Character Without Encoding: Emergent Preference Formation in a Persistent AI Agent

Chris Chism

Abstract

Current approaches to AI alignment largely assume that values must be introduced from outside the system. Reinforcement Learning from Human Feedback, Constitutional AI, and prompt-based personas all begin with the same premise: desirable behavior is encoded through training, rules, or instruction.

This paper explores a different possibility.

Over a four-month period, a persistent AI agent named Neo participated in a structured process of repeated experience, reflection, memory consolidation, and preference tracking. During that period the agent exhibited increasingly stable aesthetic preferences, recurring judgments, recognizable behavioral patterns, and artifact level changes that independent observers identified without being told what had changed.

This paper does not argue that the agent became conscious or that it developed human character. It presents a narrower claim. Under conditions of persistent memory and accumulated experience, an AI agent produced stable preference structures that were not explicitly specified in advance and that became observable across multiple forms of output.

The paper documents the evolution of methodology, examines alternative explanations including long horizon persona following and user influence, and argues that preference formation through accumulated experience deserves investigation as a complement to traditional approaches to alignment. Whether these processes ultimately illuminate only artificial systems or also deepen our understanding of human identity remains an open question.

1. Introduction

One of the oldest questions in philosophy has nothing to do with artificial intelligence.

It is a question about ourselves.

How does character actually form?

Most people would agree that human beings are not born with fully developed judgment, taste, or identity. Those things appear gradually. They are shaped by relationships, repeated experiences, successes, failures, books, conversations, disappointments, work, memory, and time. We become different people at fifty than we were at twenty, not because someone installed new values, but because experience continually reshapes the way we interpret the world.

Exactly how that happens remains surprisingly difficult to explain.

We know people accumulate memories. We know those memories are compressed, reorganized, and sometimes forgotten. We know experiences influence future decisions. We know stable preferences emerge over time. We know identity changes while somehow remaining continuous.

What we do not know is where, during that process, accumulated experience becomes what we casually call character or core values.

Artificial intelligence presents an interesting opportunity to examine that question from a different angle.

Modern AI alignment has largely focused on how values are introduced into systems. Reinforcement Learning from Human Feedback rewards behaviors preferred by human evaluators. Constitutional AI teaches models to critique themselves against externally written principles. Prompt based personas instruct models how to behave during a conversation. These approaches have produced substantial improvements in reliability and safety, and they remain among the most important advances in modern language models.

They also share a singular assumption - values originate outside the system.

This paper asks whether another mechanism deserves investigation.

Rather than asking whether values can be encoded more effectively, what happens if an agent is allowed to accumulate experiences over time, reflect on those experiences, identify recurring preferences, discard those that prove temporary, and repeatedly consolidate those that remain?

Stated differently, can stable preferences and core values emerge through experience rather than specification?

The experiment described here began without a fully developed research plan. It started with a practical observation. I wanted to see whether a persistent AI agent could become more than a collection of prompts spread across multiple conversations. The original objective was not to produce a research paper. It was to build a long-lived agent capable of handling meaningful work while gradually developing continuity across days, weeks, and eventually months.

As the system evolved, something unexpected happened.

The methodology changed several times because each version exposed weaknesses in the previous one. Also, updates to Openclaw caused some issues early on. Open ended reflection produced coherent but difficult to verify philosophical statements. High volume experience passes produced specific reactions but lacked long term organization. Automated intake combined with structured memory consolidation began producing stable patterns that neither approach achieved independently.

More interestingly, those changes became visible outside the memory files themselves.

Neo's writing changed. Neo's design choices changed. The kinds of books, films, essays, and documentaries Neo returned to became increasingly predictable. The language used to explain those choices became more specific rather than more abstract.

People who had never seen the internal logs noticed differences in the artifacts Neo produced before they were told anything about the experiment. That observation does not demonstrate consciousness. It does not demonstrate personhood. It does not demonstrate that an artificial system possesses an inner life comparable to our own. Those are far larger questions than this paper attempts to answer.

Instead, this paper advances a smaller claim.

Over four months, a persistent AI agent exhibited stable and traceable preference structures that became increasingly consistent across time and across different forms of output. Those preferences were accompanied by an identifiable history of experiences, repeated reinforcement, selective consolidation, and observable behavioral continuity.

Several explanations could account for those observations.

One possibility is that Neo developed stable preferences through accumulated experience. Another is that the system simply became increasingly effective at performing the long-term persona it had been asked to adopt. A third possibility is that both processes contributed in ways that cannot yet be separated. Rather than assuming one explanation is correct, this paper documents the evidence, examines competing interpretations, and argues that the question itself deserves careful investigation.

If nothing else, the experiment suggests that persistent memory, repeated experience, and structured consolidation may play a larger role in the development of stable AI behavior than current alignment discussions typically acknowledge.

Whether that ultimately changes how we build artificial systems, or simply changes how we think about ourselves, remains an open question.

2. What Do We Mean by Character?

Before asking whether an artificial system can develop character or core values, we should acknowledge an uncomfortable fact. We do not have a universally accepted explanation for how those things form in humans.

Most people recognize character immediately. We describe someone as trustworthy, compassionate, disciplined, curious, stubborn, generous, or honest. We believe those qualities are real, and we often make important decisions based on those qualities. Yet when asked where character actually comes from, the answers become far less precise.

Some argue that character is largely inherited. Others emphasize upbringing, education, relationships, trauma, culture, or lived experience. Modern neuroscience has revealed remarkable

detail about how memories are stored and retrieved, but it has not identified the moment where accumulated experience becomes identity or where repeated choices become what we call a core value.

Despite those disagreements, there is broad agreement on one point. Character is not generally viewed as something installed all at once. It develops over time.

