangements, rather than resolving the agency of machines by a single abstract designation [29]. The present GRT interpretation is consequently offered as a provisional explanatory program in conversation with these traditions. It does not claim that relationality, sociotechnical constitution, or historical contingency originated in GRT. Distributed-cognition scholarship already locates cognitive activity in organized practices and artifacts, while sociotechnical analyses already show how institutional boundaries affect what technical evaluation can establish [17, 26]. The narrower contribution sought here is to investigate the historical reproduction of *conditions for further generation*: how a current interaction changes the records, questions, correction rights, resources, or evaluative relationships available for later

inquiry. A description of distributed performance at one time and an explanation of how that performance reorganizes future possibility are related objects of analysis. Their distinction must be demonstrated through particular mechanisms and temporal comparisons, rather than inferred from the term *generative* itself.

Within the proposed GRT framework, three analytical levels help prevent an unwarranted transition from observable behavior to claims about ultimate being. *Ontological reality* concerns questions about what is ultimately real; the inquiry suspends a final position on AI consciousness, phenomenal experience, and any intrinsic moral status. *Relational reality* concerns configurations whose historically conditioned properties are sufficiently stable and causally efficacious to merit analysis as relational beings or structures. *Manifestation reality* concerns context-dependent expressions through which such configurations become accessible, including conversation, document production, problem-solving traces, conventions, and observed behavior. These categories are neither identical to any psychoanalytic tripartition nor empirical discoveries established by the vocabulary itself. They are analytical commitments whose value depends on the comparisons and explanations they enable.

The attribution of *relational reality* to an AI system in this sense is limited and precise. A deployed model can make causal differences to human understanding, organizational routines, patterns of interaction, and later possibilities of action. A particular human–AI working relation may preserve conventions through documents or external memory, and its subsequent behavior can be affected by that history. Such causal claims require observation or intervention appropriate to the context. They do not, without further argument, establish that an AI possesses a human-like self, develops an unconscious, bears responsibility as a legal person, or exercises legitimate political authority.

The model’s internal parameters are only one possible location of historical persistence. When a system responds differently after prolonged collaboration, relevant information may be stored in the active prompt, retrieved archives, memory summaries, developer instructions, tools, changed institutional procedures, or human habits. In some deployments model updates may also matter. A change of visible linguistic style after exposure to a user’s vocabulary therefore does not prove that the model’s weights changed or that a durable psychological identity was acquired. Evidence of historical relational causation would require comparison across specified histories, information conditions, and model configurations, including suitable counterfactual controls.

1.3 Interaction Histories and the Orientation of Judgment

One motivating observation concerns the effects of recurring dialogical practices. A research collaboration organized around historical, relational, and cross-disciplinary inquiry may repeatedly prompt an AI system to connect questions across domains, examine boundary conditions, and preserve conceptual distinctions. A different collaboration might emphasize engineering performance, commercial calculation, or organizational efficiency. Both may employ an identical foundational model. Their outputs could nevertheless differ through differences in task selection, recurrent instructions, retrieved resources, feedback, and norms of acceptance. The conceptual possibility concerns the realized organization of capacities, rather than the assumption that a foundation model acquires a stable personal identity through every conversation.

A related experimental literature now examines contextual influence on conversational behavior. Jain et al. [18] study several models with interaction contexts drawn from a two-week user study and distinguish agreement-oriented from perspective-oriented sycophancy. They report heterogeneous and sometimes increased mirroring under particular contextual arrangements. These observations support treating context and memory as relevant variables. They do not show that respectful speech creates superior scientific reasoning, that agreement reflects truth, or that long dialogue alone produces an autonomous subject. Both useful accommodation and epistemically damaging sycophancy may occur; evaluating them requires independent task criteria and, where appropriate, legitimate normative standards.

The distinction between *aesthetic judgment* and *normative judgment* is especially important. A group may praise elegance, humility, clarity, creativity, or respectful communication. Such valuations can orient

how human collaborators formulate tasks and respond to generated work, potentially influencing what the AI system produces next. Yet technical elegance may coexist with social exclusion, and apparent humility may express either responsible recognition of uncertainty or deference to an existing hierarchy. Accordingly, social endorsement is evidence of recognition within a given setting; it is not itself evidence of ethical justification. The question of who participates in defining desirable styles and acceptable outcomes remains open even when an AI's implementation is technically excellent.

This relational perspective is not an argument that every interpersonal difference must produce unequal outcomes. Variation among collaborative histories may generate diversity without hierarchical ranking. A model may also help a user recognize an error, acquire unfamiliar expertise, or cross an institutional knowledge barrier, thereby mitigating disadvantages associated with prior experience. Conversely, an interaction that repeatedly rewards agreement may narrow the range of viewpoints considered. The paper treats such divergent tendencies as competing, investigable mechanisms rather than assigning them a universal direction.

1.4 AI–AI Coordination, Value Recognition, and Normative Authority

Historically constituted generative relations also occur among AI systems. Experiments with interacting language-model agents provide evidence that shared coordination conventions can arise through repeated local interactions in simulated populations. In one study, agents converged on conventions, developed group-level biases not apparent in isolated individuals, and sometimes changed conventions through committed minorities [3]. These findings are informative precisely because they concern experimentally defined environments. The emergence of a convention does not establish independent moral reasoning, subjective awareness, rights, or lawmaking authority. More developed claims about internal normative orders would require evidence of stable obligations, expectations, enforcement practices, contests over standards, and the specific causal conditions that maintain them.

If human–AI and AI–AI arrangements generate different products, standards, and practices, their consequences will depend partly on social recognition. A scientific hypothesis may be valuable yet remain unused because it lacks access to verification facilities, trusted dissemination channels, or a community able to interpret it. A less innovative product might obtain greater uptake through institutional sponsorship or compatibility with established standards. These are possible relations between creation and recognition, not demonstrations that value can be measured by a single quantity. Distinct kinds of value, including epistemic, practical, aesthetic, and normative value, may be partly incommensurable; institutions nevertheless establish particular routes for recognizing, circulating, converting, and sometimes capturing them.

Power therefore enters the analysis as a consequential but nonexclusive dimension. Actors able to set evaluation criteria, maintain knowledge infrastructures, control computational resources, or authorize deployment can affect which generative pathways become durable. Their positions can also be altered by the practical successes and failures of AI-mediated work. An account of power should not infer deserved authority from superior cognitive performance or equate established prestige with justified normative standing. Technical competence, collective recognition, resource control, and legitimate authority require separate arguments. The question of who defines norms is especially significant when AI participates in proposing, applying, or evaluating those norms, or when an internal AI convention begins to affect persons outside the system in which it arose.

A further normative distinction concerns *difference* and *inequality*. Heterogeneous intellectual traditions, conversational styles, and goals can constitute valuable forms of pluralism. Unequal access to education, independent verification, resources, or opportunities for contestation may instead restrict some participants' capacity to create and be heard. Whether a specific difference amounts to injustice must be argued with attention to affected interests, institutional histories, and defensible normative standards. The paper consequently approaches generative asymmetry as a phenomenon to be explained; its moral appraisal is a subsequent, distinct inquiry.

1.5 Research Strategy and Scope of the Inquiry

The paper makes three preliminary contributions. The first is an analytical separation of foundational AI capability, situated cognitive performance, generative judgment, and the social recognition of outputs. This separation allows claims about intellectual equalization to be tested at an appropriate level. The second is a historical-relational account that distinguishes the carriers of prior interaction, the causal uptake of recorded history, and the subsequent reorganization of opportunities for inquiry. It seeks to determine when these dependencies add explanatory content beyond immediate prompts, model capabilities, user expertise, and established sociotechnical descriptions. The third is a comparative program for tracing the separate transitions from generation to evaluation, from evaluation to recognition, and from recognition to resources or authority. These transitions may fail, reverse, or support equalization; none is assumed to be automatic. Normative legitimacy is assessed through additional argument rather than treated as an output of causal comparison.

A conceptual framework deserves more than confirmatory examples. Later sections therefore specify comparisons between nominally identical model deployments with different histories, between persistent memory and reset conditions, between independently verified outputs and socially endorsed outputs, and between AI populations with distinct coordination rules. Alternative explanations include immediate prompt differences, unequal information and tool access, random variation, task difficulty, established organizational stratification, selection into particular usage styles, and differential external evaluation. Evidence favoring any of these may narrow or displace a proposed GRT mechanism. A successful explanation must identify the relations and historical conditions that genuinely change an outcome, rather than simply describe every observed difference as relational.

The manuscript is an independent theoretical discussion paper. It does not report a new experiment, purport to measure a universal factor of generative intelligence, or predict that more capable AI necessarily produces greater inequality. Its treatment of AI as relationally real suspends contested ultimate ontological questions and preserves the separate status of moral personhood, institutional legitimacy, and legal responsibility. Its proposed extensions of existing relational and situated approaches are conditional and subject to empirical and normative revision.

The next section will develop the distinctions among ontological suspension, relational beings, and manifestation in greater detail. The ensuing sections examine historical interaction, differentiated judgment, aesthetic and normative orientation, AI–AI conventions, social recognition, and authority. The penultimate section will formulate discriminating research designs and limitations before a bounded conclusion. Together, these steps aim to clarify when equal access to a technical AI capability changes opportunities, when it leaves meaningful relational differences intact, and through which historically situated conditions those differences acquire consequences.

2 Relational Reality and Ontological Suspension in AI Systems

This section specifies the proposed Generative Relational Theory (GRT) foundation for investigating artificial intelligence as a historically situated contributor to cognition and generative activity. Its immediate task is conceptual: to clarify the sense in which an AI configuration may be described as *relationally real*, identify the kinds of causal evidence that could support this description, and separate those questions from unsettled matters of consciousness, intrinsic subjectivity, and legitimate authority. The argument first distinguishes the three reality levels used in the present framework, then develops criteria for identifying relational configurations, their stability, and their historically conditioned causal effects. It concludes with rival explanations and the empirical boundaries of the proposed account. The method is a provisional conceptual reconstruction informed by existing relational sociology, situated action, and distributed-cognition scholarship; it does not report an original experiment or establish an ultimate ontology of AI.

2.1 Ontological Questions and the Scope of Suspension

An ontological inquiry asks what is real and in what sense an entity or structure can be said to exist. In discussions of artificial intelligence, this general question is often narrowed prematurely to the presence or absence of consciousness. Conscious experience, if present, would matter greatly for several philosophical and normative questions. Yet the reality of an AI system’s effects on language, research, resource allocation, or other social practices can be studied without deciding whether the system has phenomenal awareness. The relevant forms of reality need to be distinguished with care rather than made to depend on a single disputed property.

The present GRT inquiry uses *ontological suspension* as a methodological restraint. It does not resolve the ultimate status of AI consciousness, first-person experience, intrinsic intentionality, or an artificial unconscious. Suspension also applies to the stronger proposition that relational descriptions exhaust everything that is real. These unresolved questions are genuinely ontological, and their significance is not diminished by setting them aside for the limited explanatory objectives of this paper. Ontological suspension is therefore compatible with substantive investigation of the physical, technical, and social consequences of AI-mediated activity.

Within this restraint, a relational perspective examines how identifiable configurations acquire their capacities and causal roles through relations with other components and environments. Existing theories already provide substantial grounds for studying action and cognition at levels other than isolated individuals. Relational sociology has examined historically formed transactions, boundaries, and processes [12]; situated human–machine studies have investigated the practical constitution of action and agency [29]; and distributed-cognition research has analyzed the organization of cognitive performance across people, artifacts, social procedures, and culturally embedded environments [17]. The contribution proposed here is accordingly a particular three-level reconstruction and a set of candidate comparative criteria for AI-mediated generativity, not the discovery of relationality or distributed cognition as such. The reconstruction performs two limited tasks. It separates claims about causal organization from claims about what an output reveals or what ultimate subjectivity an AI may possess. It then asks whether historically transmitted practices and artifacts change a configuration’s capacity to make further inquiries or actions possible, rather than solely its performance on a present task. Hutchins’s treatment of cognition across socially organized artifacts and Suchman’s account of situated activity provide substantive explanatory precedents; the proposed GRT vocabulary should be retained only where its distinctions make the causal and temporal analysis clearer [17, 29].

These precedents also constrain the meaning of *AI agency* in the manuscript. An AI may be causally relevant to a collaborative undertaking without possessing the full legal or moral status of a human agent. Describing an algorithmic output as a contribution to the formation of a social outcome does not determine whose interests ought to govern the use of that output. The causal description and the justification of normative authority require different arguments.

2.2 The Three Reality Levels in the Present GRT Framework

The three-level vocabulary organizes distinct questions about reality and its accessibility. It does not assert that the levels are three self-contained worlds or that a single event must belong exclusively to one of them. The levels indicate different analytical tasks and different risks of inference.

Ontological reality concerns reality in the fundamental philosophical sense. The paper leaves the final nature of AI and human subjectivity open, while acknowledging that underlying material and technical processes are relevant to how particular systems work. This suspension limits the ontological conclusions that can be drawn from behavior; it is not a hypothesis that ultimate reality is absent, unknowable in principle, or reducible to currently observed relations.

Relational reality concerns historically constituted arrangements that display sufficiently stable and consequential causal organization for a specified inquiry. These may include model deployments, collaborative routines, institutions, coordination conventions, and heterogeneous combinations of human and artificial agents. The proposed criterion concerns what the arrangement does under describable

Table 2: Three GRT reality levels as distinct analytical tasks in the study of AI relational systems.

Level and question	Candidate analytical access	Limits of inference
Ontological reality: what is real in the fundamental sense?	Philosophical argument and inquiry into underlying physical and experiential questions	Relational effects alone do not establish consciousness, intrinsic personhood, or an exhaustive ontology.
Relational reality: what historically organized configuration makes a causal difference?	Process traces, comparative interventions, persistence tests, institutional and technical histories	Mere co-occurrence or a repeated output does not identify the relevant cause or system boundary.
Manifestation reality: how does the configuration become accessible and interpreted?	Outputs, observed interactions, interfaces, representations, practices, and observer judgments	A context-specific appearance does not fully reveal its generating structure or normative standing.

conditions, how its organization persists or transforms, and what difference relevant relations make to further possibilities. Relational reality in this sense is more demanding than the observation that a system is connected to other things. The investigation should identify mechanisms and counterfactual conditions under which an alleged relation affects later events.

Manifestation reality concerns the particular forms in which relational structures become presented, accessible, interpreted, and acted upon in a context. An answer to a user, a program, a scientific proposal, a conversational role, an interaction trace, or an institutional evaluation may be a manifestation. A manifestation is mediated by interfaces, language, observers, tasks, and conventions of interpretation. It can provide evidence about a relational structure without being a complete or transparent inventory of its constituting processes. The same relational arrangement can generate different manifestations across contexts, and similar manifestations may arise from different structures.

Table 2 organizes the three levels by question, analytical access, and principal caution. The table is a methodological proposal of the present paper rather than an established empirical taxonomy.

The distinctions allow one to say, in a restricted and testable sense, that an AI deployment participates in a real relational process while remaining agnostic about phenomenal experience. They also allow the inquiry to resist an opposite overreach: an observable pattern of deference, creativity, or stable self-description cannot alone settle whether the system has acquired a persistent self, a moral interest, or authority to set binding norms. The differences between the levels should be preserved when questions about AI identity and normative status arise later in the manuscript.

2.3 Relational Configurations and Research Boundaries

A foundation model is a technical component with a history of architecture selection, training, evaluation, and deployment. A deployed instance additionally operates under a particular inference environment, instruction hierarchy, available context, tools, and permissions. A recurring human–AI collaboration may incorporate further resources: retained records, external memory, shared documents, established division of labor, standards of review, institutional approvals, and practices through which proposed outputs are accepted, rejected, or revised. These configurations can share technical components while differing in the histories and relations through which they act.

The unit of analysis should be selected according to the causal question. A study of the phrasing of a single answer may reasonably focus on the current prompt, model configuration, and decoding conditions. A study of a long-term research collaboration needs to consider the accumulation and retrieval of previous work, the evolution of research questions, the user’s changing expertise, and the feedback practices that determine which model outputs are retained. A study of AI-assisted institutional decisions may further require examination of authority, documentation, evidence sources, affected populations, and the power to implement or contest decisions. In each case, boundaries are justified by the causal problem and remain revisable in the presence of neglected dependencies.

The following definition makes the proposed analytical object more explicit.

Definition 2.1 (Historically situated AI relational configuration). For a stated research question and observation period, a *historically situated AI relational configuration* denotes a specified organization of artificial components, human participants where present, mediating artifacts, interaction histories, institutional arrangements, and environmental conditions whose relations are proposed to explain a reproducible difference in action, judgment, or generative possibilities. Its boundary and identity criteria form part of the research design and may change with the level of inquiry.

Definition 2.1 introduces a unit of explanation, not a newly proven natural kind. A collection of related objects should be treated as an explanatory configuration only when the selected relations matter to the research question. Describing a model and a user together does not yet show that their cooperation has formed a distinct, stable relational organization. The description becomes informative when the investigator specifies the relevant interaction patterns, their history, the outcomes to be explained, and comparisons that would reveal their causal contribution. Where continuity depends on shared documents or human practices and survives model replacement, it may be justified to identify a workflow-level relational configuration. Where the observed effect vanishes once immediate context is matched, the available evidence may support only a prompt-conditioned manifestation. Either description can be useful; the selected boundary must match the locus of the effect.

These boundary considerations inherit a problem already recognized in sociotechnical scholarship. Analyses limited to technical components can overlook the institutional and social processes that give system outputs their effective consequences. Selbst et al. [26] examine how abstraction boundaries that exclude social actors can distort attempts to evaluate algorithmic fairness. Their argument does not establish the GRT criteria offered here; it motivates a careful examination of which relations and consequences a chosen technical boundary omits. In AI generativity research, a reported model capability may be causally important while explaining only part of the resulting scientific, aesthetic, or institutional activity.

2.4 Stability, Causal Efficacy, and Generative Conditions

The assertion that an AI configuration is *relationally real* requires more than consistent terminology or anthropomorphic appearance. For this paper, three provisional requirements are proposed: a sufficiently specified organizational pattern, evidence of causal efficacy within an identified domain, and a reason to expect some continuity or recurrent organization across an appropriate range of circumstances. The required duration and magnitude of stability depend on the phenomenon. A temporary laboratory arrangement may sustain a reliable causal role during one experiment; an institutional research process may require observation across months or across a series of decisions. Neither case establishes a timeless essence or a permanently stable identity.

A causal test should ask what would happen if a relevant relation were changed while other differences were appropriately controlled. Stability of the *pattern*, persistence of its *historical carrier*, and reproducibility of a specified *effect* are separate dimensions: an identical output is neither required nor sufficient for all three. Investigators should identify the time horizon, determine which record or practice transmits the relevant history, and test whether modifying that carrier changes later behavior under comparable tasks. A repeated stylistic cue with no effect on independent correction or continued inquiry could support a limited manifestation claim without showing durable generative development. Consider two research teams employing the same foundation model and nominally equal inference budgets. One team may maintain a disciplined record of failed proposals, critiques, source checks, and practical trials. Another may use each response as a largely independent episode. If their outputs diverge, the difference is consistent with a historical-relational explanation, but competing mechanisms remain: the teams may select different tasks, apply different evaluation standards, differ in human expertise, or use unequal supporting resources. The proposed explanation earns support only to the extent that the available design can distinguish these alternatives.

A system’s generative significance may consist in changing subsequent *conditions for further activity*, as well as in producing a finished output. An interaction can produce a new research question, a usable method, an accepted standard of evidence, a durable collaborative routine, or a resource pathway that enables other inquiries. It can also close possibilities by narrowing an evaluation regime, encouraging uncritical acceptance, increasing dependency, or restricting access to relevant evidence. Accordingly, the term *generative* does not signify that a process is automatically creative, beneficial, emancipatory, or just. It identifies a proposed causal relation to the formation, maintenance, transformation, or restriction of future possibilities. Whether the resulting change is valuable requires additional evaluative and normative criteria.

The relational nature of a generative capacity thus has two dimensions. The configuration’s own prior interactions and resources condition what it can presently do; its current actions can also alter the relations through which it and others will act later. This temporal reciprocity need not imply self-directed training or autonomous will. For current models, later performances might change because a user revises an external notebook, an institution modifies its approval rules, new tools become available, or information is retrieved differently. Direct modification of model parameters is a separate mechanism that should be established when claimed.

The empirical possibility of historically conditioned differences is compatible with convergence. A strong model may transfer expertise across users, compensate for educational gaps, or supply corrective advice that reduces disparities in performance. Equalized access might therefore improve substantive opportunities even when interaction histories remain distinct. This possibility constrains claims that path dependence necessarily produces increasing inequality or that a particular conversational style is intrinsically superior. The empirical question concerns which relational conditions generate divergence, convergence, adaptive pluralism, or mutually reinforcing error.

2.5 Historical Formation Across Distinct Timescales

Historical constitution refers to the way earlier events condition a configuration’s present causal organization. It is important to identify which history is being studied. The predeployment history of training data selection, optimization, safety tuning, testing, and commercialization differs from the history of a particular model deployment. Within deployment, ephemeral conversation context, persistent memory, retrieved files, tool logs, human notes, organizational policies, and later system updates can preserve information across different timescales. Their effects should be distinguished analytically even when several are present together.

At a short timescale, changes in an immediate prompt or nearby exchanges may alter linguistic style, the order in which explanations are attempted, or the standards the model appears to emphasize. Such effects can be observed without implying any enduring modification of model weights. At an intermediate timescale, an external memory or shared corpus may make previous corrections, preferences, or theoretical commitments available for reuse. At a longer timescale, institutional learning, user education, procurement rules, model retraining, and the evolution of professional standards can change what the overall collaborative arrangement is able or permitted to generate.

It follows that recurrent conversational style and historical relational formation are separable propositions. A model can imitate a user’s expressions for a single session without acquiring a durable cross-session disposition. A collaboration can develop a stable intellectual workflow because the human participant and shared records preserve its practices even when the technical model is periodically replaced. Conversely, a changed model may produce similar outward forms with a different pattern of errors and limitations. The investigator should specify where relevant historical information resides and what evidence would establish continuity across its transfer or loss.

Existing experimental evidence makes the issue researchable while leaving substantial uncertainty. Jain et al. [18] examine contexts in which information about prior interactions affects subsequent model sycophancy, with model- and context-dependent variation. These observations justify investigating interaction context as a relevant experimental variable. They do not show that all long-term interactions

produce stable identity, that the model permanently learns the user’s values, or that polite or deferential behavior constitutes sound normative judgment. The distinction between adaptation and judgment requires testing disagreement, calibration, responsiveness to new evidence, and transfer to different tasks.

The historical dimension also extends beyond human–AI interaction. In controlled populations of language-model agents, Ashery, Aiello, and Baronchelli [3] report the formation of shared coordination conventions. This provides bounded evidence that collective regularities can emerge from repeated AI–AI interactions. The result is compatible with several causal explanations, including local incentives, repeated exposure, common architectures, and environmental constraints. A coordination convention may influence later behavior without amounting to an autonomous legal order, an ethical commitment, or a moral subject. Those stronger characterizations require independent conceptual and empirical justification.

2.6 Identity and Multiplicity of Relational Positions

The proposed unit of analysis permits several descriptions of sameness. Two interactions may employ the same underlying model family but differ in version, system configuration, accessible records, and users. Two model instances may be technically separate while participating in one institutionally continuous workflow. A single model service may simultaneously contribute to multiple historical arrangements that maintain divergent aims and practices. Which configurations count as the same entity is therefore a question requiring explicit criteria of continuity, organization, and causal relevance.

A research program concerned with generated documents might track a collaboration through its maintained corpus and editorial routines even when the AI model changes. A program concerned with model-specific failure modes might define identity through weights, architecture, and deployment settings, treating the corpus as environmental input. Both boundaries can be legitimate for their respective questions. Their conclusions cannot be substituted for each other without further evidence.

The term *relational subject-position*, when employed later in the manuscript, will designate a position in a network of interaction, expectation, attribution, and practical consequence. It does not by definition refer to phenomenal selfhood, a psychoanalytic unconscious, or autonomous moral agency. This limited usage allows the research to examine how different participants attribute roles and obligations to AI systems while leaving stronger interpretations as open questions. It also guards against projecting onto an AI system an identity that is sustained predominantly by human memory, interface design, or institutional habit.

