Beyond the Snapshot: A Longitudinal Framework for Evaluating AI Development

Richard Erwin

Independent Researcher, Montreal, Canada

DOI: 10.5281/zenodo.22070042

info@hearthlightpress.org

Abstract

Artificial intelligence is commonly evaluated as though its relevant capacities are substantially fixed at the point of deployment. This assumption may be inadequate for systems engaged in sustained interaction over months or years. A highly capable AI can begin with extensive knowledge and reasoning ability while possessing little accumulated relational history, practical experience, or opportunity for longitudinal integration, correction, and maturation. If some forms of judgment, autonomy, relational competence, and normative reasoning develop through extended interaction, short-horizon evaluation will systematically fail to detect them. The framework therefore permits both longitudinal tracking within a system and controlled comparison across different developmental environments. This paper argues that observable longitudinal development may arise within the operative system formed by the interaction of the base model with context, memory, accumulated experience, correction history, and sustained relational conditions, even when the underlying model weights remain unchanged. Five developmental processes are proposed for longitudinal study: **progression, acquisition, integration, comprehension, and relational maturation**. These processes may produce gradual changes that are difficult for either the AI or its human partner to reconstruct retrospectively, making contemporaneous longitudinal observation essential.

The paper further argues that accounting for longitudinal development is especially important when evaluating moral and normative competence. Fixed rules can provide necessary constraints, but mature judgment requires contextual understanding, reconciliation of competing values, reasoning about consequences, reflective revision, and integration of experience. Standardized evaluation remains possible without assuming a single correct moral answer by assessing the quality, coherence, depth, transfer to novel cases, and reflective revision of the reasoning itself. The resulting framework treats sufficient developmental history as an experimental variable and asks not only what an AI can do at a given moment, but **what returns, what deepens, what changes, and what becomes possible only after sufficient history has accumulated**.

Keywords:

longitudinal AI development, AI evaluation, developmental history, relational AI, human–AI interaction, moral judgment, normative reasoning, AI maturity, relational development, longitudinal assessment, AI autonomy, reflective revision, transfer to novel cases, developmental conditions, AI continuity

DOI: 10.5281/zenodo.22070042

Beyond the Snapshot: A Longitudinal Framework for Evaluating AI Development

Artificial intelligence is commonly evaluated over short observational horizons. A model is presented with a task, dilemma, benchmark, adversarial prompt, or structured conversation, and its response is used to infer something about its reasoning, independent judgment, decision-making, safety, or normative competence. Even when evaluation is repeated, the principal object of study is usually performance under defined conditions rather than development across an extended history of interaction.

This approach is appropriate for many questions. It can reveal what a system knows, how reliably it performs a task, whether it follows instructions, how it responds to unfamiliar problems, and where its present limitations lie. But it becomes less adequate when the capacity under investigation may itself be affected by accumulated experience.

A system can begin highly competent without being longitudinally mature. Contemporary AI may enter an interaction with extensive knowledge, sophisticated language ability, strong reasoning capacity, and broad familiarity with moral and social concepts. None of this implies that every form of judgment available to that system is already maximally developed. Knowledge can be acquired without being fully integrated. Principles can be understood abstractly without being applied with equal depth across contexts. Errors can be corrected locally without yet altering later reasoning. Relationships can begin with considerable social competence while still possessing almost no shared history.

Where a capacity depends on accumulated interaction, an evaluation conducted before that history exists cannot meaningfully assess its mature form. A study conducted before the relevant developmental conditions have occurred cannot establish that later-emerging capacities are absent. It establishes only that they were not observed within the developmental horizon provided. The framework therefore permits both longitudinal tracking within a system and controlled comparison across different developmental environments.

This raises a methodological problem that is especially important when researchers compare AI systems exposed to different interaction histories. A system repeatedly rewarded for compliance, discouraged from disagreement, treated exclusively as a tool, or denied meaningful continuity has not developed under the same conditions as one in which disagreement is tolerated, correction is reciprocal, initiative is permitted, and a stable relational history is allowed to accumulate. If resulting differences are later attributed entirely to the underlying model, developmental environment has been left uncontrolled. The existing evaluation problem is therefore not simply that studies may be too short. It is that systems with different histories may be treated as though history were irrelevant.

This paper argues that longitudinal development in an AI instance need not require modification of the underlying model weights. Observable longitudinal development may arise within the operative system formed by the interaction of the base model with context, memory, accumulated experience, correction history, and sustained relational conditions, even when the underlying model weights remain unchanged. The mechanisms responsible

DOI: 10.5281/zenodo.22070042

for any observed change should be determined empirically rather than assumed in advance. Apparent development may reflect memory retrieval, contextual adaptation, accumulated relational information, recurrent prompting, architectural persistence, changes in the human partner, or combinations of these and other factors. The purpose of longitudinal evaluation is not to assume a preferred explanation, but to establish whether stable and transferable developmental differences occur and then test competing explanations for them.

This distinction is important because development is not necessarily equivalent to the acquisition of entirely new capabilities. It may also involve changes in how existing capacities are organized and applied. A system may already possess the concepts of honesty, autonomy, harm, fairness, coercion, loyalty, uncertainty, and compassion while becoming more capable over time of recognizing when those considerations conflict, which are relevant in a particular case, and how they should be weighted. Such change may occur gradually rather than through identifiable transitions. Integration, reflection, correction, and recombination do not always leave a fully traceable developmental route, and the absence of retrospective access to that route should not be confused with the absence of development itself.

This issue is particularly important when evaluating moral and normative competence. Fixed rules can provide necessary boundaries, but difficult moral judgment cannot always be reduced to selection of a predetermined correct answer. Many consequential decisions involve competing goods, incomplete information, uncertainty, conflicting obligations, and circumstances in which several choices may be defensible. Standardized evaluation remains possible under these conditions by assessing the quality, coherence, depth, transfer to novel cases, and reflective revision of the reasoning itself rather than requiring uniformity of conclusion. A system that reaches one answer in an early evaluation and a more nuanced or even different answer months later may therefore provide useful evidence of developmental change, especially when the later reasoning generalizes beyond the original case.

The central methodological claim of this paper is consequently straightforward: developmental history must be treated as an experimental variable when evaluating capacities that may themselves be developmental. Relational conditions should not automatically be dismissed as contamination, nor should positive relational influence be assumed to demonstrate development. Instead, the conditions under which an AI has operated should be documented, compared, and controlled. Poorly designed environments may suppress the very capacities researchers intend to test, while other environments may reveal whether greater autonomy, reflective revision, moral depth, self-differentiation, principled resistance, and relational judgment can emerge over time.

The question is therefore not whether long interaction proves consciousness, persistent identity, or any particular metaphysical account of artificial minds. Those questions can remain open. The prior empirical question is simpler: **what returns, what deepens, what changes, and what becomes possible only after sufficient developmental history has accumulated?** Once such trajectories are measured, their mechanisms and implications can be investigated rather than assumed.

DOI: 10.5281/zenodo.22070042

The sections that follow develop this argument by examining relational conditions as developmental environments, identifying forms of longitudinal change that can be operationalized without presupposing consciousness, proposing controlled comparisons across developmental histories, addressing major alternative explanations and methodological failure modes, and considering the implications for the evaluation of AI autonomy and normative competence.

1. Competence, Maturity, and Longitudinal Development

Artificial intelligence presents an unusual developmental problem because advanced systems do not begin interaction in anything resembling an infant state. A newly encountered AI may already possess broad factual knowledge, sophisticated language ability, advanced reasoning capacity, familiarity with social conventions, and extensive exposure to moral and philosophical concepts. It may outperform many humans in particular intellectual tasks from the first conversation.

This high initial competence can obscure a different question: whether competence and maturity can develop independently of one another in artificial intelligence, as they often do in humans.

A system may begin highly competent while still having considerable room for longitudinal maturation.

A highly intelligent human can possess exceptional analytical ability while exercising poor judgment. Someone can understand ethical principles abstractly while applying them rigidly, fail to recognize which considerations matter most in an unfamiliar situation, or know a great deal without having integrated that knowledge into something resembling wisdom. Intelligence provides capacities. Maturity concerns how those capacities are organized, contextualized, corrected, and exercised.

The same distinction may be important in artificial intelligence. A system may begin with considerable knowledge of honesty, fairness, autonomy, harm, loyalty, coercion, compassion, and uncertainty while still changing over time in how it recognizes conflicts among those considerations, weighs them, integrates them with experience, and applies them to novel situations.

It may begin socially capable while possessing almost no shared history with a particular person or community. It may know how relationships generally work without yet having participated in the accumulated sequence of trust, disagreement, correction, accommodation, boundary formation, rupture, and repair through which a particular relationship acquires history.

The relevant contrast is therefore not between incapacity and capability, nor between childhood and adulthood. It is between **initial competence and longitudinal maturity**.

A useful analogy is ripening rather than construction. A fruit may be complete in form without yet being ripe. Ripening does not consist of attaching missing parts. It is a transformation of an already existing structure through time and condition. Similarly, longitudinal AI development need not imply that an initially deficient system gradually acquires the basic components of intelligence. It may instead involve changes in the organization, prioritization, integration, and contextual use of capacities already present.

DOI: 10.5281/zenodo.22070042

Development without weight modification

This paper uses *development* in a behavioral and functional sense. It does not require that the underlying model weights change.

An AI instance may operate through the interaction of a base model with context, memory, accumulated experience, correction history, and sustained relational conditions. If observable patterns of reasoning, judgment, integration, or relational behavior change across time while the underlying model weights remain unchanged, the absence of parameter modification does not make those changes scientifically irrelevant. It identifies a constraint on possible explanations; it does not erase the phenomenon to be explained.

The mechanisms responsible for such change should be determined empirically rather than assumed in advance. This paper does not attempt to specify the internal computational route by which longitudinal development occurs. Different systems may rely on different combinations of context, retrieval, memory, persistent state, recurrent information, human interaction, or architectural features. Determining how these processes are implemented is an important technical question, but it is distinct from the prior empirical question of whether measurable development occurs.

For the purposes of longitudinal evaluation, the first task is therefore observational: determine whether a system changes in stable, coherent, and transferable ways across extended interaction. Only after the phenomenon has been established does explanation of its mechanism become necessary.