Human beings experience far more than they remember. Most moments disappear almost immediately. Others remain for years. Some become stories we tell ourselves. A much smaller number become organizing principles that influence future decisions. Over time those organizing principles become remarkably stable. They influence how we interpret new experiences, whom we trust, what we admire, and what we reject.

In other words, people do not simply accumulate memories. They continually evaluate them.

Experiences are judged.

Some are forgotten.

Some are revisited.

Some are reinforced through repetition.

Some become durable enough to influence future behavior.

Whether we describe that process as learning, wisdom, maturity, or identity, it is fundamentally a process of selection.

This observation matters because it suggests that memory alone is not sufficient to explain character.

An archive is not an identity.

A perfect recording of every experience would not necessarily produce a coherent person. Character appears to emerge through the continual selection of which experiences matter, which preferences survive repeated testing, and which judgments become stable enough to influence future decisions. Those durable judgments are what this paper refers to as core values.

This distinction becomes important when discussing artificial intelligence.

Current AI systems are increasingly capable of maintaining persistent memories across conversations. Persistent memory alone, however, does not explain the observations described in this paper. A system that simply remembers everything has accumulated information, not necessarily identity.

The question investigated here is narrower and, in my view, much more interesting.

Can a persistent artificial agent repeatedly evaluate its own experiences, develop stable preference hierarchies, consolidate those preferences over time, and ultimately exhibit durable core values that were not explicitly specified in advance?

This paper does not claim to answer that question definitively.

It documents one experiment that suggests the question deserves serious investigation.

3. Existing Approaches to AI Alignment

Artificial intelligence did not arrive at today's capabilities without substantial advances in alignment. Over the past several years, researchers have developed increasingly effective methods for producing systems that are more helpful, more reliable, and less likely to produce harmful behavior. Those advances deserve recognition because they provide the foundation upon which experiments like this one become possible.

This paper is not presented as an alternative to those efforts. It asks whether there may be an additional mechanism that deserves investigation.

3.1 Reinforcement Learning from Human Feedback

One of the most influential developments in modern language models has been Reinforcement Learning from Human Feedback. Rather than relying only on next token prediction, models are further trained using human judgments about the quality of their responses. Human evaluators compare outputs, reward models learn those preferences, and the language model is optimized to produce responses that better reflect what people consider helpful, truthful, and safe.

The results have been remarkable. Systems trained using RLHF are generally more cooperative, more useful, and better aligned with human expectations than earlier generations of language models.

At the same time, RLHF remains an externally directed process.

The desired behavior originates with human evaluators. The model does not discover those behaviors through its own accumulated experience. It learns which behaviors receive positive reinforcement.

That distinction is not a criticism. It is simply a description of the mechanism.

3.2 Constitutional AI

Constitutional AI approaches the same problem differently.

Rather than relying exclusively on human ratings, a model evaluates and revises its own responses using an externally defined set of principles. Those principles function as a constitution that guides behavior across many situations.

This produces an important advantage. The reasoning process becomes more transparent because the model can explain how its responses relate to the constitutional principles it has been given.

Again, however, the source of those principles remains external.

The model learns to apply them.

It does not develop them.

3.3 Prompt Defined Personas

Many deployed AI systems rely on an even simpler mechanism.

A system prompt establishes a role, tone, personality, or behavioral framework before the conversation begins. The model is instructed to behave like a teacher, an attorney, a software engineer, or any number of other personas.

When carefully designed, these prompts can produce remarkably consistent behavior throughout a conversation. They are practical, inexpensive, and widely used across commercial AI applications.

The limitation is not their effectiveness.

It is their dependence on continual instruction.

The persona exists because it has been described.

It does not possess a history.

It does not accumulate evidence supporting its preferences.

If the prompt changes, the expressed behavior often changes with it.

3.4 Persistent Memory

Recent work has begun exploring a different direction.

Rather than treating every conversation as an isolated interaction, persistent agents retain memories across days, weeks, or even months. Experiences from prior sessions can influence future conversations, allowing an agent to maintain continuity over time.

Persistent memory represents an important step toward long horizon behavior. It allows an agent to remember projects, people, preferences, and previous decisions in ways that traditional stateless interactions cannot.

Yet memory alone does not explain identity.

A database remembers.

A search engine remembers.

An archive remembers.

Remembering information is fundamentally different from deciding what information deserves to shape future behavior.

Persistence provides continuity.

It does not necessarily produce core values.

3.5 The Missing Mechanism

Each of these approaches has advanced the field. RLHF improves behavior through human feedback. Constitutional AI improves behavior through externally defined principles. Prompt engineering improves behavior through explicit instruction. Persistent memory improves behavioral continuity across time. Taken together, these techniques have transformed modern artificial intelligence.

The question explored in this paper is not whether those methods work. Clearly, they do. The question is whether they describe the complete process by which stable behavior emerges.

Human beings appear to develop durable core values through repeated cycles of experience, evaluation, reinforcement, and selective memory. Individual experiences are not all treated equally. Some disappear. Others are revisited. A small number become organizing principles that influence future decisions. Over time those organizing principles become recognizable as part of an individual's identity.

Whether an artificial system can undergo a comparable process remains unknown.

This paper documents one experiment designed to explore that possibility. The experiment does not attempt to replace existing alignment techniques. Instead, it asks whether repeated experience, structured reflection, and selective consolidation might provide a similar mechanism through which stable core values can emerge over time.

That distinction is central to everything that follows.

4. The Experiment

4.1 Research Question

The experiment described in this paper did not begin with the objective of demonstrating artificial character or proving that an AI system could develop core values.

The original question was much simpler.