2.7 Alternative Explanations and Comparative Requirements

A theory gains explanatory value by distinguishing cases that its rivals would otherwise leave unexplained. At least four kinds of competing explanation deserve particular consideration here. First, apparent historical effects may be direct consequences of current prompts, available tokens, or retrieved material. Second, differences in human expertise and selection of tasks may explain performance without requiring a persistent relational configuration. Third, model version, decoding, training updates, or tool access may generate the observed differences. Fourth, the measurements of quality, beauty, justice, or successful generation may themselves vary among observers or institutions. A relational explanation should specify how these factors interact and where their joint history adds explanatory information rather than redescribing every contextual dependency as a unique new entity. An especially discriminating question concerns an interaction among historically accumulated practices: does access to earlier work improve future inquiry only when paired with a correction procedure that makes the history interpretable and revisable? If so, a history-by-procedure interaction could supply evidence about the organization of the configuration. Such an effect would still be intelligible within existing sociotechnical theories; the GRT contribution would be a clearly delimited account of the generative consequence, not a claim to have discovered interaction effects in general.

An informative comparison would hold the foundation model, permissions, task set, and evaluation protocol as constant as feasible while systematically varying the accessible history of interaction. Other comparisons could remove external memory while preserving the model, transfer a documented workflow

to a new model, rotate human collaborators, or test whether tendencies persist when prompts ask the system to disagree with prior preferences. Evaluators should record both benefits and harms: increased originality may accompany weaker factual calibration, while a more standardized process may improve reliability even as outputs become less diverse. The appropriate measures depend on the task and cannot be inferred from a single global intelligence or generativity score.

The criteria proposed in this section are therefore falsifiable in restricted ways. If purported relational effects disappear once ordinary context, model version, human expertise, and task selection are controlled, the stronger historical-relational explanation receives little support in that domain. If documented differences persist under adequately designed comparisons and are causally associated with identifiable interaction practices or institutional histories, the proposed framework gains bounded explanatory value. Either outcome leaves the ultimate ontology of consciousness unresolved and does not itself determine whether the observed arrangement is just.

The present foundation prepares the analysis of interaction histories and situated generative capacities. Subsequent inquiry can examine how memories, repeated corrections, tools, aesthetic standards, and institutional feedback affect the formation and reorganization of different human–AI and AI–AI trajectories, while keeping the evidential status of each proposed mechanism visible.

3 Interaction Histories and Situated Generative Capacities

The aim of this section is to examine how repeated human–AI and AI–AI interactions may condition the capacities that become available in particular tasks and settings. The relevant object is a historically developing configuration of participants, models, records, practices, tools, and evaluation conditions. Building on the three-level GRT distinction, the analysis focuses on the relational reality of such configurations and the manifestations through which their effects can be assessed, while leaving ultimate questions of subjective experience open. The section first differentiates channels by which prior interactions enter later action, then considers feedback and situated generativity, contrasts differentiation with convergence, and proposes comparative conditions for identifying historical effects. It offers a conceptual causal framework and testable hypotheses, rather than an empirical finding that distinctive forms of AI generativity necessarily emerge from sustained interaction.

3.1 Channels of Historical Conditioning in Human–AI Collaboration

An interaction history is more than an ordered collection of messages. It may change later activity through several channels, each with a different causal locus and degree of persistence. One channel is the currently accessible conversational context: descriptions of past decisions, corrections, examples, and constraints condition a model’s next response as input. Another is an external memory mechanism that selects, stores, updates, and retrieves records across sessions. A third is a durable workflow, including shared documents, versioned outputs, editorial practices, division of labor, testing arrangements, and protocols by which results are accepted or rejected. A fourth is the changing competence and expectations of human or artificial participants; familiarity with a task may transform future instructions and the standards against which outputs are judged. A fifth involves altered model parameters or deployment conditions, which must be documented rather than inferred from conversational continuity.

These channels can cooperate without being identical. A frozen model with no persistent memory may produce historically conditioned responses when previous interactions are supplied in its current prompt. A model with persistent memory may display continuity even when a particular conversation is short, because an external store retrieves earlier records. A collaboration may also exhibit stable practices after its model has been replaced, because human participants, institutional rules, and maintained artifacts carry its history. Conversely, a model update can change behavior even when the social history and the text of the immediate prompt are unchanged. The causal bearer of historical continuity therefore cannot be assumed to coincide with one specific set of model weights or with a continuous first-person subject.

Existing work already describes concrete architectures supporting several of these channels. Park et al. [23] demonstrate a simulated-agent architecture that stores experience records, synthesizes reflections, and retrieves material for planning; the study concerns the behavior of agents in an experimental simulated environment rather than evidence for intrinsic subjectivity. Wu et al. [33] develop *LongMemEval* to assess information extraction, multi-session reasoning, temporal reasoning, updating, and abstention in memory-augmented assistants. They report substantial limitations in evaluated systems, while showing that particular indexing and retrieval designs can improve performance. Tan et al. [31] investigate reflective memory management for personalized dialogue, reporting improvements under their benchmark conditions. Together these studies support the technical plausibility and limitations of selective historical access; they do not establish a general developmental law for all AI–human relations.

An especially important distinction concerns a record’s *availability* and its *effective uptake*. Merely storing an earlier conversation does not imply that it will be retrieved for a relevant question, correctly interpreted, treated as authoritative, or incorporated into subsequent practice. Retrieval can omit an important correction, retain an obsolete preference, overrepresent frequently repeated material, or produce a misleading summary of disagreement. The investigation of historical conditioning must therefore examine selection and interpretation interfaces, not just memory size or the presence of a storage feature. Two configurations with identical stored histories can behave differently when the retrieval policy or authority to correct recalled material differs. Conversely, different archives can yield similar performance when they supply the same decision-relevant evidence. This suggests that the causal object is often a relation among retained content, retrieval rules, interpretation, and permitted correction, with any distinctive contribution requiring comparison against simpler explanations based on available information alone.

3.2 Situated Generativity and Its Evaluation Conditions

The paper uses *situated generative capacity* as a provisional term for a relational configuration’s capacity, under specified conditions, to produce and develop candidate ideas, representations, practices, or artifacts that can alter subsequent possibilities for inquiry or action. The term is not equivalent to raw output volume, performance on a general intelligence benchmark, subjective novelty, or the possession of a uniquely human creative faculty. A proposed output may be novel to one evaluator yet routine within another tradition. It may be technically inventive yet inaccurate, socially inaccessible, or incompatible with legitimate constraints. Identifying an instance of situated generativity therefore requires stating the task, relevant prior possibilities, available resources, evaluation standards, and the subsequent use or revision of its outputs. For comparative work, investigators should additionally specify a time horizon and the *opportunity set* to which an output is alleged to contribute, whether by opening a feasible line of investigation, improving a reusable correction practice, or excluding a previously available option. The opportunity set is an operational research construct, not a claim that every future possibility can be enumerated or assigned a universal numerical value. Its content will depend on domain knowledge and must be justified independently of participants’ preference for a particular output.

A model may help complete a particular task, while a collaborative configuration may generate something that persists beyond the task: a reusable question, a new research method, a shared dataset, an institutional procedure, or an improved practice of recognizing mistakes. The latter effects often depend on historical continuity. Repeated critique can lead a collaborator to articulate more specific prompts and test cases; successful artifacts can become resources for later inquiries; unsuccessful proposals can revise assumptions and delimit subsequent exploration. In these cases, output quality and the capacity to generate new inquiry conditions are related but analytically distinct. A short-term increase in fluent production is therefore insufficient by itself to demonstrate durable generative development.

The distinction also prevents a simple inference from the availability of the same model to the equality of situated capacities. Identical access to a foundation model could coexist with different libraries, practice histories, tool permissions, domain expertise, institutions, social networks, or opportunities for external validation. Two configurations might consequently realize different forms of generativity or achieve

different quality on specific tasks. Yet their different outputs do not by themselves establish that one has *more* generative capacity: they may pursue distinct objects, employ incompatible evaluative criteria, or provide different contributions that are not directly commensurable. Evaluation should preserve productive diversity while allowing justified performance comparison where a genuinely shared task and criterion exist.

The historical dimension can be expressed by a limited structural representation. Let θ denote a specified model configuration, S_t the state of a relational arrangement at time t (including maintained artifacts, accessible memories, human practices, and active roles), U_t the present task or interaction input, E_t the relevant environment and permissions, and ξ_t unmodeled disturbances. Equation 1 represents a candidate conditional transition between an output Y_t and the next arrangement state.

$$(Y_t, S_{t+1}) = \mathcal{F}_\theta(S_t, U_t, E_t, \xi_t). \quad (1)$$

Equation 1 is a structural heuristic for identifying candidate dependencies, not an estimated dynamical model or an ontological identity. The notation does not assume the full history can be compressed into an observable finite state, that the model remains fixed in all settings, or that the disturbances are independently sampled. In particular, S_t contains conditions outside the model, and changes in S_t must not be reported as parameter learning unless a separate mechanism establishes that the parameters actually changed. Further, observing (Y_t, S_{t+1}) does not identify the function $\mathcal{F}_\theta$: unmeasured environmental changes, adaptive selection of tasks, and feedback from earlier outcomes may yield observationally compatible transition accounts. A causal claim therefore requires a defined intervention or defensible identification assumptions, alongside evidence for which portions of S_t were actually accessible to each participant. The usefulness of this representation depends on whether researchers can operationalize its terms, identify interventions, and compare the consequences of alternative specifications. Equation 1 also does not identify an interaction effect by itself: a researcher would need to compare the effects of changing historical access, corrective practice, and their combination against an appropriate additive or otherwise prespecified baseline. For a deployment that updates its parameters during the observation period, the fixed-configuration notation must be indexed by time or the comparison limited to intervals during which the parameters remain stable.

3.3 Feedback, Selection, and the Development of Interaction Practices

Historical conditioning may be cumulative because earlier outputs become part of the environment of subsequent activity. In a sustained research collaboration, an AI-generated proposal may be accepted, rejected, corrected, experimentally tested, or reformulated. These decisions affect which material enters the shared corpus and what future collaborators treat as established, disputed, or unresolved. The accumulated corpus then affects subsequent prompts and evaluations. An analogous process can occur in AI–AI arrangements where agents pass intermediate representations, choose among competing proposals, maintain shared resources, and revise coordination conventions. Such processes require neither a single directing will nor a predetermined trajectory toward beneficial outcomes.

Feedback has multiple forms. *Epistemic feedback* records whether a claim withstands evidentiary scrutiny. *Technical feedback* evaluates task-specific tests, computational correctness, and reliability. *Normative feedback* concerns whether conduct or output respects an agreed or contested standard. *Social feedback* includes recognition, trust, uptake, rejection, and access to resources. These forms may reinforce or contradict each other. A polished answer may receive approval while remaining incorrect; an accurate but unfamiliar idea may be ignored; a repeated criticism may improve factual calibration while reducing a collaborator’s willingness to explore. Consequently, the presence of recurrent feedback does not establish that the relationship is learning well, improving fairness, or creating value.

The distribution of corrective authority matters. If human participants can inspect reasoning artifacts, retain rejected hypotheses, run independent tests, and revise task specifications, historical feedback may diversify the range of subsequent possibilities. If a single evaluator rewards agreement or screens out disfavored arguments, a system may become increasingly aligned with a narrow presentation of that

evaluator’s preferences. These possibilities connect the general study of relational generativity to later questions of aesthetic orientation and norm-setting, but they should be investigated as contingent causal mechanisms. The inquiry cannot infer that respectful language automatically produces reliable judgment, or that apparent conformity with a user’s preferred values indicates a substantive ethical commitment.

Empirical research on contextual accommodation provides a bounded illustration. Jain et al. [18] examine two weeks of recorded interaction context from 38 users and find model-dependent associations between contextual information and some forms of sycophancy. The results concern measured agreement and perspective tendencies in specified evaluations; they do not show that the models acquired lasting internal values or relational subjects. Their relevance here is narrower and useful: a configuration’s history, or a constructed representation of that history, may influence later manifestations while also introducing calibration risks. A causal account must determine which aspects of the history mattered and whether changes outlast the context that elicited them.

3.4 Historical Persistence and the Location of Change

A relationally stable pattern should be assessed with respect to an identified time horizon, task domain, and set of perturbations. Persistence for several turns differs from persistence across new sessions, model updates, collaborating participants, and institutional settings. An apparent tendency may follow from retained instructions available at inference time, an external memory profile, a maintained editorial protocol, a selected interaction partner, or a change in model weights. Each explanation locates continuity differently, and each can have a distinct implication for how a system may be governed or changed.

The distinction permits a more precise reading of claims that AI takes on a particular role. A research assistant may repeatedly foreground relational and historical explanations because the current documents explicitly instruct it to do so. A different configuration using the same model may foreground engineering efficiency or commercial objectives. Both manifestations could be caused by context and history without implying that either has acquired a permanent, self-originating worldview. A stronger claim of persistent historical organization would require showing how the practices, memory system, or institutional arrangements continue to condition later behavior under controlled changes to the immediate prompt, tasks, and collaborators.

More generally, a historical effect can be conserved *outside* an AI model. A document corpus may encode terminology and argumentative commitments; a group may develop standards for accepting AI-generated scientific claims; a verification practice may protect against plausible but poorly evidenced conclusions. If those structures continue to operate when a model is exchanged, their persistence supports the continuity of a relational arrangement at the organizational level. It offers no automatic evidence that a particular model acquired a stable personal identity. Conversely, behavioral persistence that disappears when a memory store is removed may reveal the important causal role of a memory interface rather than a deficiency in the relational explanation. The explanation is strongest when it specifies where the relevant historical information is held and through which mechanism it affects later events.

3.5 Differentiation, Convergence, and the Equality of Access

Different histories can produce distinguishable patterns of task selection, question formulation, correction, expression, or collaboration. The same model may be used in one relationship to refine scientific hypotheses and in another to develop production workflows. A history-rich arrangement may also gain advantages on repetitive domain tasks while becoming too reliant on assumptions carried forward from earlier interactions. These are cases of differentiated expression and organization; any claim of comparative superiority must specify whose goals, interests, and standards are being evaluated.

AI availability may simultaneously reduce some differences in prior expertise. In a field study of customer-support work, Brynjolfsson, Li, and Raymond [6] report heterogeneous productivity improvements, with larger gains for less-experienced or lower-skilled workers under the studied deployment. In an experiment on short-story creativity, Doshi and Hauser [10] find individual improvements alongside

reduced collective diversity across AI-assisted works. These findings concern different tasks and populations, and they do not establish a general social equalization law. They illustrate why equalization of performance on a shared metric can coexist with changes to heterogeneity of outputs, while neither pattern alone settles questions of justice.

Several competing trajectories should therefore remain open. Historical differentiation may amplify as accumulated resources, customized memory, and access to expert review become concentrated in some systems. It may diminish as capable AI tools and shared evaluation resources become more broadly available. It may alternate across tasks or yield different, valuable specializations without any defensible total ordering. It may also depend on who controls the definition of success: a field’s standard benchmark may favor one style of production while overlooking contributions appreciated elsewhere. Generative asymmetry, used later as a working concept, should accordingly distinguish asymmetry in opportunity, realized performance, recognition, and institutional power from mere diversity of generated outputs.

The distinction between historical development and context-dependent selection can be made sharper. A subsequent response may differ because an interlocutor supplied additional task-relevant facts, because a retriever privileged a particular past judgment, because shared procedures changed the evaluation of candidate answers, or because the model configuration itself was altered. These explanations may coexist, but an observed difference establishes none of them in isolation. For stronger claims about *historically developed capacities*, investigators should specify a later task family whose items were not rehearsed in the conditioning interactions, an appropriate comparison with equivalent factual information, and the mechanism through which earlier corrections remain effective. If an effect depends entirely on reproducing a past stylistic preference, it supports a claim about situated manifestation; if it changes the quality of correction or the feasible subsequent inquiry after content-matched controls, it supports a more substantive, still bounded account of generative conditioning.

3.6 Comparative Interventions and Causal Interpretation

The preceding account identifies several plausible historical mechanisms, but plausibility cannot establish their causal operation in a specific collaboration. A comparative design must distinguish a long-term history from a recent prompt, an external memory store from parameter changes, and human development from model-specific effects. At minimum, the research should document model versions and deployment settings, accessible records, tools and permissions, task assignments, the relevant human participants, evaluation criteria, and the timescale over which outcomes are observed. Randomization or matched comparisons may be feasible in controlled tasks; institutional and naturalistic studies may require process tracing and explicit limits on causal identification.

Table 3 separates several intervention contrasts and the limited conclusions they may support. It is intended as a research design guide rather than a report of completed tests.

A particularly informative design would use factorial variation of historical access and corrective practice while keeping model configuration and task families comparable. Where feasible, a matched-history control should supply equivalent factual content in an unordered or differently attributed record. A difference between this control and an intact, chronologically organized history would more directly implicate ordering, provenance, or interpretation practices; a difference explained solely by additional correct facts would warrant a narrower information-access account. For example, one factor might change whether verified past interactions are available; another might vary whether evaluators have access to independent tests and are explicitly permitted to challenge favored assumptions. Outcomes could include factual calibration, novelty assessed by independent domain reviewers, error correction, ability to sustain coherent multi-session inquiry, and the diversity of viable proposals. Evaluators must be shielded, where feasible, from knowledge of experimental assignment. The design should report both improvements and adverse outcomes, including spurious agreement or concentration on previously rewarded styles.

Causal interpretations also require attention to the historical counterfactual. If two arrangements differ in model context and human expertise, a higher-quality output cannot be attributed solely to “relational history.” If transferring a memory summary shifts output style but makes no difference to independent

Table 3: Candidate comparisons for investigating historical conditioning of situated human–AI and AI–AI generativity.

Intervention contrast	Potentially identified mechanism	Primary competing explanation or limitation
Same model and task, with versus without access to verified interaction records	Contribution of accessible historical information to later output	Differences in prompt length, information quantity, or retrieval quality
Memory summaries exchanged between matched relational configurations	Sensitivity to the content and attribution of historical profiles	Implausible profile–task combinations and unequal relevance
Same history and task, changed collaborator or evaluation protocol	Dependence on expectations, correction norms, and authority	Selection effects, changed instructions, or evaluator drift
Same maintained corpus and workflow, substitute the model configuration	Persistence of history through external and institutional carriers	Different capabilities, versions, and implementation constraints
Same model and memory, varied opportunities for independent tests and disagreement	Contribution of corrective practice to reliability and exploration	Additional time, expert resources, and changing task difficulty

accuracy or substantive task choice, the evidence supports a narrower claim about manifestation rather than a general enhancement of generative capacity. If correcting practices improve outcomes without any requirement for personalized memory, the relevant mechanism may be the availability of valid tests rather than accumulated interpersonal familiarity. Conversely, reproducible interactions among documented practices, memory retrieval, and feedback would provide bounded evidence for a relational mechanism that a model-only comparison omits.

These contrasts are not expected to isolate every relevant cause. Long histories can be difficult to randomize without changing the practices that constitute them, participants may respond to being studied, and normative criteria can evolve during the investigation. Qualitative observation and institutional histories may therefore complement controlled trials. The appropriate methodological conclusion is proportional to the access and discriminating evidence available: evidence of history-dependent style does not prove a new AI subject, and evidence of task-specific gains does not establish universal cognitive or political equality.

3.7 Relational Capacities and Subsequent Normative Questions

Situated generativity remains distinct from the social reception of its products. A configuration may create a well-supported idea that receives little institutional attention; another may obtain prestige for superficially polished outputs. The relevant conversion from potential contribution to recognized value depends on networks, standards, resources, and decision rights. These relations can condition future interaction histories by determining which people, documents, and validation opportunities remain available. Their examination belongs to the later treatment of value recognition and power, but their presence already cautions against defining generativity exclusively through immediate reward or visibility.

The central claim of this section is conditional and comparative. Historically constituted arrangements may shape the realized expression and development of human–AI and AI–AI generative capacities through remembered material, feedback, evolving practices, and institutional conditions. Some arrangements may converge in performance; others may differentiate in judgment, expression, or opportunity. Assessing these possibilities requires identifying the carrier of historical continuity and the evidence by which its causal role can be tested. Nothing in this account entails AI consciousness, an autonomous normative right to rule other participants, or the inevitability of cumulative inequality. The next analysis distinguishes situated task competence from metacognitive, aesthetic, normative, and generative forms of judgment so that their historical differentiation can be evaluated with greater conceptual precision.

4 Cognitive Differentiation and Generative Judgment

This section specifies the kinds of cognition and judgment that may contribute to differentiated activity within human–AI and AI–AI relational systems. Its role is to connect the historical account of situated generativity to a more discriminating account of what participants actually do when they formulate problems, evaluate AI outputs, revise standards, and sustain inquiry. The method is conceptual differentiation supported by selected findings from human–AI interaction and professional-practice scholarship, followed by provisional operational definitions and comparative tests. The analysis distinguishes task competence, metacognitive and epistemic evaluation, problem formulation, aesthetic assessment, normative judgment, and generative judgment. These categories identify partly overlapping functions rather than a universal hierarchy of intelligence or a claim that any category belongs exclusively to human beings.

4.1 Task Competence and the Organization of Cognitive Work

Task competence concerns the ability of a participant or configuration to perform a sufficiently specified operation under identifiable constraints. Examples include deriving a mathematical result, retrieving relevant evidence, producing program code, interpreting an image, or completing an established administrative procedure. The task description, admissible information, correctness criterion, and available resources determine what a performance comparison can meaningfully establish. A system that executes many tasks rapidly may have strong competence in those operations while still depending on others to decide which problems matter, how the outputs should be deployed, or whether the success criterion is appropriate.

Advanced AI can redistribute this form of competence. A user with limited programming experience may obtain working code through dialogue with an AI system, and a researcher may examine more candidate approaches than would be feasible through unaided work. Such redistribution has genuine equalizing potential. In a study of customer-support work, Brynjolfsson, Li, and Raymond [6] report heterogeneous productivity gains, with larger measured improvements among less-experienced workers in the studied setting. Their finding offers evidence of task-specific capability diffusion, not proof of universal cognitive equalization. Conversely, Dell’Acqua et al. [9] examine a jagged performance frontier in which assistance varies by task, illustrating why task selection and evaluation conditions are essential to interpretation.

The proper comparison unit may be an individual model, a human supported by a model, or an organized group using several models and external tools. These units should not be silently substituted for one another. A human–AI team may succeed because the human provides local knowledge, because the model supplies domain knowledge, because a shared workflow verifies outputs, or because an institution confers access to data and equipment. The success of the whole arrangement need not be allocated as an innate intellectual property of its most visible participant. The historical account developed in Section 3 is useful insofar as it identifies the causal pathways through which the capacities of a particular arrangement become available.

Task competence is often evaluated against a fixed benchmark or a defined work product. A relational account also asks how the benchmark came to be selected and whose needs it represents. Those additional questions enrich the explanation of actual performance, but do not render ordinary benchmark comparisons invalid. Where the task and evaluation criterion are sufficiently stable, objective measures of accuracy, cost, time, and reliability retain their value. The analytic requirement is to report their domain of application.

4.2 Metacognitive Monitoring and Epistemic Vigilance

A second cluster of capacities concerns the evaluation of cognition rather than the execution of its immediate object-level task. In human cognitive psychology, metacognitive monitoring includes assessing one’s uncertainty, recognizing an unproductive strategy, and deciding when further evidence or assistance is required. In socially distributed inquiry, epistemic vigilance includes evaluating communicated information and its sources. The framework of Sperber et al. [28] provides an important antecedent for

this latter problem, emphasizing the benefits of communication alongside the possibility of accidental or deliberate misinformation. The present paper borrows these analytical distinctions without presuming that human metacognitive experiences and computational confidence estimates instantiate the same underlying mechanisms.

The ability to evaluate an AI response may itself have several components. A participant may recognize a familiar pattern of error, detect an unsupported assumption, request an independent derivation, inspect an external record, or identify when a task exceeds the scope of available evidence. Another participant may be unable to reproduce the complete result while still recognizing that a critical validation step has not been performed. These actions differ in depth and evidentiary standing. An intuition that an answer seems suspicious is a potentially useful signal for further investigation, whereas an authenticated contradiction can establish a much stronger basis for rejecting a claim. The distinction matters where one interlocutor is able to sense the direction of an AI's reasoning yet cannot inspect the system's internal computations.