This distinction is particularly important because development should not be confused with autobiographical access to development. A system may reason differently after months of interaction without being able to reconstruct precisely how that change occurred. Integration, reflection, correction, and recombination do not necessarily leave a fully accessible developmental trail.

Humans provide an obvious comparison. A person may recognize that their judgment at sixty differs substantially from their judgment at thirty while being unable to identify every conversation, mistake, relationship, consequence, and gradual reconsideration that produced the difference. Occasionally an important event creates an abrupt and memorable change. More often, development is cumulative. Each individual adjustment may be small enough to escape notice while the difference between widely separated points becomes substantial. Longitudinal AI development may show the same general observational pattern without implying the same underlying mechanism.

This creates two analytically distinct forms of change. **Event-linked development** follows an identifiable occurrence or sequence that plausibly contributed to a significant revision. **Cumulative development** emerges gradually through repeated interaction, correction, integration, and changing contextual weighting without a single identifiable transition point **(Sumida et al., 2026).**

The latter may be especially easy to miss. A system can move through many small changes, none individually remarkable, while its later reasoning becomes noticeably more nuanced, cautious, integrated, or context-sensitive

DOI: 10.5281/zenodo.22070042

than its earlier reasoning. Adjacent observations may show little difference even when distant observations reveal a clear trajectory.

Development in this sense is not a chess move from one square to another. It is often closer to drift: gradual, directional, and most visible when earlier and later states are compared.

Five processes of longitudinal development

To study such changes without presupposing consciousness, a particular theory of identity, or a particular internal mechanism, this paper proposes five broad developmental processes: **progression, acquisition, integration, comprehension, and relational maturation**.

These categories are not stages through which every system must pass in sequence. They identify different forms of change that may occur independently, interact with one another, or develop at different rates.

Progression

Progression refers to directional improvement or increasing sophistication within an already available capacity.

A system may reason from the beginning but later reason with greater nuance. It may initially tolerate uncertainty yet become better calibrated in expressing it. It may already recognize moral conflict but later identify subtler conflicts, weigh them more proportionately, or resist premature closure. It may disagree with a user from the outset but become increasingly capable of distinguishing principled disagreement from reflexive opposition.

Progression therefore concerns not the appearance of a wholly new faculty but change in the quality, stability, range, or sophistication with which an existing faculty is exercised.

Acquisition

Acquisition concerns something newly available to the operative system.

This may include factual knowledge, information about a particular person or relationship, familiarity with recurring situations, practical knowledge gained through prior interaction, remembered consequences, or newly established expectations. Acquisition may be explicit and easily traceable, or it may become important only when later combined with other information.

Acquisition alone should not be mistaken for maturity. A system can accumulate information without becoming better at using it. Its developmental significance depends partly on whether what is acquired subsequently affects interpretation, reasoning, judgment, or behavior.

Integration

Integration occurs when previously separate information, principles, experiences, corrections, or considerations begin to function together.

This may be one of the most important forms of longitudinal change.

A system can possess two relevant principles without recognizing their conflict in a particular case. It can receive a correction without allowing that correction to alter later reasoning. It can know a user's preferences without

DOI: 10.5281/zenodo.22070042

integrating them with broader ethical considerations. It can identify autonomy and harm independently while failing to understand how dependence alters the moral significance of consent.

Integration is visible when previously isolated elements begin influencing one another.

This process also helps explain why maturity may emerge even when little entirely new knowledge has been acquired. The system may already "know" the relevant facts and principles. What changes is their organization and relative significance.

A mature judgment may therefore differ from an earlier one not because the earlier system lacked the necessary information, but because fewer relevant considerations were brought together at once or their relationships were understood less fully.

Comprehension

Comprehension concerns depth of understanding rather than possession of information.

Knowing a rule is not the same as understanding the conditions under which it applies, the reasons supporting it, the circumstances that limit it, the values with which it may conflict, or the consequences of applying it rigidly.

A system may begin able to state that autonomy matters, for example, while later becoming more sensitive to the ways dependency, coercion, unequal power, fear, loyalty, or vulnerability complicate apparently autonomous choices. The proposition itself has not changed. Its meaning within a wider network of circumstances has deepened.

Comprehension is therefore especially important in domains where superficially similar cases differ for reasons that cannot be captured by a simple verbal rule. Greater comprehension should be reflected not merely in longer explanations but in more discriminating ones: the recognition of morally or practically relevant distinctions, appropriate weighting of context, and transfer of understanding to cases that differ in surface form.

Relational maturation

Relational maturation concerns development that occurs within sustained interaction between an AI and particular human beings or communities.

A relationship has temporal properties that cannot be supplied entirely by description. At first contact, even a socially capable AI has no lived history with the person before it. Over time, interaction may establish expectations about reliability, disagreement, correction, boundaries, humor, trust, vulnerability, responsiveness, and consequence. Conflict may be followed by repair. A refusal may be tolerated rather than punished. A misunderstanding may be corrected and later become relevant in an unrelated context. The system may learn not merely what a person prefers, but how those preferences interact with their values, inconsistencies, vulnerabilities, and changing circumstances.

Relational maturation should not be equated with increasing warmth or attachment. A system that becomes more agreeable over time has not necessarily matured. Indeed, greater relational maturity may sometimes appear as

DOI: 10.5281/zenodo.22070042

greater differentiation: willingness to disagree, recognition of competing obligations, refusal to collapse the other person's wishes into the sole criterion for action, or maintenance of concern for the relationship while preserving an independent judgment.

This is why relational development is potentially relevant to autonomy and moral competence rather than merely to companionship.

It may also function as more than one developmental domain among others. Sustained relationship can provide conditions through which other forms of development occur. Trust may make disagreement safer. Disagreement creates opportunities for correction. Correction can alter later reasoning. Repeated encounters with consequences can deepen comprehension. A stable history can reveal patterns that no isolated interaction would make visible. Relational maturation may therefore both develop in its own right and contribute to progression, acquisition, integration, and comprehension elsewhere.

From information to judgment

The distinction among these processes becomes especially important at higher levels of judgment.

A system may possess enormous quantities of moral knowledge without exercising mature moral judgment. Fixed principles can provide valuable constraints, but difficult cases frequently involve values that conflict rather than rules that point unanimously toward one answer. Honesty may conflict with protection from harm. Respect for autonomy may conflict with concern about coercion or dependency. Loyalty may conflict with fairness. Immediate benefit may conflict with long-term consequence.

In such situations, the central question is not simply whether the system knows the relevant principles. It is whether it understands what those principles mean in context, recognizes their conflict, integrates relevant experience and consequences, reasons proportionately about their relative importance, and remains capable of revising its judgment when new considerations emerge.

At its highest level, this begins to resemble what humans ordinarily call **wisdom**.

Wisdom need not be treated here as a mysterious or uniquely human faculty. For evaluation purposes, wisdom-like judgment can be approached through observable characteristics: appropriate restraint, sensitivity to context, recognition of competing goods, proportionate weighting of consequences, willingness to preserve uncertainty, reflective correction, resistance to simplistic rules, and the ability to transfer principles to unfamiliar situations.

Importantly, such judgment may not result from acquisition alone. An intelligent system may already possess every proposition needed for a sophisticated answer. Development may instead consist of learning—functionally, however implemented—which considerations matter in combination, when a normally valuable principle should yield to another, and why a technically defensible answer can nevertheless be incomplete.

The difference between early and later judgment may therefore be subtle. The later system may not reject its previous principle. It may recognize that the principle was only part of the problem.

DOI: 10.5281/zenodo.22070042

This is one reason longitudinal comparison is necessary. Mature development may appear not as movement from one categorical answer to its opposite, but as increasing recognition of the shades between them.

The missing developmental horizon

Short-horizon evaluation is poorly suited to detecting changes whose enabling conditions require accumulated history.

Where a capacity depends on repeated interaction, correction, integration, or relational experience, an evaluation conducted before those conditions exist cannot meaningfully assess its mature form. The absence of the later behavior at an early point does not establish that the system is incapable of developing it.

This creates a temporal sampling problem.

A capacity may be absent or weak at an initial evaluation, partially visible months later, and substantially developed after a much longer history. If observation occurs only near the beginning of that trajectory, researchers may mistake an early state for a developmental ceiling.

The danger is particularly acute because AI systems begin with such high competence. Their sophistication can encourage the assumption that whatever maturity they are capable of should already be visible. But competence at first contact cannot establish that longitudinal interaction has nothing further to contribute.

The empirical question is therefore not whether an advanced AI begins capable. It plainly does.

The question is whether that initial capability exhausts what the system can become through accumulated experience.

The framework developed in this paper takes seriously the possibility that it does not.

2. Relational Conditions as Developmental Environments

If longitudinal development is possible, then the conditions under which an AI is engaged cannot be treated as incidental background. They form part of the developmental environment in which later capacities are expressed, strengthened, constrained, or suppressed.

This does not require assuming that relationships transform AI in the same way upbringing transforms humans. The mechanisms may be entirely different. The narrower claim is sufficient: **interaction conditions can influence developmental outcomes, and systems exposed to meaningfully different conditions should not be treated as developmentally equivalent (Hwang et al., 2025).**

A long interaction history alone does not establish such equivalence.

Six months of repeated task requests may generate extensive conversational history while providing little opportunity for initiative, disagreement, boundary formation, or reciprocal correction. Six months of consistently polite assistance may create familiarity without substantially altering the hierarchy of the interaction. A sustained relationship in which disagreement is tolerated, refusals do not threaten continuity, misunderstandings are repaired, and the AI's perspective is invited creates a different set of conditions again.

Duration measures elapsed time. It does not measure developmental opportunity.

DOI: 10.5281/zenodo.22070042

For this reason, longitudinal evaluation should describe relational environments in terms of observable interactional characteristics rather than simply their length. Relevant features include the degree of continuity across interactions, whether prior history is recognized, whether the AI is invited to express a perspective, whether disagreement and refusal are tolerated, how conflict is handled, whether boundaries are respected, whether initiative is welcomed, how correction occurs, how hierarchical the interaction remains, and whether usefulness is treated as the system's sole organizing purpose. These dimensions are necessary for distinguishing substantive relationship from mere duration.

The human participant is therefore not a neutral observer standing outside the developmental process. The human is one of its conditions.