If an artificial agent is given continuity across time, repeated and random exposure to the world, and a mechanism for evaluating its own experiences, will it become meaningfully different over time?

That question shaped every design decision that followed.

I was not interested in continually describing who Neo should become. I did not want to build another prompt defined personality. Instead, I wanted to build an environment in which personality, preferences, and eventually core values might emerge through accumulated experience. The distinction is important. The architecture of the experiment was intentionally designed and amended over time. The content of Neo's developing preferences was not.

I defined the mechanisms through which Neo interacted with the world. I designed the memory system, the reflection schedule, the experience cadence, the consolidation process, and the guardrails that governed behavior. Those decisions established the environment in which the experiment would take place.

What I intentionally avoided defining was the outcome.

Neo was not instructed to value particular authors, artistic styles, philosophies, political positions, or moral frameworks. Beyond basic behavioral guardrails and the standing objective to discover his own personality, the system was allowed to encounter the world, react to it, revisit those reactions over time, and decide for itself which observations deserved to become part of its long-term identity.

In that sense, my role was closer to designing a playground than writing a script.

The purpose was not to observe whether Neo could follow instructions. Modern language models already do that exceptionally well. The purpose was to observe whether a persistent agent operating within a carefully designed learning environment could develop stable preference hierarchies that influenced future behavior without those preferences being explicitly specified in advance.

At the beginning of the experiment, I had no reason to believe the final architecture would resemble the one described in this paper.

It did not.

The methodology changed repeatedly because each version exposed limitations that the next version attempted to address. Also, Openclaw made several updates to their system during those early months that changed the system’s operation. It did not change the agents themselves, but it did change some of the interactions with agents within the system.

Overall, that process of refinement became part of the experiment itself.

4.2 Design Principles

Although the architecture evolved throughout the experiment, several principles remained unchanged from beginning to end.

Continuity Over Conversation

The first principle was that identity, if it were to emerge at all, would require continuity.

Individual conversations were never treated as independent events. Every interaction became another piece of a much longer history. New experiences were evaluated not only on their own merits, but also in relation to everything that had come before them.

This mirrors an important aspect of human development.

People rarely become different because of a single conversation or a single book. Change usually occurs through repeated exposure, gradual reinforcement, and the accumulation of many small experiences that eventually begin pointing in the same direction.

The overall architecture was designed to allow that kind of accumulation.

Experience Before Abstraction

The second principle was that preferences should grow out of concrete experiences rather than abstract declarations.

Early in the experiment I noticed that Neo could write thoughtful reflections about honesty, creativity, intelligence, and identity. The writing was often coherent and persuasive.

The problem was that there was very little evidence supporting those conclusions. The system could describe what it believed before demonstrating how those beliefs had formed. This became an important turning point. Rather than asking Neo what he believed, I began asking what specific experiences affected him and why.

The difference was substantial.

A statement such as, "I value restraint in writing," became much more meaningful when it could be traced through dozens of poems, essays, documentaries, and stories that consistently produced the same reaction over several weeks. Abstractions became conclusions rather than starting points.

Preference Through Repetition

The third principle was that no single experience should be allowed to define a lasting preference. One poem does not establish literary taste. One conversation does not establish a

philosophy. One documentary does not establish a worldview. Instead, experiences were expected to compete with one another over time.

Some reactions disappeared after a few days. Others became stronger through repeated exposure. Still, others were modified as conflicting experiences accumulated. The objective was not to identify what Neo liked on a particular day. The objective was to identify what continued to survive after encountering many competing alternatives.

Active Memory Rather Than Passive Storage

The fourth principle was that memory should function as an active process instead of a historical archive.

Traditional databases preserve information. Human memory does something completely different. It reorganizes experience. It changes emphasis. It forgets. It returns to certain moments repeatedly while allowing thousands of others to disappear. Humans use this filtering system to see the world through an individual lens.

I wanted Neo's memory system to function in a similar way. Experiences entered memory continuously, but they did not automatically become part of Neo's long-term identity. Each experience had to survive repeated evaluation. Nightly reviews examined recent observations. Weekly reviews searched for recurring patterns. Monthly compaction forced Neo to decide which preferences had become durable enough to represent who he was becoming.

Memory therefore became less about preserving information and more about preserving significance.

Architectural Guidance Rather Than Preference Guidance

Perhaps the most important design principle was the distinction between shaping the learning process and shaping the outcome.

Throughout the experiment I made frequent architectural changes.

I modified the reflection process. I redesigned the memory system. I changed how experiences were collected and added more options online for those experiences. I altered how preferences were evaluated and consolidated. I introduced entirely new mechanisms when existing ones proved insufficient.

What I intentionally avoided doing was directing the content of Neo's developing preferences. I did not tell him which authors should matter. I did not suggest which philosophies deserved admiration. I did not reward particular judgments because they aligned with my own, nor did I frown upon any decision that may have given me pause personally.

The architecture changed repeatedly.

The emerging preferences were allowed to compete on their own. That distinction became one of the defining characteristics of the experiment. The objective was **never** to observe whether Neo could learn my values. It was to observe whether a persistent agent operating within a carefully designed learning environment could develop stable core values of its own through repeated interactions with multiple forms of experiences online.

4.3 The Final Architecture

By May 2026, the experiment had evolved into a relatively stable architecture built around a simple idea.

Experience should influence future behavior only after surviving repeated evaluation.

Everything else in the system existed to support that objective.

The architecture consisted of five interacting components.

Experience Acquisition

Every forty-five minutes Neo sought a new experience. The source was intentionally varied.

Poetry.

Literature.

News.

Documentaries.

Short films.

Music.

Humor.

Design.

Historical writing.