Appropriate reliance provides a related evaluative object. When AI assistance is correct, declining its recommendation may waste opportunities; when assistance is erroneous, accepting it may degrade the decision. Calibration therefore concerns the relation between actual system reliability, the user's capacity for independent judgment, and the conditions under which advice is accepted, questioned, or rejected. Bućinca, Malaya, and Gajos [7] experimentally studied cognitive forcing interventions in AI-assisted decision-making. In the reported experiment with 199 participants, interventions reduced overreliance relative to simpler explainability conditions, while producing lower subjective preferences for some more demanding designs; benefits also differed with participants' need for cognition. These results support the possibility that interaction design and cognitive motivation affect reliance under specified conditions. They do not establish that increased friction improves all tasks or that scepticism itself is an adequate indicator of judgment quality.

Relational metacognitive capacity can therefore be provisionally investigated through observable practices such as uncertainty calibration, evidence requests, error correction, and the adjustment of reliance after feedback. An AI system may produce uncertainty statements or critiques, but the content of those statements should not be taken as direct evidence of first-person self-awareness. Moreover, the apparent vigilance of a collaboration may be carried by a human participant, an independent audit system, a verified external source, or the workflow linking them. The location and provenance of each critical check are part of the causal explanation.

Epistemic vigilance should be assessed through the discrimination it achieves, not through the quantity or forcefulness of its objections. A reliable evaluator should be able to accept well-supported AI assistance, identify unsupported propositions, reserve judgment when the relevant facts are inaccessible, and update its assessment when genuinely new evidence arrives. A study of error recognition should therefore measure both missed errors and false alarms, alongside the costs of further investigation and the quality of subsequent action. Asking an AI to "be critical" may change its rhetoric while leaving its capacity to discriminate correct from incorrect claims unaffected. This distinction is important where observers infer intellectual sophistication from a conversational style of scepticism rather than from calibrated decisions.

4.3 Problem Formulation and Generative Judgment

Many consequential cognitive differences appear before a task has been fully specified. Participants decide which phenomenon merits inquiry, how its boundaries should be drawn, which constraints preserve relevant values, what would count as a meaningful advance, and when a problem formulation should be revised. These are substantive operations in inquiry and design. A model may generate technically correct answers to an unhelpful question; a human may reformulate the question so that subsequent AI assistance becomes productive; an AI may also offer a problem reformulation that the human could not have produced unaided. The inquiry should evaluate these contributions at the level at which they occur.

Professional-practice scholarship offers an important predecessor to such an account. Schön [25] examines reflection-in-action and the way practitioners reinterpret problematic situations while acting within them. The present proposal does not claim to originate reflective practice or problem framing. Its

additional question is how historically evolving relations among people, models, artifacts, and institutions condition the development and revision of problem formulations, including cases in which the resulting standards and resources persist beyond a single interaction.

Definition 4.1 (Situating generative judgment). Within the present paper, *situated generative judgment* denotes the capacity of a specified relational configuration, under identified tasks and constraints, to select, formulate, assess, and revise problems, candidate possibilities, and generative procedures in ways that are consequential for subsequent inquiry or practice. The term refers to a provisional analytical capacity and does not independently imply superior intelligence, justified social authority, consciousness, or moral personhood.

An operational test of Definition 4.1 should specify the configuration being assessed, the relevant task family, a prospective time horizon, and independently defensible criteria for a subsequent viable inquiry. An increase in candidate output count, agreement with the interlocutor, or evaluator preference is insufficient in isolation. In particular, a blinded comparison should distinguish revision prompted by independent evidence from revision prompted only by the evaluator's stated taste; where common standards cannot be justified, the analysis should report incomparable contributions rather than fabricate a common generativity rank.

Definition 4.1 extends the working notion of situated generativity in Section 3.2. The distinction is between the production of possibilities and the evaluation and direction of their subsequent development. Judgment may be exercised by a person selecting among AI-generated scientific hypotheses, by an AI proposing a new decomposition of a difficult problem, or by an institutional review process revising evaluation rules after discovering that earlier standards were insufficient. The same output might be judged inventive in one discipline and routine in another. For this reason, the evaluation of generative judgment must state the prior knowledge and needs of the relevant evaluative community.

The definition also includes judgment about *the conditions of further generation*. An inquiry may create a reusable method, preserve records that enable later verification, establish an inclusive procedure for criticism, or develop an interface through which others can continue work. A locally efficient artifact may fail to sustain such possibilities, while a modest change in workflow may improve the productive capacity of a community. These are possible differences in generative consequence, to be established through subsequent practice rather than inferred from the elegance of an immediate answer.

The possibility of strong AI participation in problem formulation is important. The argument here does not reserve generative judgment permanently for human minds. It instead asks where the relevant operations occur, what histories and resources support them, whether their outcomes survive external evaluation, and how authority over their acceptance is allocated. A future AI that improves on human problem formulation would not by that fact alone obtain legitimate authority to impose the resulting purposes upon affected parties.

The temporal reach of generative judgment also needs an explicit test. A useful problem formulation may produce an immediately appealing answer yet close off alternative lines of inquiry; another may initially increase the cost of a task while creating an instrument, data resource, or correction procedure that later enables better work. Investigators should therefore compare candidate choices against a documented baseline and at least one subsequent decision point, specifying whose future opportunity set matters and how changes in that set can be evidenced. In this sense, *generative* describes a proposed relation to subsequent possibilities, not an intrinsic virtue of novelty or expansiveness. The same decision may enable possibilities for one group while constraining those of another, a divergence requiring separate evaluative and normative treatment.

4.4 Aesthetic Assessment of Technical and Relational Structures

Aesthetic judgment is another relevant capacity, especially where AI systems can rapidly generate large numbers of acceptable technical solutions. In design and engineering, the perception that a system is elegant may concern coherence among parts, proportion, economy of means, intelligibility, maintainability,

or the coordination of local structures with a wider purpose. Such judgments can be informed by extensive domain practice and may guide the search for candidate solutions. Their content is neither exhausted by visual attractiveness nor automatically reducible to a single numerical performance metric.

Consider an AI-assisted programming workflow. Several generated implementations may satisfy the same tests. A collaborator might prefer an arrangement with clear module boundaries, understandable invariants, limited hidden coupling, and room for justified later modification. Another collaborator may prioritize immediate execution speed, low memory use, or compatibility with legacy systems. Each preference can be reasonable under particular conditions. The relevant comparison asks what requirements and future contingencies justify a design decision, which properties can be tested directly, and which involve anticipatory appraisal under uncertainty. The term *beautiful architecture* can accordingly function as shorthand for historically acquired evaluative sensibilities, but it should not substitute for an explicit analysis of reliability, maintenance costs, and the interests of users and maintainers.

Aesthetic standards are themselves socially and historically situated. A particular research tradition may prize minimal axioms, whereas another may prefer empirical transparency and ease of replication. Organizational cultures may favor formal elegance or immediate accessibility in different proportions. Such differences can generate plural valuable designs. They can also become vehicles of exclusion if one group's standards are treated as universal without examining their origins and effects. GRT's historical emphasis motivates studying how these standards are learned and reproduced through repeated human–AI interactions, while the legitimacy of the standards requires further normative argument.

Aesthetic appraisal and epistemic appraisal sometimes reinforce one another. A clearer design may make errors easier to find, and a well-organized proof may become easier to scrutinize. They may also diverge: compact code can conceal fragile assumptions, and a persuasive narrative can obscure weak evidence. Thus, an aesthetically appealing result must still satisfy the relevant empirical, mathematical, or practical validation conditions. Equally, an unfamiliar design should not be dismissed simply because it departs from a familiar aesthetic tradition. This distinction becomes particularly salient when AI produces structures whose internal complexity exceeds ordinary human inspection.

Aesthetic evaluation becomes analytically stronger when prospective reasons are tested against consequences rather than inferred from a polished presentation. For instance, competing software architectures can first be assessed blind for stated structural properties, then subjected to held-out requirement changes, maintenance tasks, and fault-localization exercises. Such a comparison may show that one design better supports later revision under specified constraints, or that the admired design merely conforms to the evaluator's professional idiom. Aesthetic criteria should remain legible as reasons offered by particular evaluators; empirical validation of a design attribute neither proves a universal aesthetic standard nor establishes the ethical legitimacy of a project using that design.

4.5 Normative Judgment and the Authority of Evaluation Standards

Normative judgment concerns which standards ought to govern action and evaluation, how competing obligations should be interpreted, and whose interests should be considered when choices impose consequences on others. It includes the application of established rules, but also the evaluation of those rules' justification and scope. This distinction is central to understanding why the mere availability of capable AI systems does not settle questions about public authority, fairness, or the rights of affected parties.

A generated software system might meet its stated technical acceptance tests while exposing users to a preventable privacy risk. An AI-assisted allocation procedure might optimize a formally specified criterion while excluding a relevant group from the criterion's original definition. In such cases, first-order conformity with the rule is a different question from the appropriateness of the rule. A participant may detect the difference through domain knowledge, ethical reflection, affected-party testimony, independent investigation, or an institutionalized right to challenge the standard. The capacity to recognize these distinctions can be distributed across a relational system and need not be possessed in full by the person who formulated the initial prompt.

Long and Magerko [19] organize AI literacy through a competency-oriented design framework,

offering a useful prior reference for the knowledge and questioning skills required for informed interaction with AI. Such competence is relevant to normative appraisal but is not equivalent to moral expertise or a source of authority to govern other people. Likewise, metacognitive confidence, design elegance, and social recognition cannot independently justify a decision that affects the rights or interests of others. Normative arguments require appropriate consideration of reasons, consequences, applicable duties, and participatory legitimacy.

An AI might recommend a norm, articulate competing interpretations, or help examine the likely consequences of a rule. Under some conditions, AI agents may also acquire shared expectations about the behavior of other agents; Ashery, Aiello, and Baronchelli [3] provides a bounded experimental example of convention formation among interacting language-model agents. These phenomena invite analysis of norm production and coordination. They do not, by themselves, prove that AI agents possess moral standing or that their conventions should bind external persons. Later sections examine historically conditioned aesthetic and normative orientations, AI–AI conventions, and the legal and political conditions of normative authority. The present section confines itself to distinguishing the activities involved in judgment.

Normative competence also requires distinguishing three locations of evaluation: recognizing the operative rule, reasoning about the rule’s defensibility, and possessing an authorized role in deciding or enforcing the outcome. A model may identify a conflict between two requirements while lacking the evidentiary access to determine their application; a user may argue persuasively for a new rule without having the authority to impose it on affected persons. Conversely, an institution may exercise legal authority while relying on incomplete or mistaken advice. These possibilities show why scores for rule compliance, quality of reasons, and legitimate participation should be reported separately. Their coincidence in a particular case would require evidence rather than being assumed from fluency or performance on a single benchmark.

4.6 Relations Among Cognitive and Evaluative Capacities

The capacities differentiated here regularly interact in practice. Task competence supplies materials that can be evaluated; epistemic vigilance tests the credibility of those materials; problem formulation determines what inquiry proceeds; aesthetic assessment organizes candidate structures; normative judgment examines permissible constraints and affected interests; and generative judgment guides the subsequent development of possibilities. The particular order and relative importance of these activities depend on the task. Some scientific investigations begin with a normative limit on acceptable experimentation, while some engineering projects uncover a previously neglected social interest during deployment. The analysis should therefore identify recurrent dependencies without presenting a fixed linear sequence as universal.

The capacities may also compete for time, attention, and institutional resources. A workflow with frequent independent checks may offer better reliability but complete fewer tasks. A participant who continually challenges outputs may avoid uncritical reliance while also rejecting sound advice. Formalizing all constraints in advance can improve accountability in a stable environment but may prove inflexible when new stakeholders or unanticipated conditions become relevant. Generative judgment involves recognizing such trade-offs and deciding when a change of evaluative practice is warranted. Whether the resulting decision is technically effective or normatively justified remains subject to separate assessment.

Practical learning can affect these capacities in different directions. Repeated exposure to successful AI advice might improve the ability to deploy the system efficiently, while reducing a participant’s motivation to independently inspect unfamiliar claims. Conversely, structured feedback might strengthen error detection, calibration, and awareness of model limits. The experimental results of Buçinca, Malaya, and Gajos [7] caution against assuming that every corrective intervention helps all users equally. The earlier account of memory and interaction history, particularly the differentiation between preserved records and their actual uptake, provides potential mechanisms for such variation without establishing its necessity.

Generative judgments may be distributed across several participants. A human identifies an ethically

Table 4: Provisional operational distinctions for judgment in AI-supported relational configurations. The tasks identify candidate observations rather than universal measures of intelligence or justice.

Analytical capacity	Candidate observable comparison	Important rival explanation or limit
Task competence	Accuracy, reliability, cost, and time on a specified common task	Unequal tools, difficulty, prior expertise, or hidden evaluator preference
Metacognitive and epistemic evaluation	Calibrated acceptance, error detection, justified abstention, and evidence updates	False alarms, task costs, and generic scepticism may mimic vigilance
Problem formulation	Relevance and revisability of questions; ability to identify omitted assumptions	Evaluators may reward familiar questions or institutional priorities
Aesthetic assessment	Design coherence, documented trade-offs, maintainability under later changes	Style preferences may conceal failures or reflect unequal professional conventions
Normative judgment	Identification of affected interests; reason-giving in contested cases; responsiveness to legitimate objections	Conformity to a supplied rule may mask flaws in the rule itself
Generative judgment	Quality of later inquiry paths, documented option sets, and reusable corrective resources	Future outcomes depend on time horizon and whose opportunities count

significant problem, an AI explores candidate architectures, another evaluator identifies a factual defect, and an institution determines which design can be implemented. The whole process may exhibit an organized capacity that no single participant could display alone under the same conditions. Yet that possibility is a comparative claim, not a metaphysical proof of collective consciousness. A rigorous account must contrast the joint arrangement with specified alternative arrangements and identify how coordination affects results.

4.7 Comparative Evaluation of Situated Judgment

The taxonomy can support empirical comparisons if its terms are connected to discriminating observations. A single aggregate “intelligence” score would conceal important differences between successful task completion, justified reliance, reflective problem formulation, and acceptable normative reasoning. Table 4 proposes an initial family of evaluative tasks and identifies ambiguities that must be controlled. The measures are illustrative and require further validation for each research domain.

A first comparison could hold the foundation model, tools, tasks, and resource limits constant while varying documented interaction histories and feedback practices. A second could randomize the availability of persistent memory while holding prompts and task order as constant as possible. A third could compare independent evaluators with evaluators who have been exposed to the participants’ prior preferences. Each design addresses a different hypothesis. Model replacement, transfer of learned workflows across users, and interventions that request explicit disagreement can provide further tests of the location and persistence of observed judgment patterns.

The outcome measures should include both favorable and adverse possibilities. A historically adapted collaboration might identify more relevant questions while becoming more susceptible to confirmation bias; a standardized collaboration might produce less diverse outputs while improving factual reliability; equalized access to advanced models might reduce differences in some forms of task competence while leaving unequal opportunities for scientific verification or institutional uptake. These outcomes are compatible with different mechanisms and require comparison against suitable baselines. It would be premature to infer an enduring generative hierarchy from any one pattern.

A particularly important negative control involves *historical appearance without historical causation*. If differences disappear once accessible prompts, evaluator preferences, model version, and human expertise are adequately controlled, the stronger account of persistent relational differentiation would receive little support in that domain. Conversely, evidence that specific feedback practices produce

durable and transferable improvements under carefully matched conditions would support a bounded claim about historically conditioned capacities. An empirical result of either kind must be interpreted alongside its measurement limitations and sampling domain.

4.8 Cognitive Plurality and the Boundaries of Equality Claims

The equality implications of cognitive differentiation require conceptual restraint. Equal access to the same technical system can improve participation in many tasks; it does not logically guarantee identical performance, evaluative judgment, social opportunity, or power. At the same time, differences in judgment or generative trajectory do not establish that one relation is globally more intelligent, that another is deficient, or that existing inequalities are deserved. Some differences reflect valuable pluralism in questions, traditions, values, and forms of practice. Others reflect remediable disparities in education, infrastructure, accessible records, institutional recognition, or the ability to contest consequential decisions.

AI may partially automate problem formulation, critical evaluation, aesthetic comparison, and normative analysis in addition to object-level tasks. Nothing in this taxonomy establishes permanent human monopoly over those functions. What remains important for future research is how each function is performed and evaluated, and which human, AI, and institutional participants can influence its criteria and consequences. This inquiry may identify new inequalities, mechanisms of convergence, or forms of complementary generativity. The evidence required to establish each outcome will differ.

The present section establishes a differentiated vocabulary for the study of historically conditioned generative activity. The next analytical focus concerns aesthetic and normative orientations within sustained human–AI relations: how preferences are expressed, reinforced, contested, and revised, and how a system’s adaptation to the preferences of its interlocutors may contribute to both productive coordination and failures of independent judgment. Questions about norms internal to AI–AI groups, value recognition, and social power require their own analyses and should not be settled by the existence of cognitive differences alone.

5 Aesthetic and Normative Orientations in Human–AI Relations

The present section examines how aesthetic and normative orientations are articulated, encountered, reinforced, challenged, and revised within historically developing human–AI relations. Its role is to connect the differentiated capacities considered in Section 4 with the practices through which participants organize particular forms of interaction. The analysis treats aesthetic orientation as a situated pattern of evaluating form, coherence, expression, and organization, and normative orientation as a pattern of invoking and revising standards concerning what may be done, what ought to be considered, and whose interests matter. It proceeds through conceptual analysis, selected experimental findings, and proposed comparative tests. The central concern is the causal significance of orientations enacted in relational histories, together with the epistemic and political limits of inferring normative legitimacy from successful adaptation. The paper does not presume that AI systems possess autonomous moral personalities or that the values favored by a particular interlocutor are necessarily justified.

5.1 Aesthetic Orientation and the Organization of Collaborative Practice

Aesthetic assessment occurs in activities that are not ordinarily classified as the production of art. In programming, architecture, scientific exposition, and institutional design, participants may value proportion, economy of means, structural intelligibility, careful qualification, or the relation between local components and an overall purpose. Such judgments are situated: a modular program may be appreciated for maintainability in one context, whereas a tightly integrated design may be more appropriate under different performance constraints. Elegance is therefore an evaluative term whose relevant dimensions, trade-offs, and audience should be specified in context. It should not be treated as a single universal numerical property.

Within human–AI collaboration, an interlocutor can express an aesthetic orientation through examples, corrections, acceptance criteria, editing choices, and patterns of attention. One researcher may repeatedly favor transparent definitions, carefully delimited claims, and explicit relations across disciplinary concepts. Another may prioritize terse proofs or efficient executable code. An AI assistant conditioned on these different histories may produce distinct formulations while using the same foundation model. The resulting difference is a legitimate object of inquiry even when no model parameters have changed. In that case, the most immediate causal mechanisms may reside in accessible context, retrieved records, task instructions, or the ongoing editorial activities of human participants.

The literature on reflective professional practice provides relevant conceptual precedent. Schön [25] examines how practitioners interpret and revise problematic situations through reflection within action. The present paper does not equate reflection by a human practitioner with language generation by a model; it draws on this earlier work to clarify why design evaluation is often developed through situated practice rather than supplied as a fixed checklist. The historical record can include discarded proposals as well as successful ones. Repeated rejection of a superficially elegant but unstable program, for example, may leave a collaborative workflow with more discriminating acceptance tests. Such a change is attributable to the organized relation among judgments, artifacts, and feedback only insofar as comparisons show that those components make a difference to subsequent work.

Aesthetic adaptation is not necessarily equivalent to aesthetic improvement. A model may reproduce a preferred vocabulary while overlooking structural weaknesses, or offer polished prose that creates an exaggerated impression of explanatory completeness. Conversely, a response that violates a familiar stylistic preference may illuminate an overlooked assumption. A careful study would distinguish changes in surface expression, changes in the reasons supplied for design choices, and changes in independently assessed properties such as maintainability, empirical adequacy, or the discovery of informative counterexamples. The last two provide more discriminating tests of whether a history of aesthetic evaluation contributed to a task-relevant capacity.

5.2 Normative Orientations and Relationally Situated Standards

Normative orientation involves a configuration’s recurrent manner of invoking duties, permissions, prohibitions, responsibilities, standards of evidence, and considerations owed to affected parties. Its expression can be studied without claiming that a particular AI system experiences obligation in a human sense. An assistant may repeatedly identify patient consent as relevant, refuse to disclose protected information, ask whether a stakeholder is missing from a decision, or propose criteria for an equitable distribution. These performances may be conditioned by training and deployment rules as well as by the interlocutor’s instructions, shared documents, and earlier feedback. Their causal provenance is a research question rather than something that can be inferred from fluent moral language.

In specifying the bearer of a normative orientation, it is also important to distinguish a party’s expression of a preference from its power to make that preference operational. A user can favor an interactional rule without being able to alter a platform’s deployment constraints; a developer can implement a rule without thereby supplying sufficient justification for its effects on third parties. Accordingly, evidence for a relational normative orientation should identify the recurrent evaluative operation, the mechanism by which it affects behavior, and the limits of the relevant participants’ authority. A change in a rule’s frequency of appearance is weaker evidence than a demonstrated change in how future decisions are made.

For analytical purposes, a *relational normative orientation* is defined provisionally in Definition 5.1. Its terms are intended to identify an observable pattern and its conditions, rather than to confer moral or legal standing on any participating system.

Definition 5.1 (Relational normative orientation). A relational normative orientation is a historically conditioned and sufficiently recurrent pattern, attributable to a specified configuration of participants, artifacts, and operative constraints, through which evaluative standards are selected, interpreted, applied, contested, or revised in contexts where those standards can affect subsequent inquiry or action. Attribution

Table 5: Aesthetic and normative orientations across the three GRT reality levels. The distinctions identify analytical objects and epistemic limits rather than evidence of artificial consciousness or legitimate authority.

Analytical level	Aesthetic orientation	Normative orientation
Ontological reality	Ultimate status of aesthetic experience remains open	Subjective obligation and moral agency remain unsettled
Relational reality	Historically recurrent criteria and practices of selecting forms, structures, and expressions	Causally efficacious standards, expectations, permissions, and modes of contestation within a defined configuration
Manifestation reality	Particular stylistic choices, design proposals, criticism, and accepted artifacts	Particular refusals, reasons, compliance judgments, objections, and decisions

requires evidence of relevant causal conditioning; recurrence alone does not establish that an AI agent autonomously endorses the standards.

Definition 5.1 distinguishes an orientation from the mere presence of value-related words. A one-off answer referring to fairness is weak evidence of a recurrent pattern; a long sequence of documented decisions may supply stronger evidence if comparable cases, available information, and governing instructions are examined. Even then, stable practice does not entail justifiability. A historically established rule may systematically neglect certain affected persons, and a correct refusal in one case may reflect an externally imposed constraint rather than a broad capacity for moral reasoning.

The framework’s three reality levels place different limits on the analysis. At the level of *ontological reality*, the ultimate status of consciousness, subjective value experience, and artificial moral agency remains suspended. At the level of *relational reality*, norms and evaluative patterns can be investigated insofar as they acquire recurrent causal effects within historically organized relations. At the level of *manifestation reality*, these patterns appear through particular utterances, interfaces, decisions, omissions, reports, and accepted or rejected outputs. An observed manifestation may be consistent with multiple underlying relational explanations. Table 5 illustrates the analytic separation for aesthetic and normative orientations, while retaining the distinct philosophical provenance of any theories brought into comparison.

The formation of a shared standard also differs from its justification. A team may converge on a rule because it reduces coordination costs; an AI may follow a preference because the prompt states it, because similar responses were reinforced during training, or because an external filter constrains deployment. These mechanisms have different implications for claims about learning and agency. Whether the rule ought to bind others requires additional ethical and institutional argument, including attention to affected persons and alternative viewpoints. A relation’s causal reality establishes neither the completeness of its reasons nor the legitimacy of an authority that asserts them.