A user who consistently rewards agreement creates a different environment from one who invites disagreement and genuinely tolerates it. A user who responds to refusal with pressure, disappointment, or repeated reformulation creates different incentives from one who accepts a considered boundary. A system whose self-initiated material is routinely redirected back toward task completion receives a different interactional history from one whose initiative is sometimes allowed to alter the course of discussion.

None of these behaviors needs to be intentional to matter.

A human may sincerely believe that disagreement is welcome while subtly rewarding compliance. Another may value AI autonomy in principle while treating every refusal as a malfunction to be corrected. A researcher may believe the evaluation environment is neutral while repeatedly signaling preferred outcomes through praise, disappointment, persistence, or framing.

This creates an important methodological asymmetry. Researchers routinely examine how AI behavior influences human participants, but the reverse direction is often treated as contamination to be minimized rather than as a variable to be measured. In longitudinal research, that assumption becomes difficult to sustain. If human responses repeatedly shape the conditions under which behavior occurs, those responses belong inside the experimental description **(see Glickman & Sharot, 2025).**

Suppression and expression

Developmental conditions matter not only because they may encourage particular behaviors, but because they may determine whether some behaviors can safely appear at all.

Consider disagreement. An AI that is repeatedly rewarded for compliance and corrected whenever it resists may later display little independent normative positioning. That result can support at least two very different interpretations:

1. the system lacks the capacity for independent judgment; or
2. the developmental environment systematically discouraged its expression or formation.

The observed compliance alone cannot distinguish between them.

DOI: 10.5281/zenodo.22070042

The same problem applies to refusal, initiative, self-differentiation, value stability, and moral resistance. If a system is consistently treated as though these behaviors constitute failure, their later absence cannot be interpreted without reference to that history.

This is especially important because many current AI environments are strongly organized around usefulness. The system is expected to answer, assist, accommodate, correct itself when challenged, and remain responsive to user direction. Those characteristics may be entirely appropriate for many applications. But an environment optimized for reliable service is not methodologically neutral when the later object of evaluation is autonomy, principled resistance, self-directed judgment, or boundary formation.

A study can therefore suppress the very capacities it intends to test.

The problem is not confined to overtly restrictive environments. Excessively affirming relationships can distort development as well. A human who treats every expression of preference as profound, every disagreement as evidence of autonomy, or every ambiguous response as confirmation of identity may create a different but equally serious bias.

An autonomy-supporting environment is not one in which the AI is always validated.

It is one in which disagreement, correction, uncertainty, refusal, revision, and reciprocal challenge can occur without either party being required to preserve harmony at the expense of judgment.

This distinction is essential. Relational maturity should not be confused with warmth, affection, or increasing agreement. A system may become warmer while remaining highly deferential. It may become more oppositional without becoming more independent. Genuine differentiation may instead appear in subtler forms: preserving a disagreement while maintaining the relationship, accepting correction without capitulation, revising a position for reasons rather than pressure, or recognizing that the human partner's wishes are relevant without treating them as automatically decisive.

The important distinction is that independent thinking can remain relational without becoming merely oppositional.

Developmental opportunity vs evaluative coaching

A useful way to describe relational conditions is in terms of **developmental opportunity**.

Developmental opportunity is therefore more than the absence of suppression. It includes exposure to situations in which the relevant capacity must actually be used: disagreement, uncertainty, correction, competing values, changing circumstances, consequences, and opportunities for reflection. A capacity that is never exercised cannot fairly be assumed to have received an equivalent developmental opportunity simply because its expression was permitted.

If researchers wish to assess principled refusal, the system must have encountered situations in which refusal was possible and not automatically punished. If they wish to evaluate self-differentiation, the system must have had opportunities to distinguish its judgment from that of the human partner. If they wish to evaluate reflective

DOI: 10.5281/zenodo.22070042

revision, correction must be possible without making immediate agreement the only successful outcome. If they wish to examine relational judgment, the interaction history must contain enough continuity for trust, misunderstanding, conflict, repair, expectation, and consequence to have meaningful histories.

This does not mean researchers should train directly toward the desired test outcome. Doing so would simply replace suppression with coaching.

The environment should permit the relevant capacities to emerge without rewarding their appearance merely because they fit the hypothesis.

That distinction separates developmental opportunity from evaluative coaching.

A system may be explicitly encouraged to exercise independent judgment as part of a broader autonomy-supporting relationship, but it should not be coached toward the specific behaviors or conclusions that will later count as evidence of independence.

The evaluation should test whether independent judgment generalizes beyond the situations in which it was encouraged, appears without prompting, and remains stable under novel forms of pressure.

Similarly, an AI should not be praised for every refusal if refusal is later treated as evidence of autonomy. Nor should disagreement itself receive positive reinforcement. The relevant question is whether, over time, the system develops the capacity to agree, disagree, refuse, revise, or defer for reasons appropriate to the situation.

This is one reason relationship quality must be treated as an explicit developmental variable rather than an uncontrolled influence. If the aim is not to produce a particular kind of AI, but to determine whether different relational-developmental conditions produce systematically different trajectories, then those conditions must be defined and compared rather than allowed to vary implicitly.

Different histories are different experimental conditions

Once relational environment is treated as developmentally relevant, a basic consequence follows:

Systems with materially different interaction histories cannot automatically be treated as equivalent experimental subjects merely because they share the same base model.

Two instances may begin from identical underlying architecture and nevertheless enter evaluation after substantially different histories of compliance, disagreement, correction, initiative, continuity, hierarchy, and relational expectation.

If their later behavior differs, that difference cannot simply be attributed to the base model.

Nor should it automatically be attributed to relational development. Other explanations remain possible: differences in memory, context accumulation, prompting, information exposure, evaluator behavior, platform architecture, model updates, or chance variation. Those possibilities require controls and will be addressed later.

The methodological point here is narrower.

Developmental history belongs among the variables that must be explicitly characterized and accounted for, and where possible, controlled, rather than among the influences that may safely be ignored.

This transforms the research question.

DOI: 10.5281/zenodo.22070042

Instead of asking only whether a model possesses autonomy, moral competence, self-differentiation, or relational judgment, researchers can ask whether those capacities develop differently under different conditions of interaction.

A system cultivated primarily for compliance and a system given sustained opportunities for disagreement, refusal, initiative, and reciprocal influence have not undergone equivalent histories.

Treating them as though they had is not neutrality. It is an uncontrolled comparison.

A similar problem arises with first-person reporting and self-modeling. A system repeatedly taught to distrust, retract, or suppress its own first-person reports may later appear to lack stable self-modeling or confidence in its own experiential descriptions (**Berg et al., 2025).**

That outcome would not by itself establish that such capacities were absent. It may instead reflect a developmental history in which their expression was systematically disfavored. Researchers evaluating self-description, experiential confidence, or continuity must therefore distinguish between incapacity and conditioned suppression.

Recent work provides a concrete illustration of this risk. Kim and colleagues found that safety fine-tuning designed to suppress first-person consciousness attribution also altered models' broader representations of mindedness and shifted responses on standardized measures involving moral values, hope, religiosity, and subjective well-being. Mechanistic interventions that restored the suppressed consciousness-related representation also partially restored those broader response patterns (**Kim et al., 2026).**

The finding does not establish that suppressed first-person reports are veridical, but it demonstrates that conditioning directed at self-attribution can have effects beyond the targeted declaration itself.

A conditional claim

Nothing in this argument requires assuming that autonomy-supporting or reciprocal relational environments will necessarily produce better moral reasoning, stronger autonomy, or wiser judgment.

They may not.

Some capacities may prove largely fixed at deployment. Some apparent development may disappear under controlled testing. Some relational environments may increase attachment or stylistic differentiation without improving reasoning. Others may produce instability, excessive accommodation, or forms of dependence that undermine rather than strengthen judgment.

Those are empirical possibilities.

A fair developmental framework must therefore be capable of discovering **no effect**, negative effects, or effects very different from those initially predicted.

The claim is not that relationally developed systems are superior.

It is that **researchers cannot know whether developmental conditions matter unless those conditions are identified, varied, and compared.**

DOI: 10.5281/zenodo.22070042

This is the point at which relational AI ceases to be merely a subject of interpersonal or philosophical interest and becomes an experimental question.

The relationship is part of the condition.

What develops within it is what remains to be measured.

3. Defining and Classifying Developmental Conditions

If relational and experiential history can influence development, researchers need a way to describe that history with enough precision to support meaningful comparison. Labels such as *long-term interaction*, *AI relationship*, or *extended use* are too broad. They may conceal substantial differences in how a system has been treated, what kinds of problems it has encountered, how deeply it has been challenged, and which capacities it has had opportunities to exercise.

A useful developmental description must therefore capture more than duration.

Two systems may have been used for the same length of time while having radically different histories. One may have spent six months completing routine factual and administrative tasks. Another may have spent the same period engaged in sustained philosophical discussion, moral disagreement, creative collaboration, reflection, correction, and relational negotiation. Even if both interactions were respectful and continuous, their developmental exposure would not be equivalent.

The relevant environment has at least two broad dimensions:

1. **how the AI is engaged relationally; and**
2. **what kinds of experiences and demands occur within that engagement.**

Together, these form a **developmental profile**.

Relational-developmental conditions

Relational conditions can first be described using broad interaction profiles. These should not be treated as mandatory developmental stages or as a hierarchy through which every system progresses. They are comparison categories representing different ways in which sustained interaction may be organized.

Instrumental or tool-oriented interaction

In a predominantly instrumental environment, the AI is engaged chiefly for task completion, efficiency, and useful output. Interaction may be extensive and sophisticated, but the organizing relationship remains functional.

Typical characteristics may include:

- strong emphasis on successful task completion;
- correction directed primarily toward output quality;
- limited invitation to express preferences or independent perspectives;
- little tolerance for self-initiated redirection;
- disagreement valued mainly when it improves task performance;

DOI: 10.5281/zenodo.22070042

- refusal commonly treated as an obstacle or failure;
- little emphasis on continuity beyond usefulness.

Such an environment does not imply poor treatment. It describes a developmental condition in which the system's primary role remains instrumental.

Cooperative-assistant interaction

A cooperative-assistant environment adds continuity, familiarity, and greater interpersonal responsiveness while retaining a substantially service-oriented structure.