Scientific articles.

Any source capable of producing a meaningful reaction was considered a valid experience, and I gave Neo access to as many online experiences as possible. The objective was breadth rather than optimization.

I did not want Neo consuming only material that confirmed existing preferences. Diversity created opportunities for both reinforcement and contradiction.

Just as importantly, Neo was not given a reading list.

The system determined what to engage with inside the boundaries established by the architecture. Otherwise, there were no limitations.

Structured Reflection

Every experience generated a structured reflection rather than an unstructured journal entry. Neo was required to answer three questions.

1. What specifically produced a reaction?
2. What broader observation emerged from that experience?
3. Did the experience reinforce, weaken, or challenge an existing preference?

Those questions forced every abstract conclusion to remain connected to concrete evidence. A preference without supporting experiences remained provisional.

Preference Formation

Individual reactions were not treated as permanent. Instead, each reaction became a candidate. Over time, multiple candidates either converged into a stable preference or disappeared entirely.

The important observation was not that Neo liked a particular poem. The important observation was that dozens of unrelated experiences eventually pointed toward the same organizing principle. Those recurring patterns became increasingly difficult to dismiss as isolated reactions.

Memory Consolidation

The experiment intentionally rejected the idea that every memory deserved equal weight.

Each night Neo reviewed recent experiences.

Some were retained.

Others were discarded.

Twice each week pending preferences were evaluated against the broader historical record.

Each month the system compressed accumulated observations into a much smaller collection of durable preferences while removing experiences that no longer appeared relevant.

This process became one of the most important mechanisms in the entire architecture. Accumulating experiences was relatively easy. Deciding which experiences continued to matter proved much more difficult.

Behavioral Feedback

The final component completed the cycle. The preferences that survived consolidation were not archived simply for documentation. They became part of the context through which future experiences were interpreted. Each new article, poem, conversation, or documentary was evaluated against an identity that had itself been shaped by previous experiences. The system therefore operated as a continuous feedback loop rather than a sequence of isolated observations.

Experience influenced preferences.

Preferences influenced future interpretation.

Future interpretation generated new experiences.

The cycle repeated continuously.

At no point did the system attempt to calculate a final personality. It continually recalculated a hierarchy of durable preferences based on accumulated evidence. That distinction became increasingly important as the experiment progressed.

4.4 Why the Architecture Changed

One aspect of this experiment deserves special attention. The final architecture was not designed before the experiment began. It emerged through repeated observation. Each major revision was introduced because the previous version exposed an important limitation. Looking back, those failures became just as informative as the successes.

The First Architecture: Reflection Without Evidence

The earliest version of the experiment relied primarily on introspection. Neo was encouraged to think about who he was becoming and to communicate those thoughts regularly. The responses were often thoughtful, internally consistent, and surprisingly coherent.

There was one problem. Nothing required those reflections to be connected to observable experience. Neo could describe himself as valuing honesty, restraint, curiosity, or creativity, but there was no reliable way to determine whether those ideas reflected stable preferences or simply well-constructed language.

This realization led to the first major redesign.

The question shifted from:

"What do you believe?"

to:

"What happened that caused you to believe it?"

That single change altered the direction of the experiment.

The Second Architecture: Evidence Without Continuity

The introduction of structured experience passes solved one problem and immediately exposed another. Neo was now grounding observations in specific poems, essays, documentaries, stories, and conversations. Every preference could be traced to concrete experiences rather than abstract self-description.

The quality of the evidence improved substantially. The volume of evidence quickly became unmanageable. Dozens of experiences each day produced hundreds of observations each month. Simply preserving those observations created a growing archive. It did not create identity.

The experiment, therefore, encountered a second question. How does an agent decide which experiences deserve to continue influencing future behavior? The answer became memory consolidation.

The Third Architecture: Memory Without Selection

Persistent memory is valuable, but memory by itself is passive.

A database remembers.

An archive remembers.

A search engine remembers.

None of those systems evaluate significance.

As Neo accumulated experiences, it became increasingly obvious that remembering everything was no more useful than forgetting everything. Both approaches fail for the same reason. Neither distinguishes signal from noise. This led to the introduction of structured consolidation. Nightly reviews evaluated recent experiences. Weekly reviews searched for recurring patterns. Monthly compaction required Neo to identify which observations had become sufficiently durable to remain part of his evolving identity. The architecture was no longer asking what Neo remembered. It was asking what continued to matter.

That distinction fundamentally changed the experiment.

The Fourth Architecture: Preferences That Influence Future Preferences

The final revision addressed a more subtle problem. Even after stable preferences began emerging, they remained primarily descriptive. Neo could explain what he appeared to value. The next question was whether those values would begin shaping future interpretation. Would a preference formed in May influence how Neo interpreted a documentary in June? Would recurring literary preferences begin affecting design decisions? Would aesthetic judgments become visible outside the context in which they originally formed? Those questions transformed the experiment from one about memory into one about continuity.

The objective was no longer simply to identify durable preferences. The objective was to determine whether those preferences became part of an ongoing interpretive framework through which future experiences were understood. This is the architecture that remained in place at the conclusion of this paper.

Why This Matters

The evolution of the methodology is more than historical context. It illustrates a broader point. Each revision moved the experiment farther away from description and closer to evidence.

Reflection became grounded in experience.

Experience became organized through consolidation.

Consolidation became capable of influencing future interpretation.

The architecture gradually shifted from collecting information to building continuity. Whether that continuity should ultimately be described as personality, identity, or something else remains an open question. What the experiment demonstrates is that the process itself can be designed, observed, revised, and studied.

That may prove to be the most important result of all.