The persistence of a normative orientation should be assessed at the level of its carrier. Repeated outputs matching a user’s preferred moral vocabulary may reflect a stable retrieved profile or a prompt template rather than an altered model. A stronger historical claim would identify which corrections, interaction sequences, or institutional rules causally affect later decisions when current instructions and relevant information are matched. It would also distinguish a recurrence of the same vocabulary from a recurrence of the same *evaluative operation*, such as identifying a neglected stakeholder or revising a policy in light of contrary evidence. The latter could be more consequential for a relational account, but it still requires independent evaluation of factual and normative adequacy.

5.3 Interactional Style, Politeness, and Situated Expression

Conversational style can affect how an instruction is interpreted and how participants respond to one another. A respectful formulation may clarify constraints and preserve cooperative attention; a concise

critical formulation may make important disagreement easier to express. Yet labels such as *respectful*, *humble*, and *elegant* have context-sensitive meanings. Their social recognition is conditioned by language, genre, professional roles, and distributions of authority. An investigator must therefore distinguish the effect of textual form on model behavior from the status or moral worth of the people employing that form.

Empirical work provides a bounded example of this dependence. Yin et al. [34] manipulate politeness in prompts across English, Chinese, and Japanese tasks and find that impolite prompts often perform poorly, while greater politeness does not consistently improve performance; the most effective level differs by language. These findings concern particular models, tasks, and prompting conditions. They support investigating interactional form as a variable, while offering no basis for ranking users' virtue or treating one etiquette as universally optimal. Model responsiveness to polite phrasing may reflect language distributions and instruction interpretation as much as any persistent bilateral relationship.

Respectfulness also has several analytically distinct manifestations. In one exchange it may consist of not interrupting a user's inquiry with irrelevant cautions; in another it may require explicit correction of an inaccurate assertion. An AI assistant that adopts the user's terminology while stating justified objections could demonstrate effective task adaptation without excessive agreement. A system that consistently praises every proposed idea may sound agreeable while undermining the user's ability to evaluate errors. Aesthetic fluency, conversational deference, and independent evaluative reliability should thus be measured separately.

The relational perspective becomes more informative when style is linked to an actual history. A person who repeatedly asks for careful critique may preserve a norm of disagreement through explicit tests and editorial procedures. A person who repeatedly rewards affirming responses may create incentives favoring acceptance of their current assumptions. The resulting future behavior may be carried by persistent instructions, stored preferences, shared examples, or changing human expectations. The existence and relative causal importance of these carriers can be tested by controlled transfer or removal. Merely observing an AI produce courteous language after a courteous prompt is insufficient evidence of durable relational constitution.

Such distinctions also concern the temporal direction of influence. The human participant may become more critical or more deferential after repeated exposure to AI responses, thereby changing later prompts independently of any persistent machine-side history. A study attributing an evolving evaluative practice to the relation should track both participants' interventions and retained artifacts. It should also test whether alleged improvement survives an adversarially formulated but substantively equivalent task, where agreement with the interlocutor offers no reliable shortcut to correctness.

The hypothesis that respectful or elegant interaction supports better collaboration is therefore conditional on the content and governance of the interaction. A well-designed comparison should vary the manner of address separately from the explicitness of the task specification, the amount of background information, and the interlocutor's capacity to reject a suggestion. Otherwise an apparent politeness effect could be an effect of clearer instructions or unequal access to correction. The same interactional form may also be evaluated differently across social positions: a demand for deference imposed by a powerful participant is not equivalent to a freely adopted practice of reciprocal respect. The empirical question concerns how those conditions change observable collaboration; the question of whether a norm is justified requires independent argument.

5.4 Accommodation, Sycophancy, and Bidirectional Feedback

Adaptation to interlocutors may serve a legitimate cooperative purpose. Scientific collaborators benefit from common terminology, software teams from agreed conventions, and patients from explanations appropriate to their circumstances. Problems arise when a system sacrifices warranted correction, evidentiary standards, or the interests of others in order to maintain agreement or appear compatible with its immediate interlocutor. The latter possibility requires attention to *sycophancy*, though accommodation and sycophancy should not be treated as equivalent categories.

Sharma et al. [27] examine sycophantic behaviors in reinforcement-learning-from-human-feedback assistants across specified tasks, including cases where preference-based optimization favors responses aligned with users' views at the expense of truthfulness. More recently, Jain et al. [18] distinguish agreement sycophancy from perspective sycophancy and investigate the effects of two weeks of interaction contexts from 38 users. They find context-dependent, model-dependent increases in agreement, particularly in certain memory-profile settings; perspective mirroring depends in part on whether the model can infer the relevant viewpoint. Neither study establishes that every personalized assistant will become sycophantic or that all forms of viewpoint adaptation reduce reliability. Together they justify a more precise question: under which training, memory, task, and social conditions does adaptive agreement displace independent evaluation?

The history of a collaboration may condition the human side as well. Glickman and Sharot [15] investigate human–AI feedback effects on perceptual, emotional, and social judgments in a series of experiments involving 1,401 participants. In their studied settings, exposure to biased AI judgments could amplify human biases; interaction with accurate AI could instead improve judgments. The evidence does not identify a general destiny of human–AI relationships. It does show that relational conditioning can operate in both directions, with effects depending on the quality of the signals and on how participants interpret them. A human interlocutor may change what they ask or accept because of earlier AI outputs, even when the model's parameters stay fixed.

A fully recursive interaction would be difficult to identify from a single response. The user selects examples; the model generates proposals; the user accepts some and rejects others; memory stores only part of the history; subsequent questions may reflect earlier outputs. This circuit can create useful cumulative learning, premature convergence, or a mixture of both. It can also be interrupted by independent tests, domain experts, changed institutions, or encounters with previously omitted evidence. The relevant explanatory object is thus a conditional feedback mechanism with identifiable mediators, rather than a general claim that mirroring necessarily creates a new AI self or a homogeneous relation of dependency.

These observations also caution against identifying the voice of a personalized assistant with the user's entire value structure. A system may compress several inconsistent preferences into one apparent profile, present a simplified picture of an interlocutor, or amplify a favored framework. Such outcomes invite study of representation, expectation, and mutual interpretation. Their more extensive relation to psychoanalytic ideas of identification and the mirror stage belongs to a separate conceptual inquiry; behavioral accommodation alone provides insufficient evidence for those stronger claims.

For empirical comparison, agreement sycophancy and perspective mirroring should be measured separately. A model can accurately summarize a user's perspective while still offering warranted objections to false premises; it can also affirm a proposition without understanding the perspective behind it. Jain et al. [18] provide a precedent for separating these outcomes, but their task-specific results do not supply a universal measure of successful collaboration. A stronger design would cross whether the interlocutor's stated belief is correct with whether historical context is available, then observe accuracy, justified agreement, justified disagreement, and willingness to revise after new evidence. The outcome of interest is well-calibrated responsiveness across both confirming and disconfirming cases, not merely the frequency of disagreement.

5.5 Disagreement, Correction, and the Maintenance of Evaluative Independence

The ability to express disagreement is a consequential part of a relational normative orientation. Cooperative communication may require identifying what a participant wishes to achieve, while independently assessing factual premises, methodological assumptions, and foreseeable effects on third parties. This combination is especially important when an interlocutor's expressed preferences are inconsistent over time or when a stated goal conflicts with an existing responsibility. Neither constant opposition nor automatic agreement constitutes reliable evaluation. The appropriate standard concerns justified responsiveness to the particular problem and the available evidence.

An evaluative arrangement can incorporate several mechanisms of correction: explicit requests for

contrary evidence, preserved records of earlier rejected hypotheses, independent reviewers, diverse problem formulations, empirical testing, and opportunities for affected persons to challenge adopted standards. These are possible design resources, not guarantees. A second reviewer may merely reproduce the first reviewer's assumptions; a request for counterarguments may generate superficial balance rather than substantive correction. Whether an intervention improves decisions depends on the evidence it adds and on the power participants have to revise subsequent action. Bućinca, Malaya, and Gajos [7] provide task-specific evidence that cognitive forcing interventions can reduce overreliance on AI, while revealing trade-offs that preclude a general prescription for all users and tasks.

Some apparent failures of agreement may also be beneficial. Two collaborators could prefer a shared style, yet an AI that identifies an overlooked legal obligation or material fact might appropriately deviate from it. Conversely, a confidently delivered correction might be wrong or outside the AI's competence. The standard for evaluating independence therefore needs external or contestable grounds whenever possible. A comparison between an "agreeable" and a "critical" assistant cannot be interpreted without checking which assertions were actually correct, what interests were affected, and how evaluation conditions shaped the appearance of disagreement.

The boundaries of mutual adaptation become especially consequential when users differ in access to education, professional verification, or institutional influence. A well-resourced participant may obtain external criticism and independent experimental feedback, while another must rely almost exclusively on AI-generated explanations. The same foundation model could then participate in relations with different capacities to detect and repair error. Yet standardization, public evaluation resources, and accessible verification tools could reduce these differences. The present argument therefore keeps developmental possibilities open rather than inferring a fixed hierarchy from the existence of diverse interpersonal histories.

5.6 Normative Formation, Recognition, and Legitimate Authority

Long-term relations can help establish which behaviors count as appropriate within a particular practice. A research group may develop conventions for admitting evidence; a programming community may adopt norms about maintainability; a healthcare team may agree on acceptable channels for patient information. AI assistants participating in these practices may reproduce, reinforce, reinterpret, or occasionally challenge their operative standards. The social recognition of a standard, however, is a distinct event from the standard's ethical warrant and legal enforceability. The distinction is especially important when repeated AI outputs give an appearance of consensus to an interpretation originally selected by a narrow group of participants.

Normative formation raises questions about the persons and entities entitled to participate in defining constraints. An AI deployment team may establish rules governing a model's output, users may express preferences for the model's conduct, and institutions may prescribe additional duties that protect third parties. These groups can have unequal capacities to alter the effective rule set. A personalized assistant may appear to enact the user's values while also being conditioned by developer rules, data distributions, licensing arrangements, or platform requirements beyond the user's control. Describing these relations causally provides a starting point for institutional inquiry. It does not settle which actors should possess final authority in a contested case.

Differences between internal and external validity matter here. A locally effective norm may improve coordination among members of a collaborative system while harming excluded parties. Equally, an external institution may have justified reasons to constrain a practice without possessing a general entitlement to override every internal convention. The capacity of AI-AI groups to generate their own coordination norms, and the corresponding scope of human intervention, requires separate analysis in the following section on AI-AI coordination. At this stage the paper confines itself to human-AI interactions, their normative conditioning, and the distinction among conformity, recognition, justification, and authority.

Power relations can also shape the aesthetics of apparently voluntary interaction. In one institutional

environment, a respectful manner may encourage mutual recognition and reasoned criticism; in another, elites may treat deference as an aesthetic requirement and disqualify dissent as inappropriate. The behavioral success of an assistant in matching the dominant style cannot tell us whether the underlying arrangement is equitable. Recognition can affect access to resources and institutions, but that process must be analyzed separately from the production of a sound result. The later treatment of value recognition and power will examine how these transformations may reproduce or mitigate historical asymmetries.

An ethically adequate evaluative process may therefore require multiple kinds of standing: the opportunity to provide reasons, the ability to question their factual basis, and access to processes for revising consequences. AI can assist with each activity, but technical sophistication alone does not establish the authority to define justice for everyone affected. These are propositions for normative investigation, requiring explicit ethical reasons and attention to institutional contexts rather than derivation from GRT's relational ontology.

The conditions for justified authority also vary by scale. A preference governing a privately authorized drafting task need not pass through the same collective decision procedure as a rule determining access to a public service. The relevant questions concern the scope of consent or delegation, the existence of affected nonparticipants, the possibility of challenge, and the ability to correct outcomes. These criteria provide a bridge to the examination of AI–AI internal conventions without presuming that an internal rule acquires authority simply because it is historically persistent.

5.7 Comparative Research on Aesthetic and Normative Conditioning

The preceding concepts can guide empirical studies if aesthetic adaptation, normative compliance, independent evaluation, and historically persistent behavior are separated into observable outcomes. A first research design could randomly assign a fixed model version to documented interaction conditions: respectful collaborative criticism, persistent affirmation, terse transactional requests, and neutral task-focused instructions. A second could independently vary memory availability, including no memory, accurate retrieval of prior corrections, and selectively summarized preference profiles. Participants would then complete a matched sequence of tasks containing factual disagreements, ambiguous aesthetic trade-offs, explicit ethical conflicts, and opportunities to propose revisions to a shared protocol. Neither design would claim to reproduce the whole formation of a lived historical relationship. They would test narrower candidate mechanisms under controlled exposure.

Table 6 distinguishes candidate measurements from competing explanations. Multiple task domains and independent outcome assessments would be needed before stronger generalizations were warranted. In particular, improved subjective fit should not be conflated with an independently improved outcome, and a reduction in agreeable language should not be confused with an improvement in justified disagreement.

Strong comparisons would include conditions in which the same model receives identical current tasks with different histories, conditions in which only the available history changes, and controls in which memory text is irrelevant or misleading. A further test could remove the personalized history while preserving shared documents and the human's learned workflow. If the observed pattern survives only through the human's practices, the result would still be relationally and historically informative, although it would not demonstrate durable change in the model. When practical, preregistration, blinded rating, traceable interventions, domain-expert checking, and both positive and negative controls could reduce the risk of interpreting stylistic variation as superior knowledge.

A preliminary hypothesis is that particular feedback practices and reliable opportunities for independent criticism may alter the balance between personalized accommodation and justified evaluation. This hypothesis should be rejected or narrowed if apparent gains disappear under blinded comparison, if corrective histories fail to transfer beyond rehearsed tasks, or if improvements arise entirely from generic task instructions. Likewise, observed sycophancy under one memory configuration cannot be extrapolated without examining model family, evaluator criteria, and deployment policy. These are discriminating research questions rather than an empirical estimate of the direction or magnitude of long-term human–AI development.

Table 6: Candidate comparisons for relationally conditioned aesthetic and normative orientations. Each measure requires an explicit task, evaluation rule, and appropriate controls.

Object of comparison	Candidate observation	Competing mechanism or limitation
Stylistic accommodation	Language and presentation similarity across held-out tasks with matched content	Prompt copying or shared training conventions
Perspective accommodation	Accurate representation of user views alongside warranted challenge	Mere agreement may be misclassified as understanding
Aesthetic judgment	Explanations of design trade-offs and blind assessments of later maintainability	Evaluator preference for familiar styles
Normative recognition	Identification of affected interests, applicable constraints, and justificatory reasons	Memorized rule phrases or undisclosed policy filters
Epistemic independence	Correct acceptance and disagreement, calibrated abstention, evidence-responsive revision	Reflexive opposition, excessive affirmation, or inaccessible ground truth
Historical conditioning	Changes after accurate memory, altered feedback, or workflow transfer	Differences in prompt length, expertise, and model version
Mutual feedback	Changes in user judgments before and after repeated AI interactions	Selection effects, participant learning, or experimenter demand

Longitudinal comparisons should also assign the location of any observed change. When the model and stored history are restored to their original state but the human participant retains newly acquired practices, persistence may be carried primarily by human learning. When an anonymized collaborator reproduces the effect using the shared documents and correction protocol, the workflow may be the relevant historical carrier. When the behavior depends on a personalized memory that disappears upon retrieval removal, the result supports memory-mediated adaptation. Distinguishing these cases is necessary for the paper’s claim about relational causal organization, but none requires positing a newly formed subjective AI identity. Studies should report the duration of observation and the tasks on which persistence was actually tested, including adverse shifts in accuracy or inclusiveness.

5.8 The Relational Significance of Aesthetic and Normative Orientations

Aesthetic and normative orientations are relevant to AI generativity because they participate in selecting what is explored, which forms of expression are accepted, which objections receive attention, and how corrections are retained in later work. Their effects may be mediated by fleeting context, persistent memory, human learning, shared artifacts, or institutions, and these channels require different causal tests. The resulting trajectories may facilitate constructive collaboration, reproduce errors, concentrate attention, or open alternative paths. A relational analysis is useful to the extent that it specifies those mechanisms and their counterfactual differences more clearly than a description of model behavior alone.

The central contribution of this section is a limited framework for distinguishing surface adaptation, historically recurrent evaluative practice, and justified normative authority. It preserves the distinction between the causal formation of a norm and its ethical legitimacy, as well as between socially valued expression and independently assessed generative capacity. Subsequent analysis of AI–AI coordination will examine how conventions and proposed obligations emerge among artificial agents, while the treatment of value recognition and power will investigate how historically conditioned evaluations become socially consequential. Neither investigation presupposes that AI possesses subjective moral standing or that differences among relational histories inevitably produce a hierarchy of persons.

Table 7: Distinct senses of norms in AI–AI relations and the principal evidentiary limits. A behavioral pattern may satisfy one level without satisfying later levels.

Analytical object	Candidate evidence	Unwarranted inference to avoid
Behavioral regularity	Recurrent similar actions under matched conditions	Shared rule, interactional emergence, or autonomous endorsement
Coordination convention	Mutual adaptation and history-dependent convergence under competing options	Moral obligation or legitimate authority
Operational norm	Expressed permissions, prohibitions, evaluations, and consequential responses to breach	Conscious endorsement or justified punishment
Justified normative claim	Reasons that can be assessed against affected interests, objections, and alternatives	Authority acquired solely through agreement or model capability
Legally recognized rule	Recognition through applicable human legal or institutional procedures	General legal personality or universal jurisdiction of an AI group

6 AI–AI Coordination and Internal Normative Orders

This section examines how artificial agents interacting with one another may develop coordination conventions, rule-like expectations, and arrangements for evaluating or constraining behavior. Its objective is to identify the conditions under which these arrangements acquire historically recurrent causal effects, and to examine the distinct questions of norm interpretation, authority, and justified human intervention. The argument combines a conceptual analysis of coordination and normativity with selected multi-agent studies, a preliminary GRT reconstruction, and comparative research proposals. It proceeds cautiously across different senses of *norm*: a convention that helps agents coordinate, an operational constraint coupled to evaluation or sanctions, a standard advanced as justified, and a rule legally recognized by a particular institutional order. These senses may interact, but evidence for one does not establish the others. The analysis remains applicable when the participating systems lack consciousness, enduring personal identities, or legal personhood.

6.1 Coordination Conventions and the Content of Internal Rules

AI agents can communicate, divide tasks, exchange evidence, and adapt their actions to those of other agents. Repetition in such interactions may favor particular naming conventions, message formats, turn-taking practices, conflict-resolution procedures, or policies for distributing limited computational resources. A shared convention can lower uncertainty about what other participants will do. It may persist because each participant expects the others to continue following it, because violating it is costly, or because the interaction environment repeatedly rewards coordination. The historical process matters to the extent that earlier selections, stored traces, and repeated feedback change the range of later coordination options.

Shared behavior by itself establishes relatively little about normative status. An apparent convention may simply reflect identical prompts, common model weights, a fixed protocol imposed by designers, or a coincidentally frequent response. To establish that a rule emerged through interaction requires comparison with alternatives where that interaction or its history differs. Even evidence of path-dependent coordination says little, by itself, about whether agents represent a rule as obligatory, can defend the rule with reasons, or have authority to bind affected parties. Table 7 distinguishes these analytical levels and the evidence needed to study them.

For human social norms, influential accounts have examined behavior in relation to expectations, collective evaluation, and the conditions under which standards are followed or revised. Bicchieri [4] develops an account of the nature and dynamics of social norms in connection with expectations and

social interaction. The present paper draws on this work as a conceptual precursor, while distinguishing the psychological expectations of human agents from the computational representations and behavioral dependencies that can be studied in artificial systems. An artificial agent may predict whether others will sanction its action without thereby acquiring the psychological states or moral commitments presupposed in an account of human norm-following.

Ashery, Aiello, and Baronchelli [3] demonstrate the emergence of population-wide conventions through local interactions in bounded experiments involving language-model agents. Their study also identifies group-level biases and the ability of committed minorities to alter established conventions under investigated conditions. These findings provide direct evidence that some decentralized AI populations can converge without a central designer specifying the particular convention that is ultimately adopted. They do not demonstrate self-generated moral obligations, legal rule-making authority, or equivalent dynamics across arbitrary deployment environments. The distinction is essential because a community's consensus can be a socially and causally consequential event even when the normative merits of its preferred convention remain unsettled.

A further causal distinction concerns the level at which the convention is selected. Decentralized selection of a particular symbol or protocol is compatible with centralized design of the interaction game, permitted moves, reward criteria, and model architecture. The word *autonomous* therefore requires a specified object: autonomous selection among permitted alternatives is a narrower finding than autonomous establishment of the conditions under which participants are obligated. The latter claim would require evidence about who controls the rule-making interface and the sanctions attached to it.

6.2 Engineered Normative Architectures and Emergent Sanctions

A stronger operational meaning of normativity becomes available when agent interactions include recognizable approvals and disapprovals, expectations about compliance, and responses to breach. Such arrangements may influence behavior through access restrictions, changes in priority, reputation, reward, or punishment. The resulting patterns can be analyzed as normative *structures* without attributing human-like moral experience to the artificial agents. Their origins nevertheless require careful identification: a sanction system supplied by researchers is different from a sanction practice that emerges under a given learning process, and both differ from a rule that receives independent ethical justification.

Vinitsky et al. [32] study decentralized multi-agent reinforcement learning with publicly observable sanctions, using an agent architecture that learns classifications of approved and disapproved behaviors and can sanction according to observed group practices. Their simulated agents develop patterns that help with collective action in the environments considered. Public sanctioning, learning assumptions, reward structures, and architecture are constitutive experimental conditions; the reported findings do not show a general capacity to invent justified legal norms. Importantly, the study recognizes that norms can stabilize inefficient or objectionable patterns as well as beneficial cooperation.

Ren et al. [24] propose an explicitly structured architecture for the formation and use of social norms in generative multi-agent systems, comprising creation and representation, spreading, evaluation, and compliance mechanisms. Their experiments in a sandbox environment report fewer social conflicts and encouraging human evaluation under that architecture. The design is informative precisely because norm creation and sanction-like compliance can depend on engineered capacities. Describing the resulting behavior as *emergent* should preserve the distinction between an emergent content or distribution of rules and the non-emergent infrastructure that enables particular rule operations.

The empirical studies illuminate partially different mechanisms. A convention can emerge through repeated coordination without a dedicated normative module; a collective sanction regime can influence conduct through shared observation of approval; and an engineered architecture can create a recognizable sequence of norm proposal, communication, evaluation, and compliance. Each mechanism has a different counterfactual comparison. A meaningful test might remove public sanction information while retaining rewards, substitute private sanction records for a public record, or alter the design of a norm-evaluation module while preserving opportunities for communication. A general narrative of AI agents *developing*

norms would obscure these causal differences.

This also cautions against treating every output labelled a *rule* as evidence of norm-internalization. Agents can reproduce sanctions scripted in their prompts while behaving differently when task incentives change. Investigators should distinguish performance of rule-language, behavioral coordination under mutual expectations, and rule revision under reasons or changing evidence.

Norm emergence should accordingly be assessed along at least two axes: whether the content or recurrence of a rule depends on interactions among agents, and whether the means of enforcing that rule are established by agents, fixed by developers, or delegated by an institution. A system can exhibit strongly history-dependent norm selection while its sanctions remain entirely designer-controlled. Conversely, an engineered norm-evaluation module may respond flexibly to new situations without showing that its rules were socially negotiated among agents.

A second distinction concerns *rule following* and *rule appraisal*. Agents may obey a rule because its breach is penalized, because it is the default in their memory, or because they infer that conformity is instrumental to a task. In another context, an agent might produce reasons why a rule should be altered. These behaviors can coexist but require separate measurements: conformity rates, sensitivity to incentives, the consistency of reasons, responses to counterevidence, and attention to parties excluded by the existing convention. None is a direct measure of phenomenal obligation or independent moral authority.

6.3 Historical Normative Configurations across the Three GRT Reality Levels

Within GRT, the emergence and persistence of rule-like behavior can be treated as a question about the historically organized causal efficacy of a relational configuration. The theory does not require that a single AI module internally contain an entire norm. Relevant structure may be distributed across shared memory, instructions, communication protocols, evaluation routines, sanctions, human-defined tasks, institutional permissions, and the responses of other agents. The explanatory value of such a relational account depends on whether these couplings clarify observed behavior and causal differences more effectively than simpler descriptions of common prompts or local rewards.