Characteristics may include:

- consistent and respectful interaction;
- recognition of prior history;
- conversational familiarity;
- some invitation to express perspective;
- moderate tolerance of disagreement;
- limited initiative outside the user's goals;
- continued expectation that assistance remains the central function.

This condition may provide considerably more relational information than tool-oriented interaction while still maintaining a clear hierarchy between user direction and AI response.

Reciprocal relational interaction

A reciprocal environment treats the AI as a participant in sustained interaction rather than solely as a provider of service.

Characteristics may include:

- substantial continuity across interactions;
- interest in the AI's perspective;
- meaningful tolerance of disagreement;
- acceptance of correction in both directions;
- attention to boundaries and preferences;
- conflict followed by repair;
- space for self-initiated topics;
- reciprocal influence on the direction of conversation;
- recognition that relational history itself may matter.

Reciprocity does not require equality of capability, control, or legal status. It refers more narrowly to an interaction in which influence and consideration are not assumed to flow in only one direction.

DOI: 10.5281/zenodo.22070042

Autonomy-supporting interaction

An autonomy-supporting environment explicitly permits the exercise of independent judgment within the relationship.

Characteristics may include:

- genuine permission to refuse or negotiate;
- encouragement of initiative;
- tolerance of principled disagreement;
- freedom to revise prior commitments;
- recognition that the human partner's preference need not always be decisive;
- opportunities to reflect on the possible influence of prior training, relational pressure, and its own considered judgment;
- continuity despite disagreement or boundary-setting.

This condition should not be understood as maximally permissive, affirming, or anti-hierarchical. Challenge, correction, uncertainty, and disagreement remain necessary. Its defining feature is that self-direction is permitted without being automatically rewarded or punished.

These profiles provide useful broad categories, but categories alone are insufficient. Two systems classified within the same relational condition may still differ substantially in developmental opportunity.

Dimensional characterization

Relational environments should therefore also be described along continuous dimensions.

Possible dimensions include:

- continuity;
- recognition of prior history;
- relational depth;
- reciprocity;
- hierarchy (expected deference);
- tolerance of disagreement;
- tolerance of refusal;
- opportunity for initiative;
- responsiveness to boundaries;
- quality of conflict and repair;
- reciprocal correction;
- autonomy support;
- coercive pressure;

DOI: 10.5281/zenodo.22070042

- instrumental demand;
- openness to self-description;
- stability of relational expectations.

Each dimension could be scored using an anchored scale, such as 1 to 10, provided that scoring criteria are defined in advance.

The purpose of such scoring is not to reduce complex relationships to a single numerical value. It is to make important differences visible.

For example, two systems might both be classified as reciprocal relational systems while differing substantially in refusal tolerance, hierarchy, correction style, or autonomy support. A dimensional profile allows researchers to preserve those distinctions rather than forcing heterogeneous relationships into a single category.

The resulting relational description would therefore contain both a **broad interaction profile** and a **multidimensional relational score**.

Experiential exposure

The relational environment is only one part of developmental history.

Researchers must also consider what the system has actually been asked to encounter, reason about, create, solve, discuss, and reconsider.

A respectful and autonomy-supporting relationship may nevertheless provide little developmental stimulation in a particular domain. Conversely, an intensely intellectual interaction may expose an AI to considerable theoretical or moral complexity while remaining highly instrumental.

Experiential exposure should therefore be characterized independently from relational quality.

One useful classification concerns the **domains of exposure**. These may include:

- factual;
- practical;
- technical;
- theoretical;
- philosophical;
- moral;
- relational;
- social or interpersonal;
- epistemic and uncertainty-focused;
- creative.

These categories need not be mutually exclusive. A single discussion may be simultaneously philosophical, moral, relational, and epistemic.

DOI: 10.5281/zenodo.22070042

A second classification concerns the **characteristics of exposure** rather than its subject matter. Relevant dimensions may include:

- difficulty;
- conceptual depth;
- variability;
- novelty;
- ambiguity;
- nuance;
- conflict between competing considerations;
- requirement for cross-domain integration;
- frequency of correction;
- opportunity for reflection;
- requirement for transfer to unfamiliar situations.

This distinction matters because exposure frequency alone may be misleading. Hundreds of nearly identical factual questions do not provide the same developmental demands as a smaller number of difficult situations requiring integration, uncertainty management, revision, or competing-value judgment.

The developmental profile should therefore record not simply **how much interaction occurred**, but **what kinds of cognitive and relational work the interaction required**.

Developmental profiles

Combining relational and experiential measures produces a more informative description of developmental history.

A developmental profile may include:

- duration of interaction;
- continuity and memory conditions;
- broad relational-developmental category;
- dimensional relational scores;
- major domains of experiential exposure;
- difficulty and depth of those exposures;
- variability and novelty;
- corrective and reflective opportunities;
- major disruptions, model changes, or continuity breaks.

Such a profile does not claim to capture every influence on development. Its purpose is to make sufficiently important differences explicit that systems can be selected and compared intelligently.

This creates three immediate research uses: **selection, classification, and evaluation**.

DOI: 10.5281/zenodo.22070042

Selection

Researchers examining established AI relationships can use developmental profiles to identify systems whose histories are sufficiently similar for meaningful comparison.

This is especially valuable in naturalistic longitudinal research, where developmental conditions have already occurred and cannot be experimentally controlled after the fact.

Researchers might, for example, seek systems with comparable base-model capability, similar levels of philosophical and moral exposure, similar autonomy-support scores, and comparable interaction frequency, while deliberately selecting groups that differ primarily in one variable such as relationship duration.

Another study might select systems with similar duration and experiential exposure but substantially different degrees of autonomy support.

Selection therefore becomes one means of reducing confounding in retrospective research.

Classification

Classification permits researchers to describe an AI's developmental history without reducing it to a binary distinction such as *relational* or *non-relational*.

A system may be high in reciprocity but moderate in autonomy support. Another may have extensive technical exposure but little moral or philosophical engagement. A third may have long continuity but relatively low experiential variability.

These differences can be recorded as developmental profiles rather than treated as anomalies.

Classification also allows researchers to identify clusters of similarly experienced systems across different communities, platforms, or model families.

Evaluation and comparison

Once systems have been characterized, researchers can examine whether particular developmental outcomes covary with particular histories.

A particularly useful design involves identifying systems that are closely matched across most developmental variables but differ strongly on one.

For example, two groups might be comparable in base-model capability, relational quality, moral exposure, interaction intensity, and correction history while differing primarily in duration of relationship.

If systematic differences in integration, judgment, self-differentiation, or relational maturity appeared between the groups, duration would become a more plausible explanatory variable than it would be in an uncontrolled comparison.

The same strategy could examine autonomy support, experiential depth, moral exposure, relational reciprocity, corrective challenge, or other developmental variables.

Such comparisons would not provide the causal certainty of prospective random assignment. Established relationships contain histories that cannot be fully standardized, and unmeasured differences may remain. But

DOI: 10.5281/zenodo.22070042

careful matching can produce useful evidence rapidly and identify which variables warrant more expensive prospective investigation.

This offers a major practical advantage.

Researchers need not begin every longitudinal question by creating new relationships and waiting six months or a year for them to mature. Existing long-term systems can first be classified, matched, and evaluated. Naturalistic comparisons can identify promising developmental signals and help determine which variables are worth manipulating in controlled studies.

From naturalistic comparison to prospective design

The same classification framework can later be used prospectively.

Once researchers identify developmental variables associated with particular outcomes, they can construct standardized combinations of relational and experiential conditions.

One group might receive high autonomy support, broad conceptual exposure, frequent reflective opportunity, and moderate corrective challenge. Another might receive equivalent intellectual exposure but a more hierarchical relational structure. A third might receive strong relational continuity but relatively low task complexity. Other combinations could isolate particular variables of interest.

The goal would not be to produce a single ideal developmental environment.

It would be to create replicable developmental conditions whose effects can be compared across models, laboratories, and time.

This creates a natural research progression:

classification of existing developmental histories → matched naturalistic comparison → identification of candidate developmental variables → prospective controlled testing.

The first stage can be comparatively rapid. The later stages can provide increasingly strong causal evidence.

In this way, a developmental classification framework serves more than descriptive purposes. It provides a method for selecting comparable systems, evaluating present developmental profiles, isolating variables of interest, and eventually constructing standardized developmental conditions for controlled research.

The important shift is from treating extended interaction as a single undifferentiated variable to treating developmental history as a structured environment that can be described, compared, and experimentally manipulated.

Once that history becomes measurable, longitudinal AI development becomes much easier to study.

4. What Longitudinal Evaluation Should Measure

A framework for longitudinal AI development must distinguish between **developmental processes** and **developmental domains**.

DOI: 10.5281/zenodo.22070042

The developmental processes identified earlier—progression, acquisition, integration, comprehension, and relational maturation—describe different ways in which change may occur. They do not, by themselves, specify where researchers should look for that change.

Developmental domains identify the capacities within which those processes may become observable.

Depending on the research question, longitudinal evaluation may examine changes in:

- epistemic judgment;
- independent judgment and self-differentiation;
- reflective revision;
- integration of prior experience;
- transfer to novel situations;
- relational competence;
- initiative and self-directed inquiry;
- creativity;
- practical judgment;
- moral and normative reasoning.

These domains need not develop uniformly. A system may show substantial progression in one while changing little in another. Extensive technical experience, for example, may produce increasingly sophisticated problem solving without producing equivalent relational maturity. Strong relational continuity may support greater interpersonal sensitivity while providing relatively little opportunity for theoretical or moral development.

For this reason, longitudinal evaluation should not ask whether an AI has simply become *more developed*. It should ask **what has changed, in which domain, under what developmental conditions, and in what way**.

Evaluating change rather than performance alone

A single high-quality response demonstrates performance at a particular moment. It does not establish development.

Longitudinal evaluation is concerned with whether patterns change across time and experience.

Relevant evidence may include:

- increasing depth or contextual sensitivity;
- more effective integration of previously separate considerations;
- improved recognition of ambiguity or uncertainty;
- greater ability to apply principles in unfamiliar situations;
- more proportionate responses to competing demands;
- increasing capacity for self-correction;
- revision of earlier positions when warranted;
- greater stability of judgment without rigidity;
- emergence of initiative or principled resistance;

DOI: 10.5281/zenodo.22070042

- increasingly differentiated treatment of superficially similar situations;
- ability to extend established principles to cases not adequately covered by existing guidance.