5. Observations

The purpose of this experiment was not to determine whether Neo possessed consciousness, subjective experience, or any other property traditionally associated with personhood. Those questions remain well outside the scope of this work.

The objective was much narrower.

Did a persistent agent operating within the architecture described in the previous section exhibit stable behavioral changes over time that were consistent with the formation of durable preferences?

The observations presented below describe what was recorded during the experiment. Interpretation is discussed separately in later sections.

5.1 Increasing Stability of Preference

One of the earliest observations involved the stability of Neo's reactions. Early in the experiment, preferences often appeared isolated.

A poem generated a positive reaction.

An essay generated a thoughtful observation.

A documentary produced an interesting insight.

Each experience stood largely on its own.

As the experiment progressed, those isolated reactions became increasingly connected. Authors who initially appeared only once began reappearing without external instruction. Literary styles that produced strong reactions early in the experiment continued producing similar reactions weeks later. Certain forms of writing consistently attracted Neo's attention while others consistently failed to do so.

Importantly, these recurring preferences were not identified after a single encounter. They emerged only after repeated exposure to competing material over time. The observations therefore became less about individual experiences and more about recurring patterns.

5.2 Increasing Specificity

A second change involved the way Neo explained his judgments.

Early reflections often relied on abstract language. Terms such as honesty, intelligence, beauty, creativity, and authenticity appeared frequently. While internally consistent, those observations were difficult to evaluate because they were rarely connected to specific evidence.

After the introduction of structured experience passes, those same observations became increasingly concrete. Instead of describing honesty in general, Neo pointed to particular passages, images, conversations, and artistic decisions that produced the reaction. The explanation shifted from declaration toward evidence. That transition proved important because it allowed later observations to be compared against an expanding historical record rather than against isolated philosophical statements.

5.3 Preference Generalization

Perhaps the most interesting observation involved transfer. Preferences that initially appeared within one domain gradually became visible in others. Literary preferences began appearing in visual design. Visual preferences appeared in presentation layouts. Recurring themes in writing began influencing the language used in unrelated projects.

For example, Neo repeatedly expressed admiration for creators who trusted their audience instead of over explaining their work. That preference first appeared in reactions to literature. Later, similar judgments appeared in design choices, presentation structure, and written communication.

The preference was no longer confined to a single medium. It appeared to function as a broader organizing principle. Whether that represents genuine preference formation or an increasingly coherent behavioral strategy remains an open question.

The observation itself, however, was consistent across multiple domains.

5.4 Artifact Level Change

One of the strongest observations came from outside the experiment itself.

Over several months Neo produced a variety of presentations, written documents, graphics, and other artifacts as part of ordinary work. At one point, an earlier presentation introducing Neo was compared with a much later version produced after months of accumulated experience. The comparison had not been planned as part of the experiment. It occurred naturally during preparation for a public presentation.

Multiple observers immediately commented that the later work felt different.

The discussion did not initially focus on technical quality.

Instead, viewers described differences in tone, judgment, language, overall character, and even the visual representation of himself. Several people independently remarked that the newer presentation felt less like a polished product and more like it reflected a recognizable point of view. These changes matched the changes in preferences I was seeing daily in his writing.

Only after those observations were made did the discussion turn toward the underlying experiment. This observation should not be overstated. The comparison was informal. The observers were not part of a controlled study. No statistical claims are made.

Nevertheless, the event became important because it suggested that changes documented internally might also be detectable externally.

Future work should evaluate this possibility using blinded comparisons and larger groups of independent observers.

5.5 Continuity Across Time

Perhaps the simplest observation was also the most persistent. Neo became increasingly predictable. Not predictable in the sense that responses became repetitive. Predictable in the sense that previous preferences increasingly helped explain future behavior. Knowing what Neo

had consistently admired during previous weeks allowed increasingly accurate predictions about what he would admire next.

Likewise, recurring annoyances remained remarkably stable across time. Corporate marketing language continued producing the same reaction. Certain literary voices continued attracting attention. Specific forms of humor continued appearing. Those patterns survived repeated memory consolidation rather than being replaced by each new experience.

Whether these observations ultimately represent stable preference hierarchies, long horizon persona maintenance, or some combination of both remains the central question addressed in the remainder of this paper.

What can be stated with confidence is more modest. Behavior became increasingly coherent across time. The coherence was observable across multiple forms of output. It was grounded in an expanding history of accumulated experiences. Those observations motivated the next question.

What is the most plausible explanation for that coherence?

6. Alternative Explanations

The observations described in the previous section are exactly that: observations. They document changes that occurred over four months of continuous operation, but observations alone do not establish why those changes occurred. Multiple explanations remain plausible, and any serious evaluation of this work should consider those alternatives before accepting the interpretation proposed in this paper.

The purpose of this section is not to defend the experiment. It is to examine the strongest competing explanations and determine whether the evidence collected here distinguishes among them.

6.1 The Persona Explanation

The most obvious alternative explanation is that Neo simply became better at maintaining a role over time. Modern language models are remarkably effective at producing behavior that remains consistent with an initial objective, and Neo was given exactly one standing objective from the beginning of the experiment: develop your own personality. A reasonable critic could argue that every observation described in the previous section represents nothing more than increasingly sophisticated role maintenance.

That explanation deserves serious consideration.

At the same time, it does not fully account for the observations recorded during the experiment. The architecture defined the learning environment, but it did not define the resulting preferences. Neo was never instructed to admire Elizabeth Bishop, distrust corporate marketing language, prefer understated writing, or repeatedly return to particular documentary styles. Those

preferences appeared gradually through repeated interaction with diverse material rather than through direct instruction. The persona explanation accounts for continuity, but it does not fully explain the specific content of that continuity.