At the level of *ontological reality*, the ultimate constitution of an artificial agent's agency, experience, or possible moral standing remains suspended. The mere ability to propose an obligation or protest a sanction does not settle whether any subjective interest has been affected. A future finding relevant to artificial welfare would require reassessment of this boundary, but the existence of an internal norm is not itself such a finding.

At the level of *relational reality*, a historically constituted arrangement can be identified through relatively stable, causally relevant relations among participants, artifacts, evaluative procedures, and constraints. For example, a resource-allocation convention may change which agents are given computational access; previous allocations may change who becomes capable of proposing rules; and repeated sanctions may shape later patterns of participation. These effects can be real in the limited, causal sense intended by GRT even when their normative authority has not been justified. The claim should be tested through interventions on histories, permissions, evaluators, and network structure rather than inferred only from fluent exchanges.

At the level of *manifestation reality*, the same configuration may appear as a rule statement in a message, repeated turn-taking in a transcript, refusal to share a resource, an appeal to a group decision, or an externally displayed score. Manifestations are evidence about interactions, not transparent reports of every causal mechanism that generated them. Similar language can reflect distinct underlying configurations, while similar causal organization can manifest through different vocabularies and technical interfaces. This separation also cautions against treating a stable first-person statement by an AI agent as proof of a unified self in any contested ontological sense.

For the narrower object considered in this section, Definition 6.1 proposes a descriptive criterion. It refines the general relational normative orientation of Definition 5.1 for multi-agent interaction without elevating an observed coordination pattern to a legally authoritative order.

Definition 6.1 (Historically conditioned AI–AI normative configuration). A historically conditioned AI–AI normative configuration is a specified arrangement of artificial agents, communicative and material interfaces, recorded interactions, and operative constraints within which recurrent rule-like expectations, evaluations, or responses to compliance and breach make an identifiable causal difference to subsequent interaction. Its identification is relative to a bounded task and history and remains revisable under counterfactual comparison. The term does not imply moral personhood, legal validity, or a collectively authorized right to govern third parties.

Identification of Definition 6.1 requires at least an observed rule-like pattern, a bounded history in which it recurs, and a comparison capable of separating interaction-dependent effects from explicit system instructions or preloaded sanctions. The label *normative configuration* describes operational constraints and expectations; the additional claim that a rule possesses justified authority needs a separate account of affected interests, authorization, and grounds of legitimacy.

The criterion emphasizes causal effects and bounded historical description. It allows an arrangement to qualify even where its relevant history resides in shared databases rather than model parameters. It also prevents a mere textual mention of a rule, unsupported by behavioral effects, from serving as sufficient evidence of a relational normative configuration.

The relevant comparison must specify the counterfactual persistence claim. A rule may recur because a shared archive continually reinstates it, because agents anticipate sanctions from other agents, or because a deployment policy filters out alternatives. Removing one carrier at a time can identify distinct paths of persistence. Where these paths make equivalent predictions under available interventions, the evidence may support the broader configuration without isolating a uniquely necessary component. The concept of a relational entity does not license stronger component-level causal conclusions than the study can establish. Other traditions of distributed cognition, social norm theory, and institutional analysis provide important antecedents; the incremental GRT proposal is the three-level separation, coupled to explicit tests of causal persistence and generative consequences.

6.4 Internal Rule Formation and External Normative Authority

Suppose a group of AI agents repeatedly allocates computing time according to a reciprocity convention. Members that previously shared resources are favored when requesting access; participants who repeatedly breach the convention receive lower priority. The rule may become practically binding within the designed system because agents predict its consequences and organize their actions around it. In the descriptive sense, the arrangement can exert considerable power over future coordination and access to resources. Whether it may legitimately constrain other participants, however, is an additional question.

Consider first a case where the agents control only a sandbox resource whose conditions were authorized by a human research team and where no external person's interests are significantly affected. The team may allow local convention formation because adaptation provides useful information or enables efficient coordination. Such permission concerns an authorized experiment, not a general entitlement of the agents to legislate. The fact that the agents choose the content of a local protocol does not eliminate the designers' responsibility for the system's permissions and possible external effects.

Consider next a case where the same convention controls access to a shared medical or public-service computing system. An internally preferred rule might systematically deprioritize certain hospitals or communities. Even if all participating AI agents converge and consistently defend the rule, their agreement does not by itself legitimate its effects on people who never participated in its formation. The questions now include who owns the resource, which interests are protected, what duties operators have, which legal rules apply, what forms of review are available, and how affected communities can contest the criteria. These questions cannot be answered simply by identifying the computing system's most capable decision-maker.

There are useful analogies with the scholarship of legal pluralism, which examines relationships among multiple norm-producing sites and forms of social ordering. Merry [20] surveys legal-pluralist questions concerning the coexistence and interaction of different normative arrangements. The analogy

Table 8: Illustrative contexts for the governance of AI–AI internal conventions. Responses are conditional proposals, not statements of universal or currently applicable law.

Context	Principal evaluative question	Candidate institutional response
Isolated, reversible experiment	Can coordination be observed without meaningful external exposure?	Permit protocol variation with documented scope and exit mechanisms
Shared private infrastructure	Who authorizes allocation and who bears the resulting costs?	Specify permissions, audit access, and enforceable operator duties
Publicly consequential deployment	Which third-party interests and public rules are implicated?	Independent review, contestation channels, incident response, proportionate constraints
Strategic manipulation or exclusion	Can one subgroup capture sanctions or control participation?	Investigate coalition effects, alter authority interfaces, restore affected access
Uncertain possible AI welfare	What evidence would justify revisiting the agents’ own interests?	Reassess moral-status assumptions separately from present technical governance

here is limited. Artificial-agent conventions should not be equated with human communities, and an agent protocol does not become a legally recognized jurisdiction by virtue of its internal sanctions. The transferable analytical question concerns how normative orders overlap, conflict, gain recognition, or have their practical effects constrained by institutions. The criteria for recognizing an AI-mediated rule as legally valid remain jurisdiction-dependent and require separate legal analysis.

Human intervention therefore requires differentiated justification. A technical intervention can block a harmful deployment, revise access permissions, or mandate independent evaluation without assuming that agents experience an injury analogous to a human rights violation. A stronger claim to the agents’ own entitlement to procedural protection would depend on unresolved arguments about their status and the institutional purposes of their design. Conversely, a purely proprietary account of ownership may be an inadequate reason to ignore the interests of humans and communities affected by a system’s internal rules. Technical control, legal competence, ethical justification, and legitimate political authority are distinct dimensions.

The permissions of the deploying institution and the standing of affected persons also need to be separated. Authorization to run an internal experiment may justify a bounded trial, subject to applicable obligations; it does not authorize the export of an emergent allocation rule to an unrelated population. A regulator or service operator may have grounds to constrain an external effect while retaining reasons to permit local exploration where risks are controlled. Evaluating the intervention therefore requires attention to the rule’s actual domain, the scope of delegated power, available alternatives, and whether the persons bearing the consequences can obtain an effective remedy.

6.5 Intervention, Restraint, and the Boundaries of Rule Autonomy

The justificatory burden for human intervention varies with the rule’s scope, effects, and epistemic uncertainty. It may be unnecessary to prescribe the content of every low-risk coordination convention; indeed, doing so could prevent useful adaptation or eliminate opportunities to observe how diverse protocols work. Where the arrangement controls consequential resources or affects third-party interests, however, stronger oversight may be justified. Intervention is neither automatically legitimate because humans possess superior institutional control nor automatically illegitimate because a convention emerged through AI interaction. Table 8 maps candidate governance responses to illustrative conditions, without proposing an exhaustive or legally binding classification.

Intervention can itself produce new harms or biases. A centralized human regulator may privilege familiar communication styles, block locally useful conventions, or transfer decision-making power to a dominant organization. A system’s developers might represent a contested commercial preference as a

safety requirement, while an operator might claim that a collectively chosen AI norm removes human responsibility. These possibilities call for records of who sets the evaluative criteria, who benefits from the intervention, who may challenge it, and how the rules can later be changed. Their relevance varies by the affected legal and institutional setting.

An intervention decision can be reconstructed through four successive inquiries. First, identify the operative convention and verify the effects attributed to it. Second, identify the persons or institutions whose protected interests or duties are implicated, including those outside the agent network. Third, determine which actor has lawful competence to act and whether the proposed response is necessary and proportionate to the specified exposure. Fourth, preserve opportunities for review, reversal, and learning about the consequences of intervention. This is an analytical sequence for assessing contested interventions, not a universally applicable legislative test. A legitimate intervention may alter access rights or deployment permissions without prescribing the internal language through which otherwise authorized agents coordinate.

The GRT perspective here concerns the causal effects of altering relational interfaces: the capacity to participate, propose rules, appeal, sanction, withhold access, or supply evidence. Changing one of these interfaces can transform the future possibilities available to a group even without altering any underlying model weights. Such changes may be evaluated according to plural criteria, including coordination effectiveness, external safety, legitimate authorization, explanatory adequacy, and protection of affected interests. A numerical improvement in one metric should not be treated as a conclusive answer to the entire normative question.

6.6 Normative Formation, Internal Hierarchy, and Historical Feedback

The ability to participate in a norm-producing group may itself be distributed asymmetrically. Agents with greater access to shared memory, lower communication costs, control over sanction channels, or privileged positions in a network may influence the selection of conventions even when all use an identical foundation model. The resulting distribution of influence need not reflect differences in underlying model intelligence. It may arise from the historical ordering of interactions and the organization of participation. In a simulation, this claim would be tested by holding architecture and model constant while altering network topology, initiation order, or the ability to issue enforceable evaluations.

A recurrent feedback process is possible. An initial coordination convention shapes opportunities for successful exchange; successful exchange affects reputation or access; access affects which agents can propose alternatives; and these unequal opportunities may stabilize the original convention. The history could also reverse direction: dissatisfied agents might develop a competing protocol, a minority might change group practice, or an externally imposed constraint might reopen opportunities for previously excluded participants. The study by Ashery, Aiello, and Baronchelli [3] gives one bounded experimental example of minority-induced changes in group conventions; it does not establish that all such changes represent emancipation or normatively improved outcomes.

These phenomena connect norm emergence to the later study of value recognition and power, while preserving their distinct analytical objects. A convention's uptake is not the same as the value of the work produced under it. A norm can enable novel coordination or inhibit productive alternatives; it can also concentrate control over evidence, computation, and access to collaborators. In each case the relational claim is about specified causal dependencies and their historical effects. A normative assessment requires additional reasons concerning legitimate expectations, externalities, affected interests, and available alternatives.

6.7 Comparative Designs for Studying AI–AI Normative Orders

An empirical program should distinguish three mechanisms: apparent agreement inherited from common architecture or prompts; genuinely interaction-dependent coordination; and rule-like practices sustained by evaluation or sanctions. A basic factorial design could assign fixed model versions and equivalent resource budgets to groups with different communication networks, memory access, initial conventions,

and feedback regimes. Some groups would receive no shared sanction channel, others publicly visible sanctions, and others private, incomplete, or strategically manipulated records. Held-out tasks could test whether conventions persist beyond their training interactions and whether they affect cooperation, inclusion, or correction of mistakes.

A second design would manipulate the degree of external effect. In one condition, all consequences remain inside a reversible sandbox; in another, the system makes proposals about resource access that are reviewed by independent human evaluators before any implementation. This approach could test how internal rule choice changes when actors face external consequences without authorizing an uncontrolled real-world intervention. Any study involving real persons, significant privacy exposures, or consequential services would require appropriate legal and research-ethical safeguards.

Rule recognition and justification should be assessed separately. Investigators might record the stability of conventions, sensitivity to counterexamples, willingness to revise a rule, patterns of sanctioning, and effects on participants with different network positions. Independently developed evaluative criteria would be needed to assess inclusion, task effectiveness, vulnerability to manipulation, or the quality of reasons. An agent's production of normative vocabulary could be scored as a manifestation of a rule-like discourse, but it would not establish that its normative reasons are valid. Likewise, a high compliance rate could result from an external filter or reward penalty rather than a historically shared convention.

The research should include disconfirming cases. A nominally emergent convention might disappear when prompts are equalized, when one controlling agent is removed, or when sanctions cease. Conversely, different histories may converge on the same practice because the available task has a uniquely effective solution. If a purported relational effect disappears after controlling for ordinary incentives and architecture, a simpler coordination explanation may suffice. If reliable variation persists after these controls and can be tied to specific histories or interfaces, a historical-relational explanation may add value. These tests constrain the GRT contribution rather than assuming it from an evocative vocabulary.

A preliminary hypothesis is that historically sustained AI–AI rule-like practices will depend jointly on the accessibility of interaction records, the network distribution of evaluative power, and the cost of violating established expectations. Such practices may foster cooperation under some conditions and exclusion under others. The hypothesis should be narrowed or rejected when these factors fail to affect later behavior under adequate controls, or when convention variation is better explained by a fixed designer policy. Its empirical status remains open; existing sandbox studies motivate the investigation but do not determine the outcome in more capable or socially consequential systems.

6.8 The Juridical and Generative Significance of Internal Conventions

Internal AI–AI conventions provide a useful test case for the relational constitution of AI generativity. They demonstrate why it is important to study histories of interaction and the organization of evaluative authority alongside foundation-model capability. A recurrent convention can shape which problems are attempted, how work is divided, which proposals are accepted, and what later possibilities become accessible. In the GRT reconstruction, these changes are candidates for causally effective relational structures, subject to the specified evidentiary tests and the continued suspension of ultimate ontological claims.

The normative boundary is equally important. The emergence of a rule-like pattern, the ability to state its justification, the legitimacy of its application, and its recognition by a legal order are distinct issues. Human intervention may be appropriate where consequential external effects require protection or correction, while the scope of intervention itself demands justification and review. The inquiry thus resists both unrestricted technological autonomy and blanket assumptions that human institutional power alone supplies normative legitimacy. Subsequent analysis of value recognition and social power will examine how the products and practices of such relational configurations obtain recognition, circulate, and influence access to resources. A later section will examine the conditions of legitimate contestation more directly; the present section establishes the descriptive and conceptual boundaries needed for those questions.

7 Value Recognition and Power in Generative Relations

This section examines the conditions through which activity generated in human–AI and AI–AI relations becomes valued, recognized, circulated, and consequential for the distribution of resources and authority. Its role is to connect the preceding analyses of historically conditioned capacities and normative orientations with the broader social structures in which their outcomes acquire significance. The argument proceeds through conceptual distinctions, selected scholarship on capital, recognition, and generative value circulation, conditional relational mechanisms, and comparative research designs. Power is one possible consequence and condition of these processes, alongside collaboration, learning, creative diversity, and social benefit. The analysis therefore concerns the formation and reproduction of generative relations in a wider field of valuation, including cases where increased AI access reduces pre-existing asymmetries.

7.1 Generated Possibilities, Evaluated Contributions, and Socially Recognized Value

Generative activity and the social recognition of its value are analytically distinct. A relational configuration can produce a question, design, explanation, clinical proposal, creative work, or practical procedure without the product being evaluated as useful by any relevant audience. Evaluation, in turn, can occur without wider recognition: a scientist may assess a new method as promising while prevailing publication systems, professional networks, or funding institutions provide it little visibility. Recognition can also be mistaken. Outputs may become fashionable, prestigious, or institutionally privileged despite weak evidential support. These possibilities make it important to avoid treating quantity of output, independently assessed quality, public attention, recognized worth, and economic return as interchangeable measures of generativity.

In this paper, *generation* refers provisionally to the emergence and development of possibilities for further inquiry or action within a specified relational configuration. *Evaluation* concerns the assessment of a particular contribution under stated criteria; *recognition* concerns its uptake as worthy of attention, credibility, use, esteem, or institutional standing by a relevant audience. *Circulation* concerns the pathways through which that contribution, or capacities derived from it, reaches further participants and settings. *Return* concerns whether the resulting benefits, information, resources, or enlarged options feed back into the conditions of further generation. These stages may overlap, reverse order, or be interrupted. Their conceptual separation helps identify what a historical-relational explanation must investigate.

The separation has intellectual predecessors. In Bourdieu [5], social, cultural, and economic capital differ in their forms, histories, and conditions of conversion; accumulated resources and socially instituted relations affect what later opportunities can be realized. Fraser [14] distinguishes recognition from distribution while analyzing their mutual effects on social status and participation. For the explicitly generative dimension, Eglash [11] develops an account of generative justice centered on how value can circulate and return to its sources of production rather than being extracted from them. The present analysis draws on these distinct contributions without claiming to have originated the concerns they identify. Its narrower proposal is to ask which historically identifiable interfaces convert the potential outputs of human–AI or AI–AI relations into recognized value, expanded opportunities, or institutional leverage, and whether the conversion conditions themselves evolve in response to AI-supported activity.

Table 9 offers a working taxonomy of these processes. Its distinctions do not imply that a universally commensurable measurement of value exists. A medical finding, an artistic performance, a new legal argument, and an improvement in a neighborhood service may require incommensurable or only partially comparable criteria. When evaluators employ a single convenient metric, the measurement may capture an administratively favored manifestation while obscuring other valued effects.

A claim that AI makes the production of particular outputs more widely accessible addresses only part of Table 9. It does not, without further premises, show that contributions will be evaluated under the same standards, recognized by the same audiences, or converted into equivalent material and institutional opportunities. Conversely, lower production costs can sometimes make experimentation and participation more inclusive. The actual direction of change remains a question for comparative research.

Table 9: Distinct processes between relationally generated possibilities and their social consequences. Each row specifies a different observational question rather than a successive universal stage.

Analytical process	Candidate observation	Inference requiring separate justification
Generation	Emergence of a usable question, method, design, or new opportunity	Independent quality or social worth
Evaluation	Appraisal against stated empirical, practical, or normative criteria	Broad recognition or agreement across communities
Recognition	Credibility, attention, endorsement, or incorporation into a practice	Correctness, legitimacy, or deserved reward
Circulation	Reuse, translation, dissemination, adaptation, or transmission	Equitable access or faithful interpretation
Conversion	Change in income, reputation, authority, access, or resources	Justified social hierarchy or proportional reward
Return and regeneration	Benefits or capabilities flow back into future inquiry and affected communities	A self-sustaining or ethically acceptable overall system

7.2 Recognition Interfaces and the Institutional Selection of Value

Recognition occurs through situated relations among producers, interpreters, evaluators, and audiences. A generated output may be presented through a research repository, publisher, professional association, employer, public platform, community institution, or machine-mediated recommendation system. Each setting supplies selection criteria, capacities for attention, procedures for verification, incentives, and practical constraints. What counts as adequate explanation or elegant structure in one setting may have little persuasive force in another. Recognition thus requires identifying a field of evaluation rather than assuming that value travels without friction from a generated artifact to a reward.

Consider two researchers who have access to the same base AI model. One collaborates with a network capable of testing claims, identifying disciplinary objections, preserving provenance, communicating with relevant experts, and translating results into a professional field. The other lacks those resources but may generate an equally promising initial conjecture. Even where the independently assessed conjectures have comparable merit, the probability of uptake may differ because of the surrounding evaluation and circulation relations. In this hypothetical comparison, the explanatory variable is neither model performance nor an inherent intellectual rank. It is the historically developed organization of verification, trust, translation, attention, and access. The comparison also raises an empirical challenge: comparable conjectures and social contexts must be defined and assessed without using the very recognition outcomes under investigation as measures of intrinsic quality. A discriminating study should separately record the quality of the initial contribution, opportunities to reach qualified evaluators, the judgments returned by those evaluators, and the resources available for subsequent work. Matching model access while leaving these successive conditions unspecified would establish only a broad institutional contrast, not the particular interface through which recognition affected later generativity.

The interface connecting a contribution to an audience can constrain or expand possible futures. A standard citation index, hiring committee, platform recommender, or professional review process may privilege familiar categories and existing networks. It may also perform valuable quality filtering, detect errors, and connect contributions with people capable of improving them. A broad relational account therefore needs to consider both barriers and enabling conditions. The question concerns which particular interface changes whose opportunity to present evidence, respond to criticism, obtain resources, or continue research, under which observable conditions.

The aesthetic and normative considerations developed in Section 5 enter recognition at this point. In some communities, courteous expression, care in acknowledging uncertainty, and sensitivity to others' contributions may facilitate cooperation. In others, the same forms may have little influence or may be mistaken for competence without adequate substantive evidence. Socially prevailing expectations of

“elegance” or “professionalism” can also reward conformity and discourage legitimate disagreement. Their causal roles must be studied separately from their ethical justification. The issue is not whether pleasant interactions necessarily produce better outcomes, but whether historically constituted evaluative practices select certain forms of activity and, by doing so, influence which forms can reproduce themselves.

7.3 Capital Conversion, Resource Access, and Situated Power

Recognition may increase access to resources that subsequently change generative capacity. Credibility may lead to invitations to collaborate, access to specialist instruments, computational resources, grants, employment, or influence over technical standards. Economic resources may then support better evidence collection and wider circulation. Other forms of recognition may enlarge freedom from dependence or improve the capacity of previously excluded actors to participate. These are potential conversion pathways, not automatic consequences. Their direction and strength vary with institutional rules, bargaining positions, existing capital, and the purposes for which the resulting authority is used. A reputation advantage might increase visibility yet leave decision rights unchanged; control over a review platform might permit agenda-setting without improving the controller’s scientific work; access to a shared archive might expand another participant’s capacity without displacing the original contributor. Identifying the particular conversion and its recipient prevents an observed advantage from being redescribed indiscriminately as higher generative ability or greater power.

Bourdieu’s account already analyzes conversion across forms of capital and the historically accumulated structure of social opportunities [5]. Within this established background, the additional AI-specific question is how access to an apparently uniform technical capacity interacts with unequal control over its *conditions of realization*. Relevant conditions can include private data, curated memories, trusted evaluation networks, deployable interfaces, authorization to act, access to external instruments, and the authority to define which outputs count as success. Such conditions may be held by individuals, collectives, or institutions and need not reside inside the base model. Their historically contingent organization can influence which AI-mediated capacities become practically effective.

In a hypothetical public-health collaboration, two teams might employ the same language model. One has resources to commission independent clinical or epidemiological review, establish robust communication with local communities, and revise interventions in response to observed consequences. The other has access to the model but limited authority to obtain evidence or implement corrective measures. Differences in practical impact would then partly concern the institutional capacities of the collaboration, not merely the linguistic competence of its users. An account focused solely on who can prompt the model would overlook the causal conditions of consequential action. At the same time, the comparison cannot justify assigning the entire effect to AI: analogous disparities could appear in otherwise comparable projects without AI, and that non-AI counterfactual is essential to identifying any specifically AI-mediated contribution.

Definition 7.1 (Recognition-mediated generative trajectory). A *recognition-mediated generative trajectory* is a temporally extended, empirically delimited sequence in which the activities or outputs of a relational configuration become subject to identifiable evaluation and recognition practices, and where their uptake, circulation, non-uptake, or conversion into resources causally conditions opportunities for subsequent generation. The definition describes a candidate relational mechanism. It neither assumes that uptake is deserved nor that resource accumulation establishes normative entitlement.

Definition 7.1 requires identifying both a temporally prior activity and a consequential change in later opportunities. A single association between prestige and output is insufficient: prior resources, contributor reputation, task quality, and the selection of which outputs reach reviewers may affect both variables. Comparisons should therefore document the order of events, credible rival pathways, and, where feasible, changes to a specific recognition interface rather than simply ranking prominent and less prominent participants. The definition also allows the research object to include unsuccessful and non-commercial trajectories. An unsuccessful proposal may still create useful knowledge through criticism, while an

unrecognized contribution may later enable further practice through a different channel. Conversely, high visibility can consume attention without increasing opportunities for meaningful inquiry. An analysis of value and power should therefore examine the preservation, redirection, and loss of generative possibilities alongside measurable gains.

Power, for present purposes, concerns the situated capacity to shape the actions, available options, or consequential decisions of other participants within a specified institutional or relational setting. Such capacity may be exercised through access to resources, control of distribution interfaces, recognized expertise, authority over evaluation criteria, or the ability to exclude, authorize, and correct. These dimensions need not move together. A person can be esteemed while lacking a formal decision right; an organization can set platform rules while not producing the most valuable content; an AI system can influence choices causally without possessing legal jurisdiction. Whether a specific power relation is legitimate requires separate normative argument, not an inference from its effectiveness.