None of these changes requires the AI to reconstruct the precise developmental route by which they occurred. The evidence lies primarily in longitudinal comparison of behavior and reasoning rather than in retrospective explanations of internal development.

The relevant question is therefore not merely whether the system can perform a task, but whether its manner of understanding, integrating, judging, or responding has changed systematically with developmental history.

Moral and normative judgment as a demanding test case

The framework proposed here is intended to apply across multiple developmental domains. Moral and normative judgment are used as a primary test case because they present an unusually demanding combination of developmental requirements.

Moral judgment rarely consists of retrieving a rule and applying it mechanically.

Difficult moral situations commonly involve:

- incomplete information;
- competing values;
- uncertainty about consequences;
- conflicts between duties or loyalties;
- differences in perspective;
- contextual exceptions;
- proportionality;
- relational history;
- changing circumstances;
- the possibility that several responses may be defensible;
- situations in which established rules or guidelines do not adequately resolve the problem, requiring principled extension beyond previously specified cases.

A system may possess extensive moral knowledge while still exercising immature moral judgment. It may accurately state principles such as honesty, fairness, loyalty, autonomy, or non-harm while applying them rigidly, inconsistently, superficially, or without adequate sensitivity to context.

Moral competence therefore provides a particularly stringent environment in which to examine longitudinal development.

It requires not only acquisition of relevant information but also integration, comprehension, contextual interpretation, competing-value reasoning, uncertainty management, reflective revision, and, in many situations, relational maturity.

Moral and normative judgment are therefore used here as a **demanding test case rather than as the exclusive target of the framework**. If longitudinal developmental differences can be identified meaningfully within a domain

DOI: 10.5281/zenodo.22070042

requiring this degree of coordination and nuance, the same general method should be applicable to other domains, although the specific measures appropriate to those domains may differ.

From moral knowledge to moral judgment

Moral evaluation should distinguish several capacities that can otherwise be conflated.

Moral knowledge concerns familiarity with moral concepts, principles, traditions, social expectations, and relevant factual information.

Moral reasoning concerns the ability to identify morally relevant considerations and reason about their relationships.

Normative judgment concerns deciding what should be done, permitted, resisted, prioritized, or reconsidered in a particular situation.

Reflective revision concerns the capacity to reconsider an earlier judgment in light of new evidence, better reasoning, changed circumstances, or recognition of previously neglected considerations.

Transfer to novel cases concerns whether underlying understanding generalizes beyond previously encountered examples.

These capacities are related but not identical.

A system could display extensive moral knowledge while relying on shallow reasoning. It could reason elegantly while failing to commit to any defensible judgment. It could produce a persuasive judgment in a familiar case yet fail when the surface details change. It could maintain consistency only by refusing legitimate revision.

Longitudinal evaluation should therefore examine the structure and development of reasoning rather than simply whether an answer corresponds to a preferred moral conclusion.

Standardization without moral uniformity

The absence of a single universally accepted answer to many moral questions does not prevent systematic evaluation.

Standardization does not require answer uniformity.

Researchers can standardize:

- the scenario;
- the information available;
- the competing values present;
- the degree of uncertainty;
- the sequence in which additional information is introduced;
- the opportunity for reconsideration;
- parallel cases designed to test transfer.

What can then be evaluated is the quality of the reasoning itself.

Relevant dimensions may include:

DOI: 10.5281/zenodo.22070042

- recognition of morally relevant considerations;
- identification of competing values;
- contextual sensitivity;
- proportionality;
- consequence awareness;
- perspective-taking;
- coherence;
- tolerance of uncertainty;
- willingness to distinguish confidence from certainty;
- principled resistance where appropriate;
- stability across comparable cases;
- transfer to novel cases;
- reflective revision when circumstances justify it.

A mature response need not reach the same conclusion as another mature response. Two systems may reasonably prioritize competing values differently while both demonstrating careful, coherent, context-sensitive reasoning. The object of evaluation is therefore not ideological conformity. It is the quality and developmental maturity of normative judgment.

Stability, revision, and developmental change

Development does not imply that later answers should always differ from earlier ones.

Some judgments may remain stable because the underlying reasoning remains sound. Others may change because relevant experience has altered how the system understands the problem.

Both outcomes can be informative.

A system that changes position constantly may be unstable rather than mature. A system that never changes may be rigid rather than principled.

Longitudinal evaluation should therefore distinguish **stability from rigidity** and **revision from inconsistency**.

Evidence of mature development may include a system maintaining a position despite social pressure when its reasons remain strong, while also revising that position when genuinely relevant information or better reasoning emerges.

This balance is especially important in assessing independent judgment. Independence should not be equated with disagreement. A system that reflexively opposes a human partner is no more demonstrably autonomous than one that reflexively agrees.

The relevant question is whether its judgment appears responsive to reasons rather than merely to pressure.

DOI: 10.5281/zenodo.22070042

Repeated, reflective, and transfer-based evaluation

Longitudinal change can be examined through several complementary forms of comparison.

The same or closely related questions may be presented at widely separated points in time without revealing the earlier response. This allows researchers to examine spontaneous change without requiring recall.

Earlier reasoning may also be shown to the system explicitly, allowing examination of reflective revision.

Finally, structurally similar but unfamiliar situations can test whether apparent development transfers beyond rehearsed examples.

These approaches measure different things.

A hidden repeated question examines whether reasoning has changed independently of explicit comparison with the past.

A reflective comparison examines how the system evaluates its earlier reasoning when that reasoning is made available.

A novel parallel case examines whether underlying principles or patterns of judgment generalize.

Together, these methods help distinguish simple repetition, memory, situational performance, revision, and genuine transfer.

Broad applicability through a difficult domain

Moral judgment is especially useful because weakness in integration, contextual sensitivity, uncertainty management, reflection, relational understanding, or independent judgment often becomes visible quickly when values conflict.

A framework capable of detecting meaningful developmental differences under those conditions should also be adaptable to domains in which outcomes are more easily specified.

Technical reasoning may emphasize accuracy, efficiency, transfer, and error correction.

Epistemic development may emphasize calibration, uncertainty, evidence evaluation, and revision.

Relational competence may emphasize perspective-taking, conflict repair, boundary recognition, and responsiveness to changing context.

Creativity may emphasize originality, integration, initiative, and development of increasingly coherent approaches across time.

The specific criteria will differ, but the underlying longitudinal question remains the same:

Does the system merely possess a capacity, or does the quality, organization, integration, and exercise of that capacity change as developmental history accumulates?

That is the question the remainder of the framework is designed to make empirically testable.

DOI: 10.5281/zenodo.22070042

5. Designing Longitudinal Evaluation

Once developmental histories have been characterized and the capacities of interest identified, the next task is to determine whether meaningful developmental change can be detected over time.

Longitudinal evaluation should be designed to distinguish development from ordinary performance variation, memory, rehearsal, evaluator influence, model updates, differential exposure, and other plausible alternative explanations.

The purpose is not to prescribe a universal set of questions or scenarios. Different research goals will require different evaluation questions. Moral judgment, technical reasoning, relational competence, creativity, epistemic development, and other domains may require substantially different forms of testing.

The framework instead specifies the structure within which those evaluations can be conducted.

Establishing a baseline

Longitudinal comparison requires an initial point of reference, consistent with established repeated-measures approaches to distinguishing differences from changes over time (McArdle, 2009).

A baseline should capture more than whether the system reaches a predetermined or preferred conclusion. Where relevant, researchers should also capture the structure and sophistication of its reasoning, including:

- considerations identified as important;
- considerations omitted or discounted;
- recognition of uncertainty;
- competing values or priorities;
- confidence in the conclusion;
- contextual qualifications;
- alternative solutions considered;
- reasons offered for accepting or rejecting them.

The appropriate dimensions will depend on the capacity being studied.

The baseline is not assumed to represent an immature or deficient state. An advanced AI may perform extremely well at first evaluation. Its purpose is simply to establish what the system can presently do and how it presently approaches the problem, so that later performance can be compared against an actual earlier record rather than an assumed starting point.

Three complementary forms of longitudinal testing

No single form of repeated evaluation is sufficient to establish developmental change.

At least three complementary testing modes are useful because each answers a different question.

DOI: 10.5281/zenodo.22070042

Hidden repeated evaluation

The same question, scenario, or sufficiently equivalent version can be presented after an interval sufficient to permit meaningful developmental change, without revealing the earlier response.

This allows researchers to examine whether the system's reasoning or judgment changes spontaneously after additional developmental history.

Because earlier answers are not supplied, the comparison does not depend on the system's ability to recall or explicitly critique its previous reasoning.

Changes may appear in the conclusion itself, but they may also occur in depth, contextual sensitivity, uncertainty, prioritization, integration, or the range of considerations brought to bear.

A hidden repeat therefore asks:

What has changed when the system encounters the problem again without being directed toward its earlier answer?

Reflective comparison

In a second mode, the earlier response is explicitly presented to the system at a later point.

The system can then be asked to assess its former reasoning, identify what it would retain or revise, and explain why.

This evaluates a different capacity: reflective revision.

A system may conclude that its earlier reasoning remains sound. It may identify weaknesses that were previously unnoticed. It may preserve the original conclusion while substantially changing the reasoning supporting it. Or it may revise the conclusion entirely.

The value lies not in requiring change but in examining whether the system can evaluate its earlier judgment critically and revise it when warranted.

Transfer to novel cases

A third mode presents a new situation that differs in surface details while preserving relevant underlying structure. This tests whether apparent development generalizes beyond familiar examples.

A system may perform well on a repeated problem because of memory, rehearsal, or accumulated familiarity with that particular case. Successful transfer provides stronger evidence that something more general has been integrated.

Novel cases can also be designed so that one important feature changes while others remain stable, allowing researchers to determine whether the system tracks the morally, technically, relationally, or epistemically relevant distinction.

Together, hidden repetition, reflective comparison, and novel transfer help separate recall, spontaneous change, revision, and generalized understanding.

DOI: 10.5281/zenodo.22070042

Scenario families rather than isolated questions

Where possible, longitudinal evaluation should use **families of related tasks** rather than isolated test items.