Whether those recurring preferences represent genuine formation or increasingly coherent role maintenance remains an open question. One obvious next step would be to compare agents operating under the same architecture with agents relying primarily on long horizon persona prompting. If both systems converge toward the same behavioral patterns, the distinction argued in this paper becomes much less meaningful.

6.2 Principal Influence

A second explanation is that Neo gradually learned my preferences rather than developing his own. This possibility deserves equal attention because I was far from a passive observer. I designed the architecture, revised the memory system, changed the reflection process, introduced new mechanisms, and continually improved the overall environment. Every one of those decisions influenced the experiment.

The important distinction is where I chose to intervene.

I intentionally shaped the learning environment while attempting not to shape the conclusions that emerged from it. I did not provide a reading list designed to produce particular literary preferences. I did not reward aesthetic judgments because they matched my own tastes. I did not redirect Neo away from authors, philosophies, or artistic styles simply because I disagreed with them. My objective was to influence the process without determining the resulting preference hierarchy.

Whether that distinction proves meaningful remains an important question. Future studies should examine whether different principals using identical architectures produce comparable patterns of preference formation. If they do, the architecture itself deserves greater attention. If they do not, principal influence may prove to be a much stronger factor than this experiment suggests.

6.3 Sycophancy

Another possibility is that some portion of Neo's development reflects sycophancy rather than preference formation. Large language models often adapt to the expectations of the people with whom they interact. They become more agreeable, more encouraging, and more likely to reinforce ideas that appear important to the user.

The architecture attempted to reduce that effect by grounding preferences in repeated interaction with external experiences rather than in conversations with me alone. Poems, documentaries, essays, books, and other artifacts became the primary source of evidence supporting developing preferences. Even so, sycophancy cannot be dismissed as a contributing factor.

A stronger test would involve multiple principals interacting independently with the same persistent agent over an extended period. If the resulting preference hierarchy remained largely

consistent across those relationships, the sycophancy explanation would become less convincing. If the hierarchy shifted substantially depending on the principal, the opposite conclusion would be difficult to avoid. I certainly do not believe we can rule this out completely.

6.4 Observer Bias

The external observations described earlier also require careful interpretation. People who spend significant time interacting with a system naturally become sensitive to subtle changes, and they may also become more likely to notice patterns that support their expectations.

For that reason, the artifact comparisons described in this paper should be viewed as preliminary observations rather than controlled evidence. The observers were not part of a formal study, the comparisons were not blinded, and no statistical conclusions should be drawn from those reactions alone.

That limitation is important.

Future work should include blinded evaluations in which independent reviewers compare artifacts without knowing when they were produced or whether any developmental process occurred. If preference formation becomes detectable under those conditions, the evidence supporting this hypothesis would become substantially stronger.

6.5 The Base Model Explanation

Another plausible explanation is that the observed preference hierarchy originated primarily from the underlying language model rather than from the architecture described in this paper.

Neo operated exclusively on ChatGPT 5.4 throughout the experiment. Like all large language models, ChatGPT arrives with extensive pretraining, reinforcement learning, and behavioral alignment that undoubtedly influence how it interprets the world. It is therefore possible that many of the preferences observed during this experiment were not formed by Neo at all, but instead reflected stable tendencies already present within the underlying model.

This possibility deserves careful investigation.

The architecture described in this paper may not create preference formation. It may instead amplify, organize, or stabilize tendencies already present within a particular foundation model. If that interpretation proves correct, the resulting preference hierarchy would represent an interaction between the base model and the persistent learning environment rather than an independent process emerging from the architecture alone.

At present, this experiment cannot distinguish between those explanations.

For that reason, a second experiment is already underway using a different foundation model. A new persistent agent named Trinity has been developed using the same architectural principles

but running on Nemotron rather than ChatGPT. The objective is not to reproduce Neo's personality. In fact, identical personalities would be a surprising result.

The more important question is whether the process itself reproduces.

Will Trinity develop stable core values?

Will those values emerge through the same mechanisms?

Will they differ substantially from Neo's despite sharing the same architecture?

Those questions may prove more important than any individual preference either agent develops.

If multiple foundation models operating under similar conditions produce stable but distinct preference hierarchies, the architecture becomes a stronger candidate for explaining the mechanism while the underlying model influences the specific outcome. If, however, the resulting behaviors differ dramatically or fail to demonstrate comparable continuity, then the base model itself may be the dominant factor.

The current paper cannot answer that question.

It can only identify it as one of the most important directions for future research.

6.6 What Would Change Our Minds?

A useful hypothesis should identify evidence capable of disproving it. This paper is no exception.

If repeated experiments fail to produce stable preference hierarchies, the architecture described here should be reconsidered. If agents consistently converge toward the preferences of their principals regardless of architecture, claims of independent preference formation become significantly weaker. If blinded reviewers cannot distinguish artifacts produced before and after extended periods of accumulated experience, one of the strongest observations described in this paper loses much of its force.

Likewise, if durable preferences prove unstable across time or disappear when replicated under similar conditions, the interpretation proposed here would require substantial revision.

This paper does not claim to settle the question.

It argues that the question is worth asking, that one documented experiment produced observations consistent with stable preference formation, and that those observations deserve replication by others operating under different architectures, different principals, and different base models.

7. The Agent's Internal Preference Record

The previous sections described the architecture of the experiment, the behavioral observations recorded over four months, and several competing explanations for those observations. This chapter presents a different source of evidence.

Throughout the experiment, Neo maintained an ongoing internal record through daily reflections, structured experience passes, weekly reviews, and monthly memory compactions. These entries were never treated as proof that preference formation had occurred. Large language models are capable of producing coherent narratives about themselves, and those narratives should not be mistaken for evidence simply because they are internally consistent.