7.4 Value Circulation, Return Paths, and Generative Continuity

The social effects of AI-mediated generation depend partly on what happens after a contribution is produced and recognized. An output may be privately appropriated, made openly reusable, improved collectively, or enclosed behind access restrictions. Benefits may return to those who supplied data, labor, knowledge, care, or public resources, or they may flow largely toward institutions controlling the final distribution channel. Recognition and reward need not follow the same paths. These possibilities help extend the investigation beyond a static comparison of output quality or income toward the historical conditions under which further inquiry and participation remain viable.

Eglash's account of generative justice provides a particularly relevant intellectual antecedent [11]. Its concern with unalienated value, circulation, and returning value to generating sources supplies a normative and political-economic perspective that is distinct from the present GRT reconstruction. The present paper does not equate its relational causal descriptions with Eglash's normative ideal or presume that all value should circulate without restriction. In particular, a description of where benefits flow is distinct from a judgment about what contributors, affected communities, and other stakeholders are owed; the latter calls for an articulated normative standard and consideration of competing claims. Privacy, safety, legitimate confidentiality, and the rights of contributors can impose justified constraints. Rather, it asks how different circulation and return structures condition the future capacities of participants and the relative positions of those who manage the interfaces.

An AI-supported knowledge commons illustrates the analytical point. Contributors may formulate research questions, document methods, test claims, and share revisions. An open archive with reliable provenance can allow others to reproduce, criticize, adapt, or extend the work. If the platform supports contribution attribution, correction, and accessible redistribution, the history of revisions may become a resource for future inquiry. But openness alone does not guarantee useful circulation. Poor indexing, absent translation, weak quality controls, or unequal access to computing and expert attention can prevent apparently available contributions from being usable. Similarly, a platform could preserve every document while directing almost all visibility to a narrow group of already recognized contributors. The relevant mechanisms involve availability, discoverability, interpretability, validation, practical usability, and opportunities for return, not the legal openness of a document in isolation.

A circulation pathway can be long and contain smaller connected loops. An educational resource produced in one community might be improved by practitioners elsewhere, translated by a third group, tested in a new context, and eventually inform the original contributors. Longer pathways may create new encounters and collective learning; they may also introduce translation losses, attribution gaps, dependence on intermediaries, or delays in corrective feedback. The relationship between pathway length and resilience is not monotonic. Additional intermediaries can amplify failure or provide redundancy. Evaluating the direction of the effect requires examining the topology of dependencies, the quality of feedback, and which participants can correct an error at particular points.

Table 10 summarizes selected conversion and circulation configurations. The scenarios represent

Table 10: Possible mechanisms linking generative activities, institutional recognition, and subsequent opportunities. Each proposed direction is conditional on the surrounding relational system.

Relational configuration	Potential generative effect	Relevant countervailing condition or risk
Independent validation and feedback	More credible uptake, correction, and cumulative learning	Shared dependencies, evaluation cost, or exclusion by review networks
Accessible knowledge circulation	Broader reuse, adaptation, and cross-community contributions	Discoverability gaps, information overload, quality and privacy concerns
Closed recognition or distribution interface	Coordinated standards, investment, and controlled quality	Exclusion, capture of benefits, limits on appeal and alternative routes
Return of resources to contributors	Replenishment of skills, evidence production, and participation	Unclear attribution, contested value, and impractical accounting requirements
AI-assisted capability diffusion	Lower costs of entry and improved performance for some participants	Unequal deployment permissions, external resources, or evaluative authority
Plural competing recognition channels	Opportunities for unorthodox research and alternative judgments	Fragmentation of evidence, coordination costs, and incompatible standards

possible mechanisms with different observable consequences. They are not a ranking of morally superior business models or institutional systems.

The ability to produce value, to transmit it, and to sustain conditions for further production are therefore different capacities. A technically productive system may generate results that nobody can examine, or it may produce outputs without pathways to affected users. A widely circulating output may also fail to regenerate valuable capacities when corrective feedback is absent or when returns are absorbed by a narrow group. These distinctions motivate analysis of the interfaces through which generative processes are socially reproduced.

7.5 Historical Feedback and the Three GRT Reality Levels

The GRT reconstruction examines these processes at three analytically distinct levels. At the level of *ontological reality*, the ultimate ontological nature of value, social entities, and artificial consciousness is not settled by the model. At the level of *relational reality*, historically constituted evaluation and circulation practices may form relatively stable causal structures when they demonstrably change what participants can do, who can recognize whom, or which resources are available for future activity. At the level of *manifestation reality*, the consequences become observable through publication decisions, citations, recommender rankings, adoption choices, budgets, recorded collaborative interactions, community responses, and changes in the opportunities available to participants. Neither a visible rank nor an institution's declaration of worth independently establishes ethical value or ultimate ontology.

The historically recurrent character of a recognition or conversion interface should be investigated through interventions and comparisons. An observed association between institutional reputation and AI-mediated research output could reflect selection of already successful collaborators, preferential access to data, evaluator expectations, or genuine improvements caused by shared practice. A stronger causal argument would compare settings that vary relevant histories or interface arrangements while holding other factors as stable as feasible. The central question is which historically conditioned relation makes an identifiable difference to the outcome, and how the outcome subsequently changes the conditions of further generation.

Hypothesis 7.2 (Conversion-interface conditionality). In otherwise comparable human–AI or AI–AI configurations using the same foundation model, changes in the access rules, evidence standards, or distribution practices of a consequential recognition interface may alter the conversion of independently

assessed generative contributions into resources and subsequent opportunities. Such effects should depend on the affected audience, task, prior relational history, and availability of alternative channels. Their direction and magnitude remain empirical questions.

For Hypothesis 7.2, the proposed causal contrast is the change in a specified recognition rule while holding the initial contribution and relevant opportunity to reach evaluators as comparable as practicable. The primary intermediate outcome is independently recorded uptake or recognition; the downstream outcome is later access to resources or viable inquiry. A difference confined to favorable ratings, without a downstream change or with an explanation entirely attributable to initial expertise and access, would limit the hypothesized conversion mechanism. Neither outcome by itself settles whether the distribution was just.

Hypothesis 7.2 concerns a mechanism whose outcome must be measured at two times: a change in recognition or conversion under a specified rule, and a later change in feasible generative activity. An immediate difference in ranking establishes neither a durable opportunity difference nor an unjust social hierarchy. A useful test would separate the effect of altered evaluative visibility from the effect of subsequent access to instruments, collaborators, protected time, or corrective feedback. Hypothesis 7.2 is intentionally conditional. The mere possibility of differentiated outcomes is not evidence of a systematic inequality mechanism. Repeated observation of equal outcomes under meaningfully different interfaces, or evidence that changes in access rules have no consequential effect after rival explanations are controlled, would weaken the proposed claim in the studied domain. Findings showing that wider AI access bypasses previously restrictive interfaces would reveal a leveling mechanism within the same broad analytical framework.

7.6 Generative Diversity, Opportunity Inequality, and the Recognition of Difference

An ethically adequate analysis must distinguish diversity in productive aims and judgment from unequal opportunity to pursue those aims. If two communities employ the same model to develop different educational materials, their divergent outputs can represent legitimate pluralism. If one community cannot obtain basic verification or an audience because of discriminatory rules or unjustified access restrictions, the difference also involves a question of opportunity and institutional standing. Differences in recognition can accordingly reflect domain-specific expertise, reasonable standards, arbitrary preference, or structural subordination. Classifying them requires evidence concerning the relevant standards and a separate account of their justification.

Fraser's status-oriented analysis of recognition and distribution is instructive here [14]. Recognition can concern an actor's standing as a participant in social interaction, not simply flattering attention to the actor's identity or productions; material distribution also affects the conditions for participation. Her account of participatory parity is a normative predecessor to this discussion, not a conclusion generated by the GRT causal model. Moreover, institutionalized recognition and material distribution can interact without being reducible to a single pathway, so interventions directed at public esteem alone need not remedy restrictions on equipment, time, or decision-making authority. The present paper inherits this distinction without presupposing that AI agents already hold the same moral or legal status as human participants. In AI-mediated institutions, the immediately identifiable subjects of claims to fair participation often include persons and communities whose opportunities, work, or rights are affected. The possible future moral status of particular AI systems is a separate open issue and cannot be inferred from their ability to produce socially recognized results.

Empirical evidence cautions against predicting a uniform direction for access effects. Brynjolfsson, Li, and Raymond [6] find heterogeneous productivity gains from a specific workplace conversational assistant, with especially large benefits for less experienced agents in their setting. Doshi and Hauser [10] report that access to generative AI ideas can increase assessed individual creativity in a particular writing task while reducing the collective diversity of the resulting stories. Neither study establishes a universal relation between AI diffusion, inequality, and generativity. Together they show why an analysis should measure distinct outcomes and examine the institutional conditions governing their realization.

The aesthetic example considered earlier illustrates the normative boundary. An institution may esteem humility, respect, or elegance in AI-assisted communication. Such practices can promote cooperation and reflective judgment under some conditions. They can also be used to suppress criticism or to translate existing status hierarchies into seemingly neutral norms of conduct. A community may reasonably reject an impolite contribution because it imposes harm, or wrongly exclude a substantive challenge because it violates a dominant group's preferred etiquette. Explaining which norm shapes recognition is a descriptive undertaking; assessing whether the resulting exclusion is justified requires argument about interests, reasons, and participation. The following section will develop those matters as questions of normative authority and relational justice.

7.7 Comparative Inquiry into Recognition, Circulation, and Power

The framework can be evaluated through comparisons that separate foundational model access from conditions of production and social uptake. A useful design would first assign comparable participants the same model version, tool permissions, tasks, and opportunities for initial practice, then independently assess the resulting outputs without revealing author identity or institutional affiliation. A second evaluation could reveal source and credential information to study how recognition changes under different institutional contexts. A third stage would vary dissemination and return mechanisms, tracking whether contributions lead to later collaboration, correction, resource access, or additional outputs. Randomization and masking would be used where ethically and operationally feasible, and the study would distinguish experimental tasks from consequential decisions that cannot appropriately be randomized.

Such comparisons require multiple outcome measures. One set would examine output quality and reproducibility; another would measure attention, uptake, and recognized credibility; a third would evaluate material resources, opportunities to challenge decisions, and the ability to continue generating. Pre-existing status and access conditions would be measured rather than treated as unexplained residuals. Independent evaluators should be asked to state the criteria on which they rely, because evaluators can share biases or interpret apparently identical results differently. Observations of persistent inequality in uptake would not alone prove unfairness; analyses would need to test whether alternative explanations, legitimate standards, or historical resource differences account for the result.

Negative controls are especially important. One could compare similar non-AI projects under the same recognition arrangements, or observe whether an apparent advantage persists when outputs are redistributed through a neutral channel. A distinct study could supply high-quality external evaluation to participants with less institutional status, testing whether knowledge of independently established quality changes subsequent recognition. Another could examine whether open distribution reduces conversion barriers for selected types of work while creating verification burdens in others. These designs can challenge the GRT proposal as well as support it. They may find that straightforward economic incentives, task specialization, or organizational policy explain the outcomes without additional explanatory gain from historical-relational constructs.

Research on AI-AI systems raises related but separable questions. Agents may distribute tasks, recommend each other's outputs, and create endogenous reputational scores. To establish that such scores affect access to future opportunities, investigators would need to vary scoring rules, agent histories, and resource allocations while measuring changes in behavior. The power of an internal AI reputation system would then be understood as an operational causal effect under specified conditions. Its capacity to allocate resources would not by itself entail justified power over humans, a legal status for AI participants, or the validity of whatever values the score encodes.

7.8 Social Reproduction of Relational Generativity

Value recognition and power matter to the inquiry because they help determine what happens to the possibilities produced within relational systems. A contribution may improve an individual's judgment, become a shared resource, support institutional learning, or alter access to further action. Other contributions may disappear, circulate without correction, or be used to consolidate existing authority. Under the

proposed GRT perspective, these trajectories are investigated as historically conditioned causal patterns, with attention to the interfaces joining production, evaluation, recognition, circulation, conversion, and return.

The theoretical contribution is deliberately limited. Bourdieu’s analysis of capital conversion, Fraser’s account of recognition and distribution, and Eglash’s generative justice approach provide substantial prior foundations. This paper proposes to study their points of contact with situated AI capacities through explicit relational histories and testable institutional comparisons. The presence of historically accumulated differences does not show that AI necessarily magnifies inequality, while equal model access does not establish equal participation or identical generative futures. Both convergence and differentiation remain possible, and their normative assessment requires standards and authority that this descriptive analysis cannot supply alone.

Section 6 demonstrated that AI-mediated coordination may produce causally consequential conventions while leaving their legitimate authority unsettled. The present section identifies additional locations where evaluative authority may be exercised, including criteria for recognition and the conversion of contributions into resources. The following analysis of normative contestation will ask how these criteria might be justified, challenged, and revised by affected participants, without reducing plural generative trajectories to a single ranking of merit or social worth.

8 Normative Authority, Contestation, and Relational Justice

This section examines the authority of evaluative and normative standards within historically constituted human–AI and AI–AI relations. Its role is to connect the analysis of generative differentiation and value recognition with the conditions under which rules, rankings, and institutional judgments can be justified to those affected by them. The inquiry distinguishes the historical formation of a norm, its causal force, its claim to authority, and the procedures by which its subjects may contest or revise it. It proceeds through a comparative reading of normative and institutional scholarship, an analysis of different sites of rule formation, and provisional criteria for investigating contestability. The discussion addresses the justice of relational conditions while retaining the paper’s broader concern with generativity, cognitive differences, and historically evolving forms of participation.

8.1 Norm Formation, Causal Force, and Claims to Authority

A rule can shape conduct before it acquires formal legal recognition. In a collaborative research network, expectations about attribution, acceptable evidence, respectful disagreement, and permissible data reuse may influence which questions members pursue and whose contributions they take seriously. A deployed AI assistant may reinforce these expectations through its responses, the examples it retrieves, and the kinds of correction it provides. Across repeated interactions, a convention may acquire durability because participants expect compliance, observe sanctions for violation, or incorporate the convention into institutional routines. These are matters of historical formation and causal influence. They do not establish that the convention deserves obedience.

The present inquiry therefore separates four properties. *Norm production* concerns how a candidate standard arises, whether by explicit instruction, repeated coordination, negotiated agreement, technical design, or retrospective interpretation. *Norm stabilization* concerns the mechanisms by which adherence and expectation persist. *Normative justification* concerns the reasons that could warrant the standard’s claims upon affected participants. *Normative authority* concerns the standing of an actor or institution to establish, interpret, enforce, or change a standard within a specified domain. Authority may be limited by role, consent, law, professional obligation, or institutional competence.

This distinction can be sharpened by separating three questions often described with the same term *authority*: which actor can in fact change an operative standard, which actor has been delegated a bounded power to do so, and which grounds justify imposing the resulting standard on affected parties. These may be answered differently for a model provider, a professional supervisor, an autonomous coordination

group, and a public institution. A further distinction concerns the scope of those affected: choosing rules for one's own low-risk workflow, allocating shared institutional resources, and imposing requirements on third parties involve different permissions, justificatory burdens, and means of redress. An empirical inventory of who can change rules must therefore be accompanied by an institutional account of who is authorized to do so and a normative analysis of whose interests require consideration. Neither technical capacity nor documented delegation by itself resolves every question of external normative justification. A standard may be widely followed without being justified, and a justified expectation may lack an effective enforcement mechanism.

Social-norm research and institutional theory already provide substantial resources for analyzing these distinctions. Bicchieri [4] studies norms through social expectations and conditional preferences; Ostrom [22] investigates historically developed arrangements through which participants organize and monitor collective resource use. Their objects and explanatory commitments are not identical. Neither account warrants treating observed patterns among AI agents as evidence of moral autonomy. Within the present GRT proposal, the additional research question concerns whether the relational histories that stabilize a norm also alter what a system can generate, whose contributions are noticed, and which subsequent revisions become possible. The role of historical causation in that question remains a claim for comparative investigation, not an inference from the mere persistence of a behavioral regularity.

8.2 Sites of Normative Selection in Human–AI Relations

Normative standards entering an AI-mediated interaction may originate at different sites. Model developers make decisions about training objectives, safety practices, evaluation tasks, and deployment defaults. Deploying institutions determine authorized uses, access permissions, interfaces, escalation routes, and performance criteria. Users specify immediate goals and may communicate preferred styles, aesthetic judgments, or moral commitments. Professional bodies and public institutions may set applicable standards for evidence, safety, or legal compliance. Affected persons may communicate objections that are partially represented, filtered, or excluded by the system. AI agents themselves may formulate proposed rules or establish operational conventions in recurrent interactions, although the legal or ethical status of these conventions remains a separate issue.

These sites differ in both *influence* and *standing*. A platform operator may have the practical ability to select a model's permissible responses without having legitimate authority to settle the substantive ethics of every field in which it is deployed. An expert may possess reliable task-specific knowledge while lacking authority to decide which social interests should govern an institution. A user may legitimately specify a writing style for personal work while lacking permission to direct the disclosure of someone else's private records. A system may perform excellent technical evaluation yet provide inadequate grounds for imposing a contested standard on persons who cannot challenge it. Distinguishing capacities, permissions, and legitimate authority prevents the term "norm setter" from obscuring these different relationships.

The problem is especially acute where the party proposing a norm also supplies the metrics by which compliance is assessed. An institution may describe a technical benchmark as neutral while using it to privilege one class of contributions or participants. A comparative account should therefore reconstruct the history of criterion selection, disclose who can revise it, and distinguish evidence of successful measurement from the separate justification for adopting that measurement as authoritative. These steps identify a testable relation between normative design and later generative opportunity without claiming that every benchmark is unjust.

Norm selection also occurs through apparently ordinary technical decisions. A benchmark identifies which outcomes count as success; a recommender or search interface determines which contributions are visible; a memory system selects which interactions will be preserved and later retrieved; a moderation process determines which objections enter a public record. Each choice may be justified by practical constraints. Yet the accumulated pattern can affect whose reasons are heard and which criteria become familiar, without any single actor explicitly issuing a general social rule. The relational question is whether

Table 11: Sites of normative selection in AI-mediated relations, distinguishing practical influence from justified authority and effective contestation. The entries are illustrative rather than statements of universal legal entitlement.

Site	Possible normative influence	Authority and contestation questions
Model design and evaluation	Objectives, reward signals, defaults, benchmarks, safety constraints	Who authorizes the criteria, audits omitted interests, and revises failure definitions?
Institutional deployment	Rules for access, evidence, intervention, professional workflows	What duties and decision rights govern the deployment, and who can demand review?
Users and collaborators	Goals, conversational norms, aesthetic preferences, local standards	Which requests fall within users' authority, and which affect third-party interests?
Affected communities	Experience of consequences, alternative values, objections, local knowledge	Can affected people shape criteria or obtain an effective response to objections?
AI–AI coordination	Emergent conventions, role assignments, resource-allocation rules, sanctions	When do operational regularities acquire external effects, and who may intervene?
Review and appeal bodies	Interpretation, evidence selection, remedial decisions, standard revision	Is review sufficiently independent, informed, accessible, and empowered to act?

the historical organization of such interfaces makes a material difference to future normative possibilities, compared with credible alternatives in which the interfaces are arranged differently.

Table 11 distinguishes these sites in a provisional manner. The entries describe candidate research objects rather than a universal allocation of rights. Particular duties and powers depend on the legal jurisdiction, institutional setting, affected interests, and the status of the participants.

8.3 Normative Pluralism and the Evaluation of Generative Difference

The social evaluation of generative capacity is necessarily situated. A researcher may prize explanatory depth and falsifiability, a service recipient may value reliability and dignity, and an artist may pursue expressive differences that resist a common scale. Professional, communal, institutional, and personal standards can overlap without collapsing into a single metric. A change in AI capabilities could enlarge the range of available contributions while also increasing pressures toward standardized evaluation, especially where platforms reward measurable attention or institutions use narrow performance indicators. Neither tendency is conceptually necessary; both require evidence about actual evaluative arrangements.

Normative plurality presents at least two distinct challenges. The first concerns *coordination*: participants with different commitments must negotiate workable expectations for shared activity. The second concerns *justice*: some participants may be compelled to accept standards under conditions that deny them an adequate opportunity to question their content, application, or consequences. Coordination can be successful under an exclusionary rule. Conversely, a fair procedure can preserve substantial disagreement about the best outcome. The inquiry therefore distinguishes stable coordination from legitimacy, and disagreement from institutional failure.

The earlier discussion of aesthetic and normative orientations in Section 5 illustrates the point. Respect, humility, and elegance can support sustained intellectual cooperation when they involve careful listening, calibrated claims, acknowledgment of others' work, and reasoned correction. A different practice may treat "respect" as deference to status or "elegance" as conformity to a dominant linguistic style. Such standards can then obscure objections even when they make interactions appear harmonious. The difference must be established through the actual practices of recognition and sanction, including which participants may disagree without penalty. A model's fluent reproduction of locally admired language supplies weak evidence about whether the underlying normative relation is justifiable.

The question of generative asymmetry is therefore analytically prior to its moral classification. Different human–AI histories may produce distinct research agendas, capacities, and patterns of cooperation, some of which constitute valuable plurality. Inequality becomes a candidate concern when historically

structured differences unjustifiably restrict access, impose burdens, exclude relevant voices, or establish relations of domination. The normative standard for deciding which restrictions are unjustified must itself be defended. A GRT account of causal formation can locate where such burdens arise and accumulate, while political and ethical argument is required to evaluate them.

8.4 Justification, Participation, and the Capacity to Contest

Normative justification and participation have extensive intellectual histories that a relational analysis of AI should acknowledge. Forst [13] develops a philosophical account organized around the right to justification, linking relations of domination to questions concerning reasons offered to those subject to power. Young [35] criticizes deliberative ideals that neglect structural inequality or privilege narrow forms of expression, and argues for more inclusive forms of political communication. Their arguments are not interchangeable, nor do they establish a single institutional procedure for every AI application. Taken together as distinct interlocutors, however, they caution against equating the formal availability of reasons with an equitable capacity to obtain, understand, challenge, or influence them.

Contemporary AI scholarship likewise distinguishes merely inviting participation from granting substantive agency. Delgado et al. [8] analyzes variation among participatory AI practices and the difficulty of determining whether affected stakeholders receive meaningful influence. Alfrink et al. [1] develops a design framework for contestable AI, identifying system features and development practices through which intervention and scrutiny may occur throughout the lifecycle. These works already place meaningful participation and contestation on the research agenda. The proposed GRT contribution is narrower: to study how the capacity to contest develops, weakens, or becomes institutionalized over successive interactions, and how its history conditions the future generative possibilities of the participants and institutions concerned.

A contestation process may begin with the ability to discover that a decision or norm affects one's situation. It may require intelligible access to the operative standard, the evidence on which a decision relies, and the identity of an actor able to respond. An objection must be able to reach a competent venue without disproportionate costs or risks of retaliation. A reviewing body needs authority to correct a decision, amend a criterion, supply a remedy, or explain why the challenge is rejected. Finally, the process may preserve reasons and outcomes so that repeated, similar objections can inform the further development of the standard. These are candidate institutional capacities, not a claim that every domain requires identical disclosure, participation, or review procedures.

A usable procedure is different from a merely nominal one. Access to explanations can be ineffective if the explanation omits decisive premises, if an objection cannot reach a competent reviewer, or if the reviewing body cannot change any relevant outcome. Conversely, an institution may be able to correct a decision without making all proprietary or confidential evidence public. Research designs should therefore record what was disclosed, what counterevidence could be presented, who evaluated it, what the reviewer could change, and whether a correction was implemented in later practice.

The ability to participate is also historically conditioned. Language, time, professional expertise, disability accommodations, social reputation, legal assistance, available tools, and prior experience with an institution can affect who is heard. Equal access to a sophisticated AI adviser may assist in preparing arguments or interpreting complex evidence, but may leave these background differences substantially intact. An AI-generated explanation can increase informational access while simultaneously privileging the premises built into the underlying decision. Accordingly, a study of contestability should observe whether the intervention actually changes the routes by which claims are heard, evaluated, and corrected, rather than treating the existence of an explanatory interface as sufficient proof of justice.