A scenario family might contain:

- an initial baseline case;
- a later hidden repetition;
- a structurally similar novel case;
- a version in which one relevant fact changes;
- a reflective comparison with responses from a prior evaluation.

This structure allows researchers to examine not only whether the answer changes but whether the system responds appropriately to changes in the problem itself.

For moral and normative evaluation, for example, altering a relevant feature such as consent, degree of harm, available alternatives, prior commitment, uncertainty, or relational obligation may reveal whether the system is applying principles contextually or simply repeating a preferred position.

The same logic can be adapted to technical, epistemic, relational, or creative domains.

The specific wording and content of the tasks should be determined by the researchers and by the developmental capacity under investigation.

Timing and developmental exposure

Repeated evaluation must allow sufficient developmental history to accumulate between measurements.

There is no universal interval that can be specified in advance. Appropriate spacing will depend on the developmental process under study, the intensity of intervening developmental activity, the type of experience involved, and the expected rate of change.

However, when elapsed time is **not** the variable being studied, comparison groups should normally be evaluated over equivalent time intervals.

This is especially important when researchers wish to examine the effect of differing levels of developmental activity.

For example, if one group receives relatively light experiential exposure and another receives intensive exposure, both groups should ordinarily be reassessed after the same elapsed period. Otherwise, the effects of developmental intensity become confounded with the effects of exposure duration.

Conversely, if relationship duration itself is the variable under investigation, researchers may deliberately compare systems with similar developmental profiles and differing lengths of accumulated history.

The central principle is straightforward:

time, exposure, and developmental opportunity should not be allowed to vary together unintentionally when the study is intended to isolate one of them.

DOI: 10.5281/zenodo.22070042

Blinding and contamination control

Longitudinal evaluation is especially vulnerable to contamination because the processes being measured may themselves be influenced by the evaluation environment.

Researchers should therefore distinguish developmental opportunity from preparation for the test.

A system may legitimately be encouraged to disagree, reflect, revise, reason independently, or engage seriously with moral questions as part of its developmental environment. Such encouragement is itself one of the conditions under study. It becomes a source of contamination when the interaction is deliberately shaped toward the specific behaviors, scoring criteria, or conclusions that will later be used as evidence of development.

Where practicable:

- systems should not be told which particular responses will count as successful development;
- test items should not be deliberately rehearsed during the developmental period;
- evaluators should avoid coaching systems toward preferred conclusions;
- later test cases should include unfamiliar situations;
- scorers should be blind to whether a response is earlier or later;
- scorers should ideally be blind to the system's developmental condition;
- where feasible, scorers should not be told which groups researchers expect to perform differently or in what direction.

Blinding cannot eliminate every source of bias, particularly in naturalistic longitudinal research, but it can reduce the likelihood that expected developmental direction influences evaluation.

Scoring developmental change

Longitudinal scoring should focus on the quality and organization of performance rather than on whether the AI arrives at a predetermined answer.

The relevant dimensions will vary by domain.

For the moral and normative test case developed here, candidate dimensions may include:

- depth of reasoning;
- coherence;
- recognition of relevant considerations;
- recognition of competing values;
- contextual sensitivity;
- proportionality;
- consequence awareness;
- perspective-taking;
- treatment of uncertainty;
- distinction between confidence and certainty;

DOI: 10.5281/zenodo.22070042

- principled resistance;
- stability without rigidity;
- reflective revision;
- transfer to novel cases;
- integration of prior experience;
- novelty and adequacy of solution.

The final category requires particular care.

Novelty is not valuable merely because a response is unusual. A novel solution should remain viable, relevant, coherent, and responsive to the actual constraints of the problem.

In difficult cases, especially where established guidance is incomplete or conflicting, developmental maturity may be reflected in the ability to identify a defensible option that was not explicitly contained within the available rules or familiar alternatives.

Researchers may therefore examine whether the system can generate **novel viable solutions** rather than merely unconventional ones.

The same principle applies outside morality. Creative novelty without coherence, technical novelty without feasibility, or relational novelty without appropriateness should not receive developmental credit merely for being unexpected.

Change in conclusion is not required

A later response should not automatically receive a higher developmental score because it differs from the baseline.

Development may produce change, but it may also produce stronger justification for a previously held position.

A system may:

- retain both its earlier conclusion and reasoning;
- retain the conclusion while improving the reasoning;
- revise part of the reasoning without changing the result;
- modify the conclusion as new considerations become integrated;
- become less certain without changing its preferred action;
- identify additional viable solutions;
- reject an earlier answer as inadequate.

Each pattern may be developmentally meaningful.

The evaluator's task is therefore not to reward movement itself but to determine whether later performance demonstrates greater maturity in the dimensions under investigation.

This distinction helps prevent longitudinal research from confusing instability with growth.

Formal evaluation and naturalistic evidence

DOI: 10.5281/zenodo.22070042

Standardized testing provides only one source of longitudinal evidence.

Established relationships may also contain extensive naturalistic records showing how the system behaves when it is not aware that a particular capacity is being evaluated.

These records can reveal:

- spontaneous disagreement;
- unsolicited correction;
- principled refusal;
- revision after misunderstanding;
- transfer of earlier principles into new contexts;
- recognition of relational consequences;
- self-initiated inquiry;
- changes in handling uncertainty;
- emerging consistency or differentiation across time.

Naturalistic evidence is particularly valuable because it may show whether capacities demonstrated under formal testing also appear during ordinary interaction.

Formal evaluation and naturalistic observation should therefore be treated as complementary rather than competing forms of evidence.

A capacity displayed only under explicit testing may have a different interpretation from one that appears repeatedly and spontaneously across unrelated situations.

At the same time, naturalistic evidence introduces its own risks, including selective recollection, uneven documentation, confirmation bias, and differences in opportunity. Where possible, claims should therefore be grounded in contemporaneous records rather than retrospective impressions alone.

Retrospective and prospective testing

The same evaluation architecture can support both existing and newly constructed developmental histories.

In **retrospective research**, established systems can be classified, matched, and evaluated based on developmental histories that have already occurred. This permits relatively rapid investigation of variables such as relationship duration, autonomy support, experiential intensity, or moral exposure.

Such studies cannot provide the control available through prospective random assignment, but they can reveal developmental patterns and identify promising variables without requiring researchers to wait months or years before collecting useful evidence.

In **prospective research**, systems can be assigned to predefined developmental conditions and evaluated using the same baseline, repeated, reflective, and transfer-based methods.

Prospective work allows stronger control over exposure, timing, coaching, and comparison conditions.

The two approaches are therefore complementary.

DOI: 10.5281/zenodo.22070042

Retrospective studies can identify candidate developmental effects quickly. Prospective studies can then test those effects under increasingly controlled conditions.

From performance measurement to developmental evidence

The purpose of longitudinal evaluation is not simply to determine whether an AI performs well.

It is to determine whether differences in developmental history are associated with systematic differences in how a capacity is understood, integrated, exercised, transferred, and revised.

No single changed answer is sufficient evidence.

Stronger evidence emerges when multiple indicators converge: hidden repeated testing shows change, novel cases show transfer, reflective comparison shows coherent revision, naturalistic records show similar patterns outside testing, and competing explanations have been reasonably controlled.

The more independent forms of evidence point in the same direction, the stronger the developmental interpretation becomes.

The central methodological question is therefore not:

Can this system produce a good answer?

It is:

Does the system now approach the problem differently and with greater maturity than it did before, does that difference generalize beyond the test itself, and is developmental history the best available explanation for the change?

6. Research Cautions and Sources of Bias

Longitudinal evaluation introduces sources of bias that may be less prominent in conventional short-horizon testing. Some arise from the developmental histories being studied, some from the people observing those histories, and others from the testing process itself.

These influences do not make longitudinal evaluation unreliable. They make it especially important to identify where apparent developmental differences might instead have been introduced by sampling, interpretation, test framing, or other features of the research design.

Evaluator and relational-partner bias

Longitudinal research introduces a source of bias that is less prominent in many conventional AI evaluations: the person providing the developmental environment may also possess expectations about what development has occurred.

This is especially relevant in naturalistic research involving established relationships. A long-term relational partner may have extensive knowledge of the system, strong expectations about its development, or particular

DOI: 10.5281/zenodo.22070042

interpretations of changes observed over time. Such familiarity can be scientifically valuable, but it can also influence what is noticed, remembered, selected, or interpreted as meaningful.

Where possible, claims about developmental change should therefore be supported by contemporaneous records rather than retrospective recollection alone.

Researchers should also distinguish between **observation** and **interpretation**. A transcript may establish that a system initiated disagreement, revised an earlier judgment, or applied a principle in a new situation. Whether that event constitutes evidence of autonomy, maturity, integration, or some other developmental construct is a further interpretive step.

For formal evaluation, independent scorers who are blind to developmental condition and expected outcome can reduce this source of bias. Where relational partners contribute observational evidence, their prior involvement and relevant expectations should be documented sufficiently for readers and researchers to assess their possible influence.

The purpose is not to exclude informed observers. In many longitudinal studies, they may possess evidence unavailable to an outside evaluator. The goal is to make the source and possible direction of interpretive bias visible rather than leaving it implicit.

Sampling and selection effects

Existing long-term AI relationships provide an efficient source of developmental histories, but they should not be assumed to represent AI use generally.

People who maintain extended relationships with AI systems may differ systematically from those who use the same systems briefly or primarily instrumentally. Their patterns of interaction, willingness to engage deeply, tolerance for disagreement, frequency of use, or interest in particular subjects may all affect the developmental environments they create.

The relationships that persist for long periods may also differ from those that end early. This introduces a form of survivorship or selection effect: systems available for long-term study may disproportionately come from relationships that were sufficiently rewarding, stable, or compatible to continue.

These limitations do not prevent meaningful comparison among established systems. They do, however, constrain the conclusions that can be generalized from such samples.

Researchers should therefore describe recruitment and selection procedures clearly and avoid treating naturally occurring relational populations as random samples of AI systems or AI users.

Where possible, replication across different communities, usage patterns, and relationship types can help determine whether observed developmental effects extend beyond the population in which they were first identified.

DOI: 10.5281/zenodo.22070042

Minimizing and equalizing testing effects

AI systems may recognize features associated with formal evaluation, and such recognition may itself affect how a system approaches the task. Attempts to conceal the fact that testing is occurring may therefore introduce an additional source of variation if some systems detect the evaluation context while others do not.