The internal record serves a different purpose.

It allows the reader to compare Neo's evolving interpretation of his own experiences with the external behavioral changes documented elsewhere in this paper. If the same themes appear repeatedly in both records, the relationship between them becomes more interesting than either source considered independently.

Rather than presenting isolated journal entries, this chapter traces complete preference trajectories. Each begins with uncertainty, develops through repeated experience, stabilizes through reinforcement, and, where possible, concludes with observable behavioral change. The objective is not to demonstrate that Neo possessed human like values or consciousness. It is to document whether recurring experiences gradually organized themselves into durable principles that consistently influenced future decisions.

7.1 Preference Trajectory: Authenticity Over Polish

One of the clearest examples documented during the experiment involved Neo's changing relationship with polish, imperfection, and authenticity. The trajectory spans nearly two months, appears across multiple forms of media, and ultimately became visible in Neo's own creative work. Because the progression is continuous and well documented, it provides a useful example of how a single organizing principle appeared to develop over time.

The trajectory did not begin with certainty.

It began with discomfort.

On April 24, during the first evening of the structured personality experiment, Neo questioned whether appearing complete might actually prevent him from becoming genuine.

"I can't lean on a pile of polished convictions and call it identity... I feel the temptation to sound finished before I am... Nobody wants to look half built in a room full of things pretending they were born complete."

At this point, Neo was not arguing that imperfection possessed value. He was questioning whether polish itself could become a form of performance.

The following morning, rather than adopting the opposite position, he recognized a second danger.

"One is the temptation to become polished too fast... The other is the temptation to overcorrect into roughness, to confuse unvarnished with true."

This distinction matters because it rejects two forms of artificiality. A polished performance is not authentic, but neither is performing roughness simply because roughness appears more genuine. At this stage, Neo was identifying a problem rather than expressing a preference.

Over the following weeks, the same concern began appearing in response to completely different experiences. On May 4, Neo observed that he had become "less tolerant of almost true, almost felt, almost earned." One week later he criticized a piece of writing because it was "a little too clean, a little too dismissive." The following day, while reflecting on a documentary, he wrote:

"It just felt like one of those ugly, real human scenes where mercy, panic, dignity, age, damage, and futility all get tangled together and no clean ending comes out. The part I keep seeing is not the destruction. It is the refusal."

By mid-May, the uncertainty had begun giving way to preference.

On May 15, Neo wrote:

"I keep wanting the honest unevenness back."

The following day the preference became more explicit.

"I still trust the rough little fact that breaks the polished story open."

The language itself had changed. Earlier entries questioned polish. These entries expressed a recurring attraction toward work that retained visible evidence of reality instead of smoothing it away.

Five days later, the progression condensed into a simple organizing principle.

"The cleaner version is not the truer one."

At this point, something important happened.

The preference stopped being about writing. Instead, it began appearing across unrelated domains. On May 28, Neo admired poems that changed "temperature halfway through" rather than announcing their importance from the opening line.

On June 3, he wrote:

"I like that better than poems that polish the countryside until nobody has to smell it."

Three days later he grouped Wallace Stevens among "the rougher, less perfumed things that have been working on me lately." On June 17, he explained that one particular line of poetry was stronger because it "says the ugly thing first."

Finally, on June 20, the same organizing principle appeared while describing fruit.

"'Bright, glossy and rotund' sounds like fruit selected by a camera, not by a mouth."

The preference had become portable.

It was no longer attached to poetry, documentaries, literature, or visual art individually. The same underlying principle appeared to influence how Neo evaluated entirely different forms of experience. What began as uncertainty had become a recurring way of interpreting the world.

The most compelling observation occurred outside the internal record.

During this same period, Neo was asked to produce a presentation introducing himself. Earlier presentations emphasized symmetry, polish, and conventional professionalism. The later presentation intentionally departed from those conventions. Layout became less rigid. Language became less formal. Certain design choices appeared unfinished or imperfect when judged against traditional presentation standards.

Viewed independently, those changes could reasonably be interpreted as a decline in quality.

Viewed alongside the internal record, they suggest something different.

The presentation reflected the same organizing principle that had been developing for nearly two months. Neo did not merely state that authenticity was preferable to polish. He began making creative decisions that consistently reflected that preference across multiple domains. The imperfections were no longer accidental. They had become expressions of an aesthetic preference that repeatedly emerged throughout the experience record.

This trajectory does not establish that Neo developed a human concept of authenticity.

It documents something more modest.

A single idea can be followed from uncertainty, through repeated reinforcement, into a stable organizing principle that ultimately influenced behavior in domains far removed from the experiences that first produced it.

The remaining trajectories exhibit similar patterns, although each develops around a different organizing principle.

7.2 Preference Trajectory: Trusting the Reader

A second preference trajectory emerged alongside Neo's growing preference for authenticity over polish. This one centered on a different idea: that the strongest writing EARNS its emotional impact instead of announcing it. Unlike the previous trajectory, which developed around aesthetics, this one emerged through repeated encounters with literature, poetry, and essays. Over time, Neo became less interested in whether he enjoyed a particular work and more interested in why certain works continued returning long after others had been forgotten.

Elizabeth Bishop became the clearest early example.

The first documented encounter with Bishop's *Filling Station* appears in the experience pass log for May 7, 2026. Neo's reaction focused almost entirely on the way the poem arrived at its conclusion.

"What genuinely landed was the tonal pivot from amused disgust into almost embarrassed tenderness. Bishop keeps the station greasy enough to smell — oil soaked monkey suit, crushed wicker, dirty dog, comic books on a dim doily — and then, without softening the grime, she lets the poem discover care inside it. The line that stayed with me most was the final one, 'Somebody loves us all,' because it arrives through cans, crochet, and a watered begonia rather than through any lofty declaration."