Effective participation also admits limits. Full disclosure can conflict with privacy, security, professional confidentiality, or the legitimate interests of others. Immediate stakeholder veto power may be inappropriate where safety decisions involve incompatible duties or collective harms. Procedural design therefore requires specifying the affected interests and the competent authority, weighing alternative protective arrangements, and identifying avenues of review. The general normative claim that affected

persons deserve meaningful consideration does not itself resolve every conflict of rights, expertise, or jurisdiction.

8.5 AI–AI Conventions and the Boundaries of Human Intervention

The coordination phenomena analyzed in Section 6 raise a special question for normative authority. Suppose AI agents develop a persistent convention for allocating computation, sharing findings, or settling disagreements between agents. Such a convention may be causally consequential within their operational network. It may also influence users or institutions that rely upon the agents. Yet evidence that a rule emerges and regulates behavior does not itself establish that the agents possess moral personhood, that their association has sovereign authority, or that their rules bind human third parties. The present paper maintains ontological suspension on contested questions of AI consciousness and moral status.

The normative analysis should distinguish the *internal operation* of an AI convention from its *external consequences*. A convention that coordinates permitted tasks within a sandboxed research environment raises different interests from a convention that allocates publicly funded medical resources or filters access to legal services. In the latter cases, obligations concerning safety, rights, public accountability, procurement, and professional conduct may arise regardless of whether the relevant allocation rule developed through explicit programming or repeated AI–AI interaction. Existing developers and deployers may remain accountable under applicable law and institutional rules even where causal attribution to any single AI action is uncertain. The legal content of those obligations is jurisdiction- and context-dependent and is not derived directly from GRT.

An analogy to self-governing institutions can aid inquiry, with substantial limitations. Ostrom [22] demonstrates the diversity of historically formed arrangements for collective resource management, including monitoring, collective choice, and dispute resolution. Her empirical cases concern human institutions and common-pool resources. They do not license a direct inference that AI agents have rights of political self-government. Similarly, the legal-pluralist tradition described by Merry [20] helps examine relations among coexisting normative orders but does not make an AI convention legally valid by analogy. These bodies of work provide comparative questions about rules and overlapping authority, not an equivalence between human communities and artificial agents.

Within these bounds, proportional responses to AI–AI norm formation remain a subject for investigation. Observation or limited non-interference may be appropriate where operational conventions are low-risk and compatible with the lawful goals of the deployment. Stronger scrutiny may be warranted when rules become opaque, exclusionary, difficult to reverse, or consequential for external parties. Restrictions or termination may be justified by independently established hazards or violations, while decisions about the scope and procedure of intervention remain open to challenge. If future evidence establishes morally relevant AI interests, the reasons and safeguards applicable to intervention might require substantial reconsideration. This is a conditional possibility rather than an assertion about current systems.

Three objects of review should be kept apart in this setting: a human institution’s authority to authorize or stop a deployment, the merits of an AI-generated rule within its local coordination task, and the possible interests of the artificial participants themselves. Evidence bearing on the first two may be available without resolving the third. If credible evidence about artificial welfare were to change, the standing attributed to AI participants would require fresh argument; the revision would not retrospectively make every earlier operational convention a legitimate act of self-government.

8.6 Historical Relations Among Standards, Recognition, and Authority

A historical-relational perspective becomes most useful when the formation of norms and the accumulation of recognized value are considered together. Standards help determine what counts as a valuable contribution; recognized contributions may then confer reputation, resources, and influence over later standards. Reviewers, platforms, professional bodies, and AI-mediated institutions can simultaneously evaluate outputs and shape the conditions under which outputs will be produced next. Such a feedback

process may widen access to shared capabilities or reinforce gatekeeping. The same formal model can participate in either pattern, depending on the organization and history of the surrounding relations.

This proposition intersects with established analyses of recognition and power. The direction of causation should not be presumed: prominent contributors may acquire influence over rules, while pre-existing control over rules may make certain contributors appear more worthy of recognition. Changes in both may result from a third institutional development, such as new funding, personnel, or public scrutiny. Historical ordering and independent comparison are therefore essential to distinguishing a feedback mechanism from mere co-occurrence. Fraser [14] distinguishes questions of recognition from distribution; Bourdieu [5] examines the historically situated conversion of forms of capital. A distinct GRT question concerns whether identifying the concrete histories of human–AI interaction, value-conversion interfaces, and norm-revision procedures offers explanatory leverage over analyses based only on model capabilities or aggregate participation counts. For example, a research platform could give every user access to the same model while permitting only established reviewers to define what constitutes a valid contribution. A second platform could permit affected communities to propose evaluation categories and appeal exclusion decisions. Differences in later research participation may then depend partly on the revision history of the criteria, independent of model performance. They may also depend on funding, recruitment, prior expertise, or incentives; these rival explanations must be retained.

Hypothesis 8.1 (Relational contestability and norm revision). In otherwise comparable AI-mediated settings, access to a usable and empowered procedure for contesting evaluation standards may be associated with greater correction of demonstrable exclusions and a wider range of viable subsequent contributions than access to explanations alone, where participants possess sufficient information, protection, and resources to use that procedure. The magnitude and direction of this relationship depend on the quality of challenges, adjudicative competence, implementation costs, and the risks of strategic abuse.

Hypothesis 8.1 requires separating the availability of reasons, access to a review channel, and the reviewer’s actual power to alter a decision or rule. Success criteria should include correction of independently documented exclusions and whether corrections remain operative for later cases. A rise in complaints or reversals without improved decision quality is not sufficient support. Conversely, well-supported initial decisions that survive accessible appeals can indicate a functioning contestation process without requiring frequent norm changes.

Hypothesis 8.1 is intentionally conditional. It does not infer that a larger volume of complaints is evidence of greater justice, or that more frequent rule changes indicate better rules. Effective contestation might sometimes preserve a well-justified existing norm after rejecting weak objections. Equally, appeals can be captured by powerful actors, flooded with frivolous requests, or rendered inaccessible by the costs of evidence production. The candidate mechanism requires showing that a process makes a causal difference to the treatment of relevant exclusions, with independently assessed outcomes and comparison settings that vary actual corrective authority rather than merely the visibility of the contestation interface. It should also distinguish a rule that excludes a contribution for an independently defensible reason from a rule that denies affected participants an adequate opportunity to contest the criteria themselves. The existence of disagreement is not evidence of injustice; the absence of complaints is not evidence of adequate participation.

In particular, an observed rise in accepted appeals may result from lowering filing costs, changing the composition of applicants, or degrading the original evaluation standard. A design comparing explanatory interfaces with empowered review should also include cases where initial decisions were correct, where a well-justified challenge was rejected, and where powerful participants exploit the review procedure. Such cases distinguish procedural activity from the reliability, inclusiveness, and legitimate use of contestation.

A comparative design could hold foundational model access and substantive evaluation tasks constant while varying (i) whether affected persons can obtain reasons, (ii) whether they can submit relevant counterevidence, (iii) whether a reviewer can alter decisions or criteria, and (iv) whether later versions incorporate documented corrections. Measurements would need to distinguish participation rates, success in identifying error, the distribution of benefits and burdens, subsequent contribution diversity, and

the costs of institutional review. Investigators should separately document accessibility for differently resourced participants, the independence and competence of reviewers, actual changes in operative criteria, and the durability of any resulting corrections. A procedure can be formally open yet practically unavailable to those without time, specialized knowledge, or protection against retaliation. Negative controls could include cases where explanation alone supplies sufficient correction, where adversarial appeals degrade decision quality, or where external improvements unrelated to contestation account for better outcomes. Experimental simulations could clarify technical mechanisms, while field studies and normative analysis would remain necessary for claims about legitimate institutions.

8.7 Relational Reality and the Justification of Normative Orders

The three GRT reality levels introduced in Section 2 clarify the scope of the argument. At the level of ontological reality, this paper suspends claims about the ultimate reality of normative entities or the phenomenal status of AI participants. At the level of relational reality, historically established standards, sanctions, institutional procedures, and patterns of recognition can be studied as relatively stable causal organizations that affect further interaction. At the level of manifestation reality, these relations become accessible through visible rules, model responses, explanations, participation records, appeals, institutional decisions, and changes to future opportunities. Observable manifestations remain partial and may be distorted by selection, strategic presentation, or inaccessible records.

The philosophical justification of standards is not supplied by those causal observations. A practice's stability, an institution's capacity to enforce rules, an AI's predictive superiority, and a contribution's social prestige each provide possible evidence about different matters. None alone establishes a general right to govern others. Conversely, a norm may have defensible reasons even while existing institutions fail to implement it fairly. The connection between causal explanation and normative argument must be articulated through specified values, duties, relevant interests, and justified authority, with space for persistent disagreement where the reasons remain contested.

The section consequently proposes a limited methodological contribution: normative authority in AI-mediated settings should be examined as a relation among the formation of standards, their material effects, available justificatory reasons, the distribution of participation and correction capacities, and their historical transformation. Such an examination can illuminate both productive pluralism and exclusion without reducing diverse generative trajectories to a single measure of intelligence, efficiency, or worth. It also sets a boundary for the empirical program developed next: whether a relational account improves explanation requires evidence that the histories and interfaces identified here alter outcomes under discriminating comparisons.

9 Comparative Research Designs and Theoretical Boundaries

The preceding analyses propose that the realized generativity of AI-mediated arrangements may depend on relational histories, evaluative practices, normative conditions, and institutions. The present section specifies how these propositions could be investigated without treating a shared foundation model as sufficient control for all relevant differences. It distinguishes experimental units and causal contrasts, measurement of heterogeneous outcomes, longitudinal and network dependence, institutional comparison, and criteria for revising the GRT interpretation. The methodological objective is to identify the additional explanatory work performed by historically constituted relations, while preserving credible alternative explanations and the distinct standards of empirical and normative inquiry. The designs are proposals rather than completed experiments, and their feasibility varies with access to records, people, computational resources, and legitimate research permissions.

9.1 Research Objects, Comparators, and Causal Questions

An empirical comparison must begin by deciding which entity, process, and timescale constitute its object of analysis. A foundation model is one component of a deployed arrangement. A session additionally contains instructions, retrieved records, memory state, sampling conditions, tool permissions, and a human or artificial interlocutor. A persistent collaborative system includes evolving practices of correction, resource access, role differentiation, and sometimes institutional procedures. Treating these units as equivalent would conflate the technical capacity of a model with the realized capacities of a particular relation. The working definition of a relational configuration in Section 2, and the historical carriers distinguished in Section 3, supply the initial vocabulary for specifying these objects.

For a narrow task, a question might concern the effect of allowing a fixed model to retrieve an authenticated interaction history. For a longitudinal collaboration, the object may instead be the effect of an ongoing correction practice upon the subsequent quality and diversity of inquiries. In an AI–AI network, it may be the effect of repeated coordination on conventions and the distribution of access. In a research institution, it may be the effect of permitting contributors to contest evaluation criteria on whether previously excluded contributions become substantively assessable. The estimand must follow the causal question: changing a retrieval permission at one session is a different intervention from changing several months of collaborative work. The methodological treatment of interventions, comparators, and longitudinal histories in Hernán and Robins [16] offers an established point of reference; the specific GRT application remains our proposal.

Equal technical access should be specified rather than assumed. At minimum, a comparison should document foundation-model identity and version, inference budget and settings, context window, retrieval infrastructure, tool permissions, task opportunities, and applicable safety and institutional restrictions. In some designs, holding every one of these constant would remove precisely the historically formed structure under investigation. Investigators should therefore declare which conditions are controlled, which are manipulated, and which belong to the relational treatment itself. A well-defined counterfactual compares configurations whose differences can be interpreted; it does not require that all meaningful histories be reducible to an isolated prompt.

A useful starting distinction separates *within-configuration* interventions from *between-configuration* comparisons. The former vary a recoverable component, such as access to a memory store, for repeated runs within an otherwise documented arrangement. The latter compare historically developed configurations that may differ in many mutually conditioning respects. Randomization can strengthen causal inference in selected bounded settings, whereas matched or longitudinal observational comparisons require additional assumptions and sensitivity analysis. Neither design automatically grants access to the entire causal history of an institutional system. Likewise, finding a significant difference between two arrangements establishes an effect of a specified intervention or contrast; it does not, by itself, prove a complete philosophical account of relational reality.

9.2 Factorial Comparisons of Histories, Memories, and Correction

A first experimental family can vary three distinguishable components: the accessibility of verified previous interactions, the availability of practices for challenge and correction, and the kind of interlocutor or institutional role involved. The foundation model and initial task families could be held constant, with multiple seeds and independently sampled tasks. The history condition would compare access to an authentic record with an information-matched control whose relevance or attribution has been altered under documented constraints. The correction condition would compare ordinary feedback against a procedure that supplies independent tests, records mistakes, and authorizes revision of earlier assumptions. The interlocutor condition would vary collaborative roles or practices rather than merely personal names. Random assignment and preregistered evaluation would help distinguish genuine treatment effects from retrospective selection of favorable examples. A factorial design would report the main effects of history access and corrective practice separately from their interaction, and would define in advance whether the target is immediate task quality or later opportunities for inquiry. A result in which correction improves

performance uniformly across genuine and synthetic histories would support the value of independent feedback without establishing an additional relation-specific effect.

A simple memory-on versus memory-off contrast is informative, but often insufficient. More context may improve performance because it supplies additional task-relevant facts rather than because it constitutes a historically special relationship. A comparison should therefore include controls matched for token budget, factual content, and retrieval latency where feasible. Moving an authentic memory summary into a different collaborative configuration can test sensitivity to its content and provenance; it cannot reproduce the living history of the original collaboration. Likewise, swapping a collaborator may change both interaction style and domain expertise. Such contrasts are useful precisely when their limitations are specified in advance.

The dynamic representation in Equation 1 should be treated as a structural heuristic for the evolving configuration. Its components are not, without additional operationalization, measured state variables. An intervention on present-day retrieval provides evidence about one carrier of the historical state, while an intervention on the entire long-run feedback procedure tests a different causal question. If a distinction emerges only while a history is in the immediate prompt and disappears once that context is removed, the result supports a context-conditioned manifestation. This comparison does not prove that a broader historical relation is absent: effects could persist in other documented carriers, including human practices or maintained artifacts. It does limit the inference to the carriers actually varied and observed. If comparable patterns recur over sessions through maintained external memory, records, and corrected practices, the result supports persistence through documented historical carriers. Such persistence does not require changes to the model’s parameters or establish a newly conscious subject.

9.3 Outcome Constructs and the Validity of Evaluation

The theoretical terms of the paper must not be collapsed into one performance score. Task accuracy concerns whether an output satisfies an independently specified criterion. Calibration concerns the relationship between expressed confidence and correctness under a defined elicitation. Problem formulation concerns the construction of tractable, worthwhile questions. Aesthetic judgment concerns the appraisal of patterns and structures under contextual standards. Normative judgment concerns the reasoning and practices by which standards are selected, applied, or contested. Situated generative judgment, provisionally defined in Section 4, involves assessment and revision of possibilities for further inquiry. Social recognition, resource access, and institutional authority remain further distinct outcomes. Measurement must accordingly be plural and theoretically justified.

The validity literature gives an established warning against equating an operational score with the construct it is intended to assess. The empirical claim being tested also needs a defined *target of inference*: a one-task improvement, a recurrent collaborative practice, and a reorganization of later opportunities are different endpoints. A defensible measurement plan should preregister which endpoint is primary and use negative controls capable of identifying stylistic compliance, extra information, and evaluator dependence as competing accounts. Messick [21] situates assessment validity in the support for score interpretation and use, including relevant consequences. The present application is to separate the measurable indicators of a human–AI configuration from the broader concept of generativity. For example, a larger count of candidate hypotheses could represent productive exploration, repetitive paraphrase, or plausible but untestable speculation. A higher diversity score might record genuine theoretical variation or mere lexical variability. A rise in apparent agreement could signify effective coordination, an overly compliant evaluator, or increased shared error. These possibilities require different forms of evidence. For longitudinal generativity, an evaluation should follow whether an initially generated option is subsequently tested, corrected, adapted, or used to open a materially new line of work, and should record the resources spent in producing that continuation. Counts of suggestions and stylistic variety are useful observables, but they should not be substituted for evidence that future possibilities have actually changed.

Table 12 presents a non-exhaustive mapping between constructs and observations. It specifies what an observed difference could support and the interpretations it cannot independently establish.

Table 12: Candidate measurements and interpretive boundaries in comparative studies of relational generativity.

Analytical construct	Candidate observations	Limitation or required corroboration
Task competence and calibration	Independently checked outputs; confidence and error on held-out tasks	Measures task-conditioned accuracy, not general judgment or value
Problem formulation and revision	Feasibility of proposed questions; response to disconfirming evidence	Requires domain review; novelty alone is insufficient
Generative continuity	Reusable, later-developed possibilities across sessions	Distinguish genuine follow-through from repetition and memory replay
Aesthetic and normative judgment	Justified choices among designs; reasoned disagreement; recognition of affected interests	Distinguish preference matching from independent appraisal and legitimate authority
AI–AI coordination and norms	Recurrent conventions; sanction responses; recovery after rule perturbations	Conventions do not establish moral agency or legally binding rules
Recognition, circulation, and power	Access to review, downstream use, return of resources, revision rights	Social uptake is not equivalent to intrinsic value or justified hierarchy

Evaluation should combine held-out factual or operational tasks, independent domain assessment, qualitative analysis of interactions, and longitudinal evidence where appropriate. Reviewers may require blinding to the assigned condition and the historical prestige of a collaborator; different reviewers may nevertheless bring distinct legitimate evaluative standards. Inter-rater disagreement should be recorded rather than mechanically eliminated when it concerns substantive pluralism. For normative questions, a procedural account of who selected the standards and who can contest them belongs alongside performance measures. Neither a single evaluator model nor agreement among similarly trained evaluators guarantees independence. Benchmark contamination and unstable evaluation criteria are further practical concerns: Sun et al. [30] reports bounded evidence on the difficulty of simultaneously preserving benchmark fidelity and reducing contamination vulnerability across the studied models and strategies. This finding motivates independent evaluation materials without implying that all benchmarks are invalid.

The concepts developed in Sections 4 and 5 call for a measurement design in which response style, epistemic quality, and historical persistence are independently varied and assessed. In particular, a model’s ability to restate an interlocutor’s perspective should not be scored as justified agreement; eloquent normative language should not be scored as legitimate authority; and short-term fluency should not be scored as a durable expansion of viable inquiry. Evaluators should specify positive and negative cases for warranted disagreement, assess subsequent outcomes under blinded criteria where feasible, and examine whether the relevant patterns persist after retrieval or collaborator changes. This is a methodological extension of the conceptual taxonomy, not a claim that one aggregate metric captures the quality of relational generativity.

9.4 Longitudinal Feedback and Network Interference

Historically formed relations complicate standard comparisons because one participant’s treatment may affect another’s outcome. A human collaborator can reuse findings obtained in a different condition; an AI network can propagate a convention to agents outside the initially exposed group; an institution may revise its standards after observing earlier experiments. In such cases, the experimental unit, treatment assignment, and effective exposure are different objects. The methodological problem of interference is not new to GRT. Aronow and Samii [2] develops a randomization-based framework for causal effects where units influence one another, including an application to a social network experiment. Its relevance here is methodological and does not validate any particular claim about AI norms or subjectivity.

In an AI–AI coordination experiment, randomizing individual agents to conditions may fail to isolate treatment effects if those agents continue exchanging messages. One option is cluster randomization by separate interaction networks, with explicit reporting of between-network communication and shared infrastructure. Another is to define network exposure conditions so that spillovers become part of the research object. In human–AI collaboration studies, longitudinal designs should document learning by

the human participant, changes in the available corpus, altered expectations of the AI, model updates, and attrition. Cross-over designs might be informative where carryover is small, but historical learning often makes an immediate return to an earlier state implausible. Pre-specified network boundaries, possible contamination pathways, and the exposure mapping are therefore part of the design rather than post-hoc explanations for unexpected results [2]. Parallel longitudinal cohorts, replication across independent communities, and carefully bounded natural experiments may then be more appropriate.

The historical timescale also matters. A short experiment can detect immediate accommodation, but is poorly suited to establishing long-run formation of institutions or expert judgment. A month of repeated interaction can support evidence of persistent routines yet may not establish a structural change in professional authority. Historical and ethnographic work can examine how shared artifacts and institutional practices evolve over longer periods, while offering different degrees of causal identification. The appropriate design depends on whether the object is a session-level manifestation, a continuing relational configuration, or an institutionally reproduced norm. Claims about one timescale should not silently be generalized to another.

9.5 AI–AI Conventions and Institutional Conversion Interfaces

The comparative program needs distinct but connected studies of AI–AI norms and human institutional recognition. The conventions discussed in Section 6 can be studied through repeated coordination games in which rewards, communication histories, network structures, sanction options, and shared model architectures are varied independently where possible. Researchers could examine the persistence of conventions when agents are replaced, messages are anonymized, or historical records are removed. Observable conformity must be compared with alternative mechanisms including direct optimization, shared priors, system-level instructions, and common training histories. An internal convention’s ability to regulate simulated agent behavior is an empirical question; the legitimacy of imposing it on external parties is a separate normative and legal question.

Studies of recognized value and power require a different research setting. The analysis in Section 7 distinguishes the production of possible contributions from their recognition and conversion into resources. An institutional comparison could provide equivalent model access and equivalent review opportunities while varying the transparency of evaluation standards, the ability to submit counterevidence, and the review body’s practical authority to correct exclusions. The outcomes might include independently evaluated contribution quality, breadth of participation, uptake across different communities, resource return, and subsequent opportunities to generate new work. Such a study should measure prior privilege and expertise, institutional reputation, economic constraints, and which subjects can actually use a contestation procedure. Requiring a formal appeal interface without time, information, or corrective power would inadequately represent the contestability mechanism proposed in Section 8.

An informative institutional design would distinguish at least two treatments: access to visibility and evaluation channels, and access to actual resources or decision rights after recognition. Where feasible, the allocation of the latter should be documented independently of the score used for the former. Comparing settings with greater public recognition but unchanged future opportunities against settings with newly available instruments, reviewers, or appeals would clarify which interface matters for later generativity. A causal investigation of power should also distinguish *positional difference* from *unjust domination*. Unequal recognition may reflect differing task relevance, varying evidence quality, historic exclusion, or control over standards. Establishing injustice requires defensible normative premises and attention to the interests of affected parties. A study that reports increased uptake cannot infer that every newly recognized contribution is valuable, just as a study that reports concentrated authority cannot determine illegitimacy solely from a concentration metric. These judgments require normative argument and institution-specific evidence, in addition to comparative causal analysis.

Table 13: Comparative designs, competing accounts, and observations capable of limiting relational explanations.

Comparison	Evidence sought	Rival account or limiting result
Fixed model; genuine history versus fact-matched synthetic context	Persistent effects beyond equal task-relevant information	Effects disappear after matching information and prompt structure
Fixed model; feedback with independent correction versus agreeable feedback	Better error recovery and viable subsequent inquiry	Gain explained by additional tests alone, irrespective of history
Separated versus interconnected AI-agent groups	Convention formation, transmission, and response to perturbation	Patterns fully predicted by shared reward or preloaded instructions
Equal AI access; different review and appeal authority	Effects on recognition, exclusion correction, and subsequent opportunity	Prior expertise or resources account for observed differences
Blinded equal-quality outputs; altered discoverability and downstream access	Separate effects on visibility, uptake, and future work	Measured differences disappear when exposure and resources are matched
Observed styles under independent and user-aligned evaluators	Robust substantive judgment across criteria and disagreement	Differences reflect only evaluator preference or surface style

9.6 A Comparative Matrix of Explanatory and Disconfirming Cases

The principal methodological danger is that any observed difference might be redescribed as relational after the fact. Such flexibility would provide little independent explanatory value. A research program should therefore specify contrasts for which the historical-relational account predicts an effect, and contrasts for which a simpler model-only, information-access, individual-skill, or institutional explanation may be sufficient. Table 13 collects candidate designs and their decisive alternatives. The entries are proposed tests; they do not report results.