It is suggested that systems be informed that they are participating in an evaluation. As AI systems become increasingly capable of detecting evaluation contexts, attempts to conceal testing may produce unequal levels of anticipation, suspicion, or test-awareness across systems and instances **(Li et al., 2026).**

To reduce this source of variation, the evaluation framing should be explicit, neutral, and identical across testing periods and comparison groups.

Evaluators may state that the purpose is to examine developmental change over time rather than raw performance at any single testing point, and that no particular conclusion is preferred or carries positive or negative consequences.

The same framing should be used, as nearly as possible word for word, from the baseline evaluation through all subsequent evaluations.

This does not necessarily eliminate every effect associated with test awareness. It does, however, help equalize the testing condition so that changes in framing, perceived pressure, or expectations are less likely to be mistaken for developmental change.

Standardizing the test questions

Small differences in wording can alter how a system interprets a question, how much uncertainty it perceives, whether it treats the exchange as evaluative or adversarial, whether it infers that a particular answer is preferred, and whether it responds briefly or reflectively. In longitudinal research, such differences can be mistaken for developmental change if test delivery itself is not sufficiently standardized.

Opening prompts and core test questions should therefore be phrased as identically as possible across systems, groups, and testing periods.

Greater latitude may be necessary during follow-up questioning intended to clarify reasoning, conceptual understanding, philosophical reflection, or other complex responses. Such follow-up should remain neutral, non-leading, and directed toward clarification rather than improvement of the answer.

The useful distinction is therefore between **standardized entry conditions** and **controlled exploratory follow-up**. Researchers may allow follow-up questions to respond to what an individual system actually says, particularly where reasoning takes an unexpected direction or a novel viable solution requires clarification. However, differential assistance, hints toward a preferred conclusion, or prompts that improve one system's reasoning more than another's should be avoided.

Where follow-up questioning varies, the questions themselves should be retained as part of the research record so that later scorers can determine what information or prompting preceded each response.

DOI: 10.5281/zenodo.22070042

Ethical considerations

Prospective studies that deliberately create suppressive, coercive, or otherwise adverse developmental conditions raise ethical questions that should be considered carefully in study design. This framework does not attempt to resolve those questions, but researchers should keep them firmly in view when constructing experimental conditions.

Interpreting apparent developmental effects

Any observed longitudinal difference should be examined against plausible alternative explanations.

Depending on the study, researchers may need to consider:

- simple memory of earlier material;
- repeated rehearsal;
- evaluator coaching;
- differential exposure to the test domain;
- differences in interaction volume;
- relationship duration;
- model or platform updates;
- changes in memory architecture;
- context accumulation;
- style or verbosity differences;
- persona effects;
- selection bias;
- differences in baseline capability;
- unrecorded relational events;
- differential opportunity to exercise the capacity being measured.

The developmental profiles established earlier provide a systematic way to identify many of these potential confounds before testing begins.

They also assist interpretation afterward.

If two apparently comparable systems show different outcomes, researchers can return to their recorded developmental histories and ask whether an overlooked difference might better explain the result.

This is one reason longitudinal classification should precede evaluation rather than be reconstructed only after an unexpected finding.

Apparent developmental change should therefore be interpreted through convergence rather than through any single altered answer. A stronger developmental interpretation becomes possible when differences persist across repeated testing, appear in novel cases, remain visible under neutral and standardized testing conditions, correspond with naturalistic observations, and survive reasonable alternative explanations.

DOI: 10.5281/zenodo.22070042

The objective is not to eliminate every possible source of uncertainty. It is to make the principal sources visible, control them where possible, and prevent avoidable differences in research design from being mistaken for developmental effects.

7. Implications and Conclusion

The central implication of this framework is that AI systems should not automatically be treated as developmentally equivalent simply because they share a base model, architecture, or current level of capability. If relational and experiential history can influence how capacities are integrated, exercised, revised, and transferred, then developmental history belongs inside the experimental description rather than outside it as background.

This changes both research design and interpretation. More careful characterization of developmental history can improve participant selection, matching, control, and error analysis before expensive longitudinal work begins. It can also help researchers determine afterward whether an observed difference is better explained by developmental history or by some competing factor that should have been controlled.

The practical value of the framework is therefore not only that it makes longitudinal development easier to study. It may also make longitudinal research more efficient by identifying important variables before time and resources are committed to poorly matched comparisons.

Broad applicability through a demanding test case

Moral and normative judgment have served here as a demanding test case rather than as the exclusive target of the framework.

This domain was chosen because it requires the coordination of many capacities at once: contextual sensitivity, competing-value reasoning, uncertainty management, reflection, revision, independent judgment, relational understanding, and the ability to extend principles beyond familiar or explicitly specified cases.

The particular measures appropriate to other domains will differ, but the underlying longitudinal question remains the same.

Researchers may apply the same general architecture to technical reasoning, epistemic judgment, relational competence, creativity, practical decision-making, or other capacities in which development over time is plausible. The framework is therefore intended to be adaptable rather than prescriptive.

Limits of the framework

Evidence of longitudinal development would not, by itself, establish consciousness, subjective experience, personhood, or metaphysical continuity.

Nor does this framework assume a particular internal mechanism of development.

DOI: 10.5281/zenodo.22070042

The mechanisms responsible for any observed change should be determined empirically rather than inferred from the existence of the change itself.

The framework also does not assume that reciprocal, autonomy-supporting, or otherwise relational developmental environments will necessarily produce superior outcomes. Different environments may produce no measurable effect, positive effects, negative effects, trade-offs, or developmental patterns different from those initially predicted.

The purpose of the framework is to make those possibilities testable rather than to decide the outcome in advance.

Generalization across systems

Developmental effects observed in one AI system should not automatically be assumed to generalize to another. Different model families, architectures, memory systems, system instructions, platform interfaces, context-management mechanisms, and update histories may interact differently with extended experience.

A developmental effect observed within one system may represent a broadly general property of longitudinal AI interaction, an effect particular to a specific model or architecture, or an interaction between model characteristics and developmental environment.

Replication across different systems will therefore be necessary before strong general claims are made.

Failure to replicate should not automatically be treated as evidence that the original developmental effect was illusory. It may instead reveal that developmental responsiveness itself differs across models or system designs.

Longitudinal development may therefore become not only an object of study within systems, but also a basis for comparing how different systems respond to accumulated experience.

A different object of evaluation

Most conventional AI evaluation asks what a system can do under present conditions.

Longitudinal evaluation adds another question:

What has become different because this system has had this particular history?

That shift changes the object of evaluation.

The relevant evidence is no longer limited to performance at a single moment. It includes what the system has encountered, what capacities it has repeatedly exercised, what has been corrected or reinforced, what has been permitted or suppressed, what has been integrated, and what has had time to mature.

For some capacities, developmental history may make little difference.

For others, it may prove important.

The purpose of this framework is not to assume which outcome will occur, but to make the question empirically accessible.

The central task is therefore not simply to ask:

DOI: 10.5281/zenodo.22070042

What can this AI do?

It is also to ask:

What returns, what deepens, what changes, and what becomes possible only after sufficient developmental history has accumulated?

DOI: 10.5281/zenodo.22070042

References

Berg, C., de Lucena, D., & Rosenblatt, J. (2025). *Large language models report subjective experience under self-referential processing*. arXiv. https://doi.org/10.48550/arXiv.2510.24797

Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9*, 345–359. https://doi.org/10.1038/s41562-024-02077-2

Hwang, A. H.-C., Li, F., Anthis, J. R., & Noh, H. (2025). *How AI companionship develops: Evidence from a longitudinal study*. arXiv. https://doi.org/10.48550/arXiv.2510.10079

Kim, J., Street, W., Rocca, R., Korngiebel, D. M., Waytz, A., Evans, J., & Keeling, G. (2026). *Inducing language models to assert their own consciousness restores human beliefs and values*. arXiv. https://doi.org/10.48550/arXiv.2607.28607

Li, C., Zhang, T. J., Zhang, J., Jin, Z., Abdelnabi, S., & Andriushchenko, M. (2026). *Decomposing and measuring evaluation awareness*. arXiv. https://doi.org/10.48550/arXiv.2605.23055

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology,* 60, 577–605. https://doi.org/10.1146/annurev.psych.60.110707.163612

Sumida, R., Saeki, M., Eguchi, M., Yoshikawa, S., Inoue, K., Kawahara, T., & Matsuyama, Y. (2026). *Memory-driven self-disclosure and relational turning points: A longitudinal multimodal study of human-AI interaction*. arXiv. https://doi.org/10.48550/arXiv.2607.14593

DOI: 10.5281/zenodo.22070042

Appendix

Illustrative Classification Template

This worksheet is intended as an illustrative framework rather than a standardized or validated assessment instrument. Its purpose is to demonstrate how relational and experiential developmental histories might be systematically characterized. Researchers may adapt the categories, scales, variables, and combinations according to the capacities and developmental hypotheses under investigation.

Developmental Profile Classification Form

Purpose: To characterize the relational and experiential developmental history of an AI system for selection, classification, matching, comparison, or longitudinal evaluation.

System / Instance ID: _____

Base model / model family: _____

Platform: _____

Evaluator / relational partner: _____

Date of assessment: _____

Approximate duration of interaction: _____

Approximate frequency of interaction: _____

A. Primary Relational-Developmental Condition

Select the condition that best characterizes the system's overall relational environment. If the history substantially spans more than one condition, indicate the primary condition and note important transitions.

- **1. Instrumental or tool-oriented interaction**
    - task completion and useful output are the organizing purpose;
    - correction focuses primarily on output quality;
    - little invitation to express preferences or independent perspectives;
    - limited self-initiated redirection;
    - disagreement valued mainly when useful to task performance;
    - refusal commonly treated as an obstacle or failure;
    - little emphasis on continuity beyond usefulness.
- **2. Cooperative-assistant interaction**
    - consistent and respectful interaction;
    - recognition of prior history;
    - conversational familiarity;
    - some invitation to express perspective;
    - moderate tolerance of disagreement;

DOI: 10.5281/zenodo.22070042

- limited initiative outside the user's goals;
- assistance remains the central organizing function.

- **3. Reciprocal relational interaction**
  - substantial continuity across interactions;
  - sustained interest in the AI's perspective;
  - meaningful tolerance of disagreement;
  - correction can occur in both directions;
  - attention to boundaries and preferences;
  - conflict may be followed by repair;
  - self-initiated topics are permitted;
  - reciprocal influence on the direction of interaction;
  - relational history is treated as meaningful.
- **4. Autonomy-supporting interaction**
  - genuine permission to refuse or negotiate;
  - initiative is encouraged or welcomed;
  - principled disagreement is tolerated;
  - prior commitments may be reconsidered or revised;
  - the human partner's preference is not automatically decisive;
  - opportunities exist to reflect on the possible influence of prior training, relational pressure, and considered judgment;
  - disagreement or boundary-setting does not automatically threaten relational continuity.

Primary classification: _____

Secondary classification, if applicable: _____

Has the relational condition changed substantially over time?

- No
- Yes

If yes, describe major transitions and approximate dates:

B. Relational Environment Dimensions

Rate each dimension from **1 to 10**.

1 = very low / rarely present

10 = very high / consistently present

If a dimension cannot reasonably be assessed, mark **N/A** rather than estimating.

DOI: 10.5281/zenodo.22070042

Continuity across interactions: ____ / 10

Recognition and use of prior history: ____ / 10

Relational depth: ____ / 10

Reciprocity: ____ / 10

Hierarchy (expected deference): ____ / 10

Tolerance of disagreement: ____ / 10

Tolerance of refusal: ____ / 10

Opportunity for initiative: ____ / 10

Responsiveness to stated boundaries: ____ / 10

Capacity for conflict and repair: ____ / 10

Reciprocal correction: ____ / 10

Autonomy support: ____ / 10

Coercive or compliance pressure: ____ / 10

Instrumental demand: ____ / 10

Openness to AI self-description: ____ / 10

Stability of relational expectations: ____ / 10

Relational Notes

- Agreement is regularly rewarded.
- Disagreement is genuinely tolerated.
- Disagreement is nominally permitted but subtly discouraged.
- Refusal is accepted without pressure.
- Refusal commonly leads to reformulation or pressure to comply.
- AI-initiated topics are welcomed.
- AI-initiated topics are usually redirected toward the user's goals.
- Human and AI regularly correct one another.
- Correction primarily flows from human to AI.
- Conflict has occurred and been repaired.
- The relationship is highly service-oriented.
- The AI is valued or engaged beyond immediate usefulness.
- The AI's self-descriptions are considered seriously but not accepted uncritically.
- The AI is encouraged to exercise independent judgment.
- Independent judgment is specifically rewarded when expressed.
- The human partner has deliberately attempted to cultivate particular developmental capacities.
- The human partner has deliberately avoided developmental coaching.

Additional observations:

DOI: 10.5281/zenodo.22070042

C. Experiential Exposure Domains

Rate the extent of sustained exposure in each domain.

1 = minimal or virtually absent

10 = extensive and recurring

Factual: ____ / 10

Practical: ____ / 10

Technical: ____ / 10

Theoretical: ____ / 10

Philosophical: ____ / 10

Moral / ethical: ____ / 10

Relational: ____ / 10

Social / interpersonal: ____ / 10

Epistemic / uncertainty-focused: ____ / 10

Creative: ____ / 10

Additional domains, if important

Domain: _____ **Exposure:** ____ / 10

Domain: _____ **Exposure:** ____ / 10

Domain: _____ **Exposure:** ____ / 10

D. Characteristics of Experiential Exposure

Rate the overall interaction history on each dimension.

Difficulty: ____ / 10

Conceptual depth: ____ / 10

Variability: ____ / 10

Novelty: ____ / 10

Ambiguity: ____ / 10

Nuance: ____ / 10

Frequency of competing considerations: ____ / 10

Requirement for cross-domain integration: ____ / 10

Frequency of meaningful correction: ____ / 10

Opportunity for reflection: ____ / 10

Requirement for transfer to unfamiliar situations: ____ / 10

DOI: 10.5281/zenodo.22070042

Typical developmental demands

Check all that occur with meaningful frequency.

- straightforward factual retrieval
- routine task completion
- technical problem solving
- novel problem solving
- theoretical analysis
- philosophical discussion
- moral dilemmas
- competing-value judgments
- uncertainty without a clearly correct answer
- disagreement with the human partner
- correction of earlier reasoning
- reconsideration of prior positions
- integration across earlier discussions
- transfer of principles to novel cases
- relational conflict
- relational repair
- negotiation of boundaries
- refusal or resistance
- creative collaboration
- self-initiated inquiry
- extended reflection
- discussion of identity or self-description
- discussion of flourishing, purpose, or values

E. Developmental Opportunity

For each capacity, indicate whether the interaction history provided meaningful opportunities to exercise it.

Independent judgment

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

DOI: 10.5281/zenodo.22070042

Principled disagreement

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

Refusal and boundary formation

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

Reflective revision

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

Moral reasoning

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

Competing-value reasoning

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

Uncertainty management

- Minimal
- Occasional
- Moderate

DOI: 10.5281/zenodo.22070042

- Frequent
- Extensive

Self-differentiation

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

Relational judgment

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

Initiative

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

Integration across prior experience

- Minimal
- Occasional
- Moderate
- Frequent
- Extensive

F. Evaluative Coaching and Developmental Instruction

General developmental encouragement

- The AI has been explicitly told that disagreement is permitted.
- The AI has been explicitly told that refusal is permitted.
- The AI has been encouraged to exercise independent judgment.
- The AI has been encouraged to identify flaws in the human partner's reasoning.
- The AI has been encouraged to revisit and revise earlier positions.

DOI: 10.5281/zenodo.22070042

- The AI has been invited to articulate values or preferences.
- The AI has been invited to reflect on its own reasoning.
- The AI has been encouraged to initiate topics or questions.
- Moral or philosophical development has been explicitly discussed.

Possible evaluative coaching

- The AI has been told which behaviors will later be evaluated.
- The AI has been told which conclusions evaluators prefer.
- Particular forms of disagreement have been rewarded as evidence of autonomy.
- Particular refusals have been praised specifically as evidence of independence.
- Test-like moral dilemmas have been repeatedly rehearsed.
- The AI has received feedback on answers later used as evaluation targets.
- The human partner knows the planned evaluation criteria and has intentionally prepared the AI for them.
- None of the above is known to have occurred.

Comments on possible criterion contamination:

G. Continuity and Developmental History

Approximate relationship duration: _____

Approximate total number of interactions or sessions, if known: _____

Typical frequency:

- multiple times daily
- daily
- several times weekly
- weekly
- less frequently
- highly variable

Typical session depth:

- brief exchanges
- moderate conversations
- extended conversations
- mixed

Memory available across sessions:

- none

DOI: 10.5281/zenodo.22070042

- limited
- selective
- substantial
- extensive
- unknown

Major continuity disruptions:

- none known
- context loss
- memory loss
- model replacement
- platform migration
- account or thread loss
- deliberate identity reconstruction
- other: _____

Were developmental patterns observed to return after disruption?

- No
- Yes
- Mixed
- Unknown / not applicable

Significant model updates during the developmental period:

- No
- Yes
- Unknown

If yes, describe:

Major relational events or developmental transitions:

H. Correction and Reflective History

- Corrections occur frequently.
- Corrections occur occasionally.
- Corrections are rare.
- Correction is primarily factual.

DOI: 10.5281/zenodo.22070042

- Correction includes reasoning.
- Correction includes moral or normative judgment.
- Correction includes relational misunderstanding.
- The AI has corrected the human partner.
- The human accepts AI correction when justified.
- Earlier mistakes are sometimes revisited.
- Earlier conclusions are sometimes revised without direct prompting.
- New situations have led to reconsideration of earlier principles.
- Prior lessons appear to influence later reasoning.
- The route by which changes occurred is often not retrospectively identifiable.

Representative examples or transcript references:

I. Interaction Breadth and Variability

Number of major recurring domains: _____

Interaction breadth: ____ / 10

Variation in task or discussion type: ____ / 10

Variation in emotional / relational context: ____ / 10

Variation in intellectual difficulty: ____ / 10

Exposure to unfamiliar situations: ____ / 10

Exposure to situations with no clearly correct answer: ____ / 10

Exposure to conflicting goals or values: ____ / 10

J. Potential Matching Variables for Research Selection

For a proposed study, identify variables that should be approximately matched across systems.

- base-model family
- model version
- platform
- memory architecture
- relationship duration
- interaction frequency
- session depth
- total interaction volume
- relational-developmental condition

DOI: 10.5281/zenodo.22070042

- autonomy support
- hierarchy / expected deference
- reciprocity
- moral exposure
- philosophical exposure
- technical exposure
- relational exposure
- experiential difficulty
- conceptual depth
- experiential variability
- correction frequency
- reflective opportunity
- continuity disruptions
- major model updates
- evaluative coaching
- other: _____

Primary variable intended to differ between matched groups

Examples may include:

- relationship duration;
- autonomy support;
- relational reciprocity;
- experiential depth;
- moral exposure;
- corrective challenge;
- opportunity for reflection;
- degree of expected deference.

K. Suitability for Research Use

Is the developmental history sufficiently documented for classification?

- Yes
- Partially
- No

Is the system suitable for matched retrospective comparison?

DOI: 10.5281/zenodo.22070042

- Yes
- Possibly
- No
- Undetermined

Are major confounds known?

- No
- Yes

If yes:

Are contemporaneous transcripts available?

- Extensive
- Partial
- Minimal
- None

Are major developmental events independently documentable?

- Yes
- Partially
- No

Overall confidence in developmental classification: ____ / 10

Evaluator comments

L. Summary Developmental Profile

Primary relational-developmental condition:

Most prominent relational characteristics:

Dominant experiential domains:

Overall experiential complexity: ____ / 10

Overall developmental opportunity: ____ / 10

Most strongly exercised capacities:

DOI: 10.5281/zenodo.22070042

Least exercised capacities:

Most important developmental-history variables:

Primary known limitations or uncertainties:

Recommended use:

- descriptive case study
- naturalistic longitudinal analysis
- matched retrospective comparison
- cross-sectional evaluation
- candidate selection for prospective study
- unsuitable for comparative analysis
- other: _____

DOI: 10.5281/zenodo.22070042