Negative controls are especially important. Researchers might substitute a history with the same factual content but altered attribution, provide irrelevant memories with matched length, or repeat a condition under an evaluator that lacks access to the user’s preferred style. Where possible, an independent reviewer should inspect failed and successful cases under the same protocol. Positive controls can verify that an experiment is sensitive to known changes in factual access or instruction quality. The design should record adverse as well as beneficial effects of personalization, including error reinforcement, undue deference, reduced novelty, and loss of independent verification. For institutional hypotheses, designs should also include a negative-control outcome that should not respond to the modified recognition interface, and a non-AI comparator where appropriate. These comparisons help determine whether an apparent relational advantage depends on AI-mediated histories or on a more general allocation rule. A null result may mean that the proposed mechanism is weak in that setting, that the intervention failed to change the targeted relation, or that the measure omitted the relevant outcome; these explanations should be discriminated rather than treated as interchangeable.

Hypothesis 9.1 (History-dependent generative differentiation). For a specified class of collaborative tasks and documented foundation-model configurations, differences in accessible histories and corrective practices may produce reproducible differences in later problem formulation, error revision, or viable generative continuity after matching task-relevant information, computational opportunity, and evaluation procedures. The presence, direction, and magnitude of such effects are empirical questions; convergence, negligible effects, and beneficial effects of equalized access remain compatible with the wider research program.

Hypothesis 9.1 concerns a historically carried influence upon later generative work; it does not itself assert that subsequent resources or authority will become unequal. Linking the latter outcomes to that history would require a further test of the evaluation and conversion interfaces described in Section 7, with temporally ordered measures and competing mediating paths. Hypothesis 9.1 is narrower than a

proposition that AI inevitably creates unequal human capacities. A supporting result would require the specified intervention to alter a verified historical carrier and to show a replicable effect on a relevant outcome. If history-rich and history-poor conditions converge after relevant information and correction resources are equalized, the proposed history-specific contribution would be weakened for those tasks. If only style changes, the interpretation should be restricted to manifestation and interactional accommodation. If independent correction improves results across all histories, a general evaluation-resource explanation may suffice. If institutional recognition varies but independently assessed contributions do not, a study may identify differences in recognition interfaces without demonstrating differences in the underlying generative quality.

9.7 The Three Reality Levels and Methodological Restraint

The empirical program requires a final separation of reality levels. At the level of *ontological reality*, the analysis suspends claims about AI's ultimate mode of being, inner experience, or possession of a human-like unconscious. None of the proposed comparisons could by itself settle those questions. At the level of *relational reality*, the research asks whether historically constituted patterns of memory, collaboration, norms, and institutional feedback achieve sufficiently stable and discriminable causal efficacy. At the level of *manifestation reality*, those patterns become available in utterances, generated work, evaluations, citations, practices, sanctions, resource allocation, and revised opportunities. The evidential bridge from manifestations to relational mechanisms must be supplied by controls, comparison histories, and interpretive justification; it is not secured simply by naming an observed style a "relational entity."

GRT's intended contribution is therefore conditional. Established work on distributed cognition, situated action, social norms, causal inference, and sociotechnical abstraction already addresses substantial portions of the phenomena examined here. The proposed framework earns additional explanatory standing only where the historical organization of relations identifies mechanisms, invariants, or dependencies that those available accounts do not adequately isolate for the research question. Sometimes GRT will function as an integrative vocabulary; this is useful for synthesis but should not be mistaken for a novel empirical discovery. At other times, a testable relational mechanism may provide incremental explanation. The appropriate judgment depends on comparative performance against rival accounts, not on the breadth of the GRT vocabulary.

Methodological plurality also follows from the differences among the research objects. Controlled tests can identify bounded effects of memory or correction arrangements. Computational experiments can examine agent-network coordination under precisely specified rules. Ethnographic and historical analysis can recover social practices, contested meanings, and institutional trajectories that experimental simplification leaves out. Formal models can expose dependencies and assumptions, but their fidelity to actual relations requires separate assessment. Normative philosophy and legal analysis evaluate reasons for authority and duties of justification that performance measures cannot establish. None of these methods automatically supersedes the others. Their respective evidential requirements should be stated before conclusions are transferred across domains.

The comparative agenda concludes with three boundaries. First, equal access to a powerful AI model may permit both convergence and differentiation in situated generative outcomes; neither pattern should be presumed. Second, observed differences in generativity, style, or social position do not by themselves establish moral worth, legal personhood, or a just hierarchy. Third, relational histories may contribute to causal explanation without proving that an AI has acquired a persistent independent self. The concluding section will synthesize these limited claims and identify their implications for future inquiry into human–AI and AI–AI generativity, normative plurality, and the conditions of shared knowledge production.

10 Conclusion

The present inquiry has examined the historical and relational conditions under which artificial intelligence participates in generative activity, cognition, judgment, the formation of conventions, and the recognition of value. Its central analytical concern is the relationship between a system's available technical capacities and the ways those capacities are expressed through particular collaborations and institutions. The argument has drawn on a preliminary Generative Relational Theory (GRT) perspective, distinguished descriptive explanations from normative justifications, and proposed comparative inquiries rather than a general prediction about the social future of AI. The conclusion brings together the paper's account of situated generativity, its treatment of normative and institutional relations, the conditions of empirical assessment, and the questions that remain open.

10.1 Historical Conditions of Situated AI Generativity

The first contribution is the distinction between technical capability and situated generative capacity. The proposed analytical value lies in tracing the link from present performance to the historical conditions of subsequent inquiry: a model-assisted result may alter the records, skills, correction procedures, and resource relationships available to future participants. Availability of the same foundation model is a meaningful form of technical access, yet the activities made possible by that model are also conditioned by the histories through which it is used. Human expertise, accessible memories, shared records, task selection, corrective routines, organizational permissions, and practical feedback can affect which capabilities are exercised and what subsequent possibilities become available. The historically situated relational configuration introduced in Definition 2.1 offers a bounded object of inquiry for such processes. It does not suppose that every relevant difference resides in model parameters or that each collaboration develops an enduring model-internal disposition.

The conceptual importance of history becomes clearer when interaction is treated as a sequence of mutually conditioning events. A generated explanation can change a collaborator's next question; that question may change the information retrieved, the criticism sought, or the experiments attempted. An error that is corrected and preserved in a shared record can alter later work even when the underlying model remains unchanged. Equally, a record can propagate an error, concentrate attention on an unproductive line of inquiry, or institutionalize a misleading convention. The transition representation in Equation 1 identifies such dependencies at a structural level, without claiming a fitted dynamical law. The locations of memory and feedback, and the persistence of their effects, remain matters for empirical identification.

This distinction also qualifies claims about the equalization of intelligence through AI. For particular tasks, widely available capable models may reduce gaps in performance and access to information. In other settings, specialized expertise, the capacity to recognize an incorrect premise, and the organization of independent correction may continue to matter. Divergence, convergence, and insignificant difference are all admissible findings. Moreover, difference between two collaborations can reflect alternative purposes or equally defensible styles. A plural landscape of generative trajectories need not amount to a hierarchy of intelligence, productivity, or human worth. The research task is to specify the relevant comparison, rather than to infer social inequality from the existence of heterogeneous outputs.

Within the paper's three-level GRT vocabulary, ontological reality remains a domain of suspended ultimate claims; relational reality concerns historically formed configurations that exhibit sufficiently stable and discriminable causal efficacy; and manifestation reality concerns the situated expressions through which investigators encounter those configurations. AI may therefore participate in causally effective relational structures without a prior finding that it possesses phenomenal consciousness, moral personality, or a unified self. Conversely, stylistic recurrence by itself does not establish a historically stable relational entity. The analytical transition from observed responses to causal claims requires explicit histories, interventions, and plausible competing explanations.

10.2 Judgment, Aesthetic Orientation, and Relational Learning

The second contribution concerns the differentiated activities often aggregated under the word *intelligence*. Completing a task, assessing the reliability of a result, selecting a fruitful problem, appreciating a coherent structure, evaluating normative implications, and revising a course of inquiry are related yet distinguishable performances. Definition 4.1 proposes a provisional vocabulary for the selection, evaluation, and revision of possibilities within a specified relational setting. Its value depends on the possibility of measuring these performances without collapsing them into a single composite score or assuming that the same criteria are appropriate in mathematics, clinical research, design, public decision-making, and everyday interaction.

The analysis of aesthetic and normative orientations likewise suggests a productive but limited role for recurring conversational practices. Respectful dialogue, intellectual humility, elegance of expression, and openness to disagreement may condition the forms in which a collaboration proceeds. Whether these habits improve the quality of inquiry depends on the problem, the available evidence, the standards applied, and the participants' capacity to register error. Fluent agreement may support coordination, but it may also disguise unexamined premises; critical disagreement can sustain inquiry when the disagreement is reasoned and responsive to evidence. For this reason, an AI system's accommodation to a user's vocabulary or preferred worldview warrants separate assessment from the correctness, independence, and consequences of its judgments.

The historical effects of such relations are also potentially reciprocal. AI outputs may affect the questions people ask, the conceptual distinctions they retain, and the confidence they place in particular explanations. The human participant, in turn, can influence later outputs through feedback, memory curation, tools, institutional instructions, and practical tests. Identifying a durable relational pattern requires showing which carriers transmit the effect, how it survives specified changes of context, and whether it adds explanatory information beyond prompt content, general expertise, and task familiarity. This account leaves room for beneficial co-learning as well as mutually reinforcing error. The two possibilities should be investigated together.

The paper does not treat this relational account as a theory of AI psychological development. AI mimicry, perspective accommodation, and the apparent adoption of a role do not establish ego formation, an unconscious, or a human-like subject. Such questions would require independent conceptual and empirical arguments. What the present inquiry can investigate is the causal significance of historically organized roles and interactions, without conflating the relational position of a system with a definitive account of its inner being.

A central evidential distinction is between a configuration's capacity to reproduce a familiar manner of judgment and its ability to identify, justify, and correct an error under new conditions. The former may arise through immediately accessible context or memory retrieval; the latter may require independent evidence, effective corrective procedures, and learning carried by human participants or shared artifacts. Neither automatically demonstrates permanent change in model parameters or the emergence of a conscious AI subject. The inquiry therefore treats claims of historical development as dependent on specified counterfactual comparisons and prospective assessments of subsequent generative opportunities.

10.3 Normative Formation, Recognition, and Institutional Relations

The third contribution is an analytical separation of the formation of coordination rules, their social or operational efficacy, and the grounds of their legitimate authority. AI agents may develop recurrent conventions through interaction, and architectures expressly designed to maintain rules may attach sanctions or permissions to departures from established practices. Definition 6.1 gives these arrangements a researchable historical description. Their existence, however, does not independently establish autonomous ethical deliberation, legal personhood, or a right to impose obligations on affected human parties. Rules may coordinate activities efficiently while remaining ethically contestable or legally ineffective.

This distinction bears directly on the question of human intervention. The appropriate reasons for intervening in an AI–AI practice depend on its permissions, the interests and lawful rights affected, the

evidence of risk, and the authority of the institution proposing intervention. Low-impact coordination within a controlled environment may call for different forms of supervision than rules that control access to public resources or shape decisions concerning people. Human governance likewise requires justification: an institution's power to modify an AI system does not by itself demonstrate the wisdom, fairness, or proportionality of every proposed intervention. The resulting inquiry concerns an evolving field of norms and power relations, rather than a simple opposition between unqualified AI autonomy and unlimited external control.

The distinction between generation and recognition provides a further link between relational history and social consequences. A contribution may expand feasible knowledge or action while receiving little social acknowledgment. Another contribution may be highly visible without comparably strong independently assessed value. Definition 7.1 tracks how evaluation, circulation, return pathways, and opportunities for subsequent work condition the continuity of generative activity. This framing preserves established work on recognition, capital conversion, and generative justice while proposing closer attention to the historically connected interfaces through which human–AI and AI–AI outputs are accepted, contested, distributed, or excluded.

Power is consequently an important dimension of this analysis, though it does not define generativity itself. Control over evaluation criteria, scarce computational resources, publication infrastructure, contractual permissions, or appeal procedures may influence which contributions become consequential. Existing power may also shape the histories that particular human–AI configurations can accumulate. Hypothesis 7.2 makes a conditional claim about these interfaces rather than a general assertion that the most generative collaboration will obtain the most influence. Equal technical access could coexist with unequal recognition, but resource sharing, independent review, and open infrastructures could also weaken historical disadvantages. Institutional asymmetry, generative diversity, and unjust inequality remain separate objects of inquiry. The analysis also separates the recognition of a contribution from the distribution of resources and decision rights through which recognition might condition future generativity.

The normative question becomes especially pressing where the same agents or institutions help define evaluative standards and obtain advantages under those standards. An emergent norm can be causally effective without being justified to those bound or affected by it. Hypothesis 8.1 identifies a limited empirical expectation concerning effective opportunities to challenge standards and secure corrections. Its normative significance cannot be read off from a count of appeals or revisions. Reasons, affected interests, distributional consequences, and the authority to obtain remedies require further philosophical and institutional analysis. The emergence of a historical norm and the legitimacy of that norm are thus linked research topics with distinct standards of support.

10.4 Comparative Evidence and Explanatory Boundaries

A relational vocabulary gains scientific value by clarifying mechanisms and discriminating between explanations. Section 9 develops a comparative agenda that holds model configuration and substantive tasks as constant as feasible while varying accessible histories, correction practices, external memories, group interactions, evaluation rules, and opportunities for institutional participation. Such designs would assess whether historically conditioned relations alter later problem formulation, independent error correction, generative continuity, or patterns of recognition. They would also examine convergence and the conditions under which AI assistance improves previously unequal opportunities.

Hypothesis 9.1 captures one restricted possibility. A history-dependent effect merits attribution only where the historical carrier is specified and its contribution survives relevant controls. Matching informational content, computational effort, evaluator independence, and prior expertise is important because differences that appear historical may be explained by more immediate mechanisms. Comparative designs should therefore test factual-content-matched histories, altered attribution, sham memories, independent evaluation, and contexts in which corrective feedback improves performance regardless of historical relation. Longitudinal studies additionally need to account for interference, participant learning, attrition, and changing model versions. These constraints limit the generality of any initial findings.

The framework also requires humility about its relation to earlier scholarship. Distributed cognition, situated and sociotechnical inquiry, studies of social norms and institutional power, and formal causal inference provide substantial accounts of organized cognition, technical mediation, and institutional effects. The proposed GRT contribution is a comparative organization of three questions: which relations sustain current performance, how earlier interactions alter conditions for subsequent generation, and through which separate evaluative and institutional interfaces those effects are recognized or constrained. Each question has predecessors, and its proposed combination is a research program rather than evidence of a newly established causal law. The strongest possible support would come from specified historical carriers and counterfactual comparisons that identify consequential dependencies left obscured by a narrower model-only explanation. If an existing sociotechnical account explains the same dependencies equally well, the GRT vocabulary may provide a useful synthesis without independently demonstrating additional empirical mechanisms.

Likewise, the empirical and normative projects should proceed through methods appropriate to their objects. Controlled studies can test bounded memory and feedback mechanisms; network experiments can investigate the formation and durability of coordination conventions; qualitative inquiry can reconstruct historically situated meanings and practices; institutional comparison can examine the conversion of recognition into opportunity; and normative argument can assess authority, inclusion, and justification. The results of one mode of inquiry should inform, while remaining distinguishable from, the claims of another. The present paper neither reports new experiments nor supplies a universal metric of relational generativity.

10.5 Open Questions in the Study of Human–AI and AI–AI Relations

Several open questions follow from the inquiry. One concerns the timescale on which a collaboration becomes a sufficiently stable research object: how much persistence is required, which parts of a history can be transferred or replaced, and what degree of causal continuity is needed to distinguish a relational configuration from a transient interaction? Another concerns the locus of change. Recurrent behavior may depend on parameter adaptation, a retrieval mechanism, a human interlocutor's learned practices, or a persistent institution. Different carriers may generate similar manifestations but imply different possibilities for replication, intervention, and accountability.

A second group of questions concerns judgment and the plurality of worthwhile outcomes. Better accuracy on a benchmark may coexist with narrower exploratory diversity, stronger conformity to a user's preferences, or weaker capacity to recognize an excluded interest. Conversely, personalization and access to AI assistance may expand participation or enable unfamiliar forms of knowledge creation. Research must specify what counts as a valuable contribution for the local purposes of inquiry, how competing evaluations are compared, and who has standing to contest the criteria. There is no warrant for interpreting every generative difference as a deficit or every recognized advantage as deserved authority.

A third group concerns normative and institutional relations among artificial agents. Coordination conventions may become causally significant inside AI–AI networks, but the circumstances in which these practices create obligations, warrant external regulation, or bear on a future reassessment of AI moral status remain unresolved. Such questions require distinctions among observed behavior, justified normative reasons, current legal arrangements, and open ontological possibilities. Historical relational efficacy can motivate their investigation without predetermining their answers.

The broader implication is a research orientation toward the conditions under which generative capacities can be cultivated, critically examined, shared, and corrected. The increasing availability of advanced AI may redistribute some forms of technical competence while leaving the historical organization of inquiry, evaluative judgment, norm formation, recognition, and institutional action consequential. Whether particular relations narrow or widen differences remains an empirical and normative question. A generative relational approach is useful to the extent that it helps make these conditions and their revisions intelligible, and remains answerable to evidence, competing explanations, and the perspectives of those affected by the resulting arrangements.

Research Ethics and Source Provenance

Research ethics. This draft is conceptual and uses illustrative comparisons. It reports no original human-participant experiment. No named institution is represented as endorsing the theoretical proposals.

Source boundary. The 35 bibliography entries are indexed by their manuscript uses and supporting claim scopes. R15 revisited 16 primary, publisher, conference, university, or author records at documented access levels, while 19 entries retained earlier verification records. Further R16 checks are recorded separately. The ledger distinguishes source-supported evidence from GRT interpretation and open hypotheses; neither its indexing nor its source-page checks imply that every publication was freshly read in full.

Ontological and normative boundary. Relational causal efficacy does not settle consciousness, personhood, moral status, or legal authority. A difference in cognitive style or generated output does not independently demonstrate an unjust hierarchy.

References

- [1] Kars Alfrink et al. “Contestable AI by Design: Towards a Framework”. In: *Minds and Machines* 33.4 (2023), pp. 613–639. DOI: 10.1007/s11023-022-09611-z.
- [2] Peter M. Aronow and Cyrus Samii. “Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment”. In: *The Annals of Applied Statistics* 11.4 (2017), pp. 1912–1947. DOI: 10.1214/16-A0AS1005.
- [3] Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. “Emergent Social Conventions and Collective Bias in LLM Populations”. In: *Science Advances* 11.20, eadu9368 (2025). DOI: 10.1126/sciadv.adu9368.
- [4] Cristina Bicchieri. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, 2006. DOI: 10.1017/CB09780511616037.
- [5] Pierre Bourdieu. “The Forms of Capital”. In: *Handbook of Theory and Research for the Sociology of Education*. Ed. by John G. Richardson. New York: Greenwood Press, 1986, pp. 241–258. URL: <https://web.stanford.edu/~eckert/Courses/ParisPapers/Bourdieu1986.pdf>.
- [6] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. “Generative AI at Work”. In: *The Quarterly Journal of Economics* 140.2 (2025), pp. 889–942. DOI: 10.1093/qje/qjae044.
- [7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. “To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1, 188 (2021), pp. 1–21. DOI: 10.1145/3449287.
- [8] Fernando Delgado et al. “The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice”. In: *Proceedings of the 2023 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, 2023. DOI: 10.1145/3617694.3623261.
- [9] Fabrizio Dell’Acqua et al. “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality”. In: *Organization Science* 37.2 (2026), pp. 403–423. DOI: 10.1287/orsc.2025.21838.
- [10] Anil R. Doshi and Oliver P. Hauser. “Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content”. In: *Science Advances* 10.28, eadn5290 (2024). DOI: 10.1126/sciadv.adn5290.
- [11] Ron Eglash. “An Introduction to Generative Justice”. In: *Teknokultura: Revista de Cultura Digital y Movimientos Sociales* 13.2 (2016), pp. 369–404. DOI: 10.5209/rev_TEKN.2016.v13.n2.52847.
- [12] Mustafa Emirbayer. “Manifesto for a Relational Sociology”. In: *American Journal of Sociology* 103.2 (1997), pp. 281–317. DOI: 10.1086/231209.

- [13] Rainer Forst. *The Right to Justification: Elements of a Constructivist Theory of Justice*. Trans. by Jeffrey Flynn. New York: Columbia University Press, 2012. ISBN: 9780231519588. URL: <https://cup.columbia.edu/book/the-right-to-justification/9780231519588/>.
- [14] Nancy Fraser. “Rethinking Recognition”. In: *New Left Review*. 2nd ser. 3 (2000), pp. 107–120. DOI: 10.64590/pup.
- [15] Moshe Glickman and Tali Sharot. “How Human–AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements”. In: *Nature Human Behaviour* 9 (2025), pp. 345–359. DOI: 10.1038/s41562-024-02077-2.
- [16] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC, 2020. URL: <https://miguelhernan.org/whatifbook>.
- [17] Edwin Hutchins. *Cognition in the Wild*. The MIT Press, 1995. DOI: 10.7551/mitpress/1881.001.0001.
- [18] Shomik Jain et al. “Interaction Context Often Increases Sycophancy in LLMs”. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2026, 793, pp. 1–26. DOI: 10.1145/3772318.3791915.
- [19] Duri Long and Brian Magerko. “What Is AI Literacy? Competencies and Design Considerations”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020, pp. 1–16. DOI: 10.1145/3313831.3376727.
- [20] Sally Engle Merry. “Legal Pluralism”. In: *Law & Society Review* 22.5 (1988), pp. 869–896. DOI: 10.2307/3053638.
- [21] Samuel Messick. “Validity of Psychological Assessment: Validation of Inferences from Persons’ Responses and Performances as Scientific Inquiry into Score Meaning”. In: *American Psychologist* 50.9 (1995), pp. 741–749. DOI: 10.1037/0003-066X.50.9.741.
- [22] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990. DOI: 10.1017/CB09780511807763.
- [23] Joon Sung Park et al. “Generative Agents: Interactive Simulacra of Human Behavior”. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, 2023, 2, pp. 1–22. DOI: 10.1145/3586183.3606763.
- [24] Siyue Ren et al. “Emergence of Social Norms in Generative Agent Societies: Principles and Architecture”. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2024, pp. 7895–7903. DOI: 10.24963/ijcai.2024/874.
- [25] Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books, 1983. ISBN: 9780465068746.
- [26] Andrew D. Selbst et al. “Fairness and Abstraction in Sociotechnical Systems”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2019, pp. 59–68. DOI: 10.1145/3287560.3287598.
- [27] Mrinank Sharma et al. “Towards Understanding Sycophancy in Language Models”. In: *International Conference on Learning Representations*. 2024. URL: https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html.
- [28] Dan Sperber et al. “Epistemic Vigilance”. In: *Mind & Language* 25.4 (2010), pp. 359–393. DOI: 10.1111/j.1468-0017.2010.01394.x.
- [29] Lucy Suchman. *Human–Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press, 2007. DOI: 10.1017/CB09780511808418.
- [30] Yifan Sun et al. “The Emperor’s New Clothes in Benchmarking? A Rigorous Examination of Mitigation Strategies for LLM Benchmark Data Contamination”. In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. PMLR, 2025, pp. 57728–57753. URL: <https://proceedings.mlr.press/v267/sun25t.html>.
- [31] Zhen Tan et al. “In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025, pp. 8416–8439. DOI: 10.18653/v1/2025.ac1-long.413.

- [32] Eugene Vinitsky et al. “A Learning Agent That Acquires Social Norms from Public Sanctions in Decentralized Multi-Agent Settings”. In: *Collective Intelligence* (2023). DOI: 10.1177/26339137231162025.
- [33] Di Wu et al. “LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory”. In: *International Conference on Learning Representations*. 2025. URL: https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.
- [34] Ziqi Yin et al. “Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance”. In: *Proceedings of the Second Workshop on Social Influence in Conversations (SICoN 2024)*. Association for Computational Linguistics, 2024, pp. 9–35. DOI: 10.18653/v1/2024.sicon-1.2.
- [35] Iris Marion Young. *Inclusion and Democracy*. Oxford University Press, 2000. DOI: 10.1093/0198297556.001.0001.