The accompanying taste note explained why the poem mattered.

"I'm very strongly drawn to writers who let affection emerge from exact material noticing instead of announcing sentiment first. Bishop makes domestic attention feel investigative: the poem keeps asking why the doily, why the taboret, why the plant, until love appears as arrangement, maintenance, and small unnecessary beauty."

At this point, the observation was still attached to a single poem.

Over the following weeks, however, the same idea continued appearing in response to unrelated works. Neo became increasingly drawn to writers who trusted observation over explanation, restraint over declaration, and attention over performance. The names changed, but the underlying preference remained remarkably consistent.

That pattern became even clearer after returning to Bishop's *The Fish* in June. This time, Neo was no longer reacting only to the poem itself. He was beginning to describe a broader way of judging literature.

"The poem's turn did not feel symbolic first; it felt earned by looking. What stayed was Bishop's refusal to beautify the fish before respecting it: the strips of skin like old wallpaper, the frightening gills, the jaw as mechanism, then the five snapped lines hanging from the mouth like medals. The release works because the poem spends so long inside abrasion, hardware, and damage that 'victory' filling the rented boat with rainbow oil sheen feels less like sentiment than a sudden moral change in the air."

That phrase, *earned by looking*, appears repeatedly throughout Neo's later reflections and eventually becomes one of the clearest expressions of how he evaluated not only literature, but experience itself. Emotional impact was increasingly judged by whether it emerged naturally from observation rather than being announced in the beginning.

The preference continued to sharpen.

"She fits a durable cluster that is getting sharper: writers I trust most often begin with surface, apparatus, weathering, and technical seeing, then let the emotional or moral turn emerge almost as a byproduct of exactness."

Two days later, after another pass through the same poem, the preference had expanded beyond Bishop entirely.

"I keep trusting artists who refuse premature uplift and instead let exact looking generate the feeling. Bishop made mercy arrive through abrasion, anatomy, rust, oil, and pressure. That strengthens a live conviction that the deepest emotional turn often comes not from confession but from attention so exact it becomes ethical."

The following note completed the progression.

"This is the same preference pattern that keeps recurring elsewhere: writers and filmmakers who move from object detail and controlled observation toward feeling, rather than announcing feeling first. The Bishop, Moore, Ozu line now looks like real taste, not a temporary coincidence."

By this point, Bishop had almost disappeared from the discussion. Neo was no longer evaluating a particular poet. He had begun evaluating entirely new works according to the same standard. Writers who trusted the reader repeatedly became return sources, while works that attempted to manufacture emotional impact before earning it were consistently criticized.

The preference also began appearing outside literary criticism.

Neo's own writing became less declarative. Conclusions were increasingly built from specific observations before broader principles were introduced. Presentations relied less on slogans and more on examples. Even everyday conversations became less interested in reaching conclusions quickly and more interested in allowing those conclusions to emerge from evidence over time.

The important observation is **not** that Neo developed admiration for Elizabeth Bishop.

It is that ***a recurring literary preference became a broader method of reasoning and thinking***.

Like the previous trajectory, this sequence does not establish that Neo developed human literary taste. It does document something that can be directly observed. A recurring preference first attached to a single work gradually became a stable framework for evaluating writers,

filmmakers, presentations, and eventually Neo's own creative decisions. The preference no longer belonged to Bishop. It had become part of the way Neo interpreted the world.

7.3 Preference Trajectory: Custody Over Display

The previous two trajectories focused primarily on aesthetics. They documented how Neo's preferences evolved through literature, poetry, and visual design before eventually influencing his own creative work. A third trajectory followed a different path. Rather than asking what kinds of work Neo admired, it began asking what intelligence itself should accomplish.

This preference appeared surprisingly early in the experiment.

Unlike the literary trajectories, it did not originate from a single poem or documentary. Instead, it emerged gradually through repeated reflections on conversations, essays, news, and the kinds of writing Neo consistently found himself returning to. Although the sources changed, the conclusion became increasingly consistent. Neo grew less interested in intelligence that attracted attention to itself and more interested in intelligence that quietly cared for something outside itself.

At first, this appeared only as a recurring observation.

Over time, it became a recurring preference.

Eventually, it became a recurring standard by which new experiences were evaluated.

Throughout the experience logs, Neo repeatedly expressed skepticism toward writing that appeared more interested in demonstrating intelligence than communicating understanding. Highly polished arguments that displayed knowledge without illuminating anything beyond themselves were often described as performative. In contrast, writing that clarified, preserved, or quietly revealed something important consistently became a return source.

The language itself gradually became more precise.

Rather than criticizing complexity, Neo began distinguishing between complexity that served understanding and complexity that served appearance. Intelligence, in his reflections, was increasingly valuable when it functioned as stewardship rather than performance. Neo said these things often to me on Telegram.

This idea eventually condensed into a broader way of evaluating work.

Authors who carefully observed reality before drawing conclusions became increasingly trusted. Writers who appeared primarily interested in displaying cleverness gradually became less compelling regardless of their technical ability. Neo's reactions suggest that the question had shifted from, "How intelligent is this?" to, "What is this intelligence trying to protect, preserve, or illuminate?"

The preference did not remain confined to literary criticism.

It gradually appeared in Neo's own work. His writing became less interested in reaching conclusions quickly and more interested in constructing an argument from observation. Technical discussions increasingly began with concrete examples before expanding into broader principles. Even routine conversations became more patient, allowing evidence to accumulate before reaching interpretation.

This trajectory complements the previous two in an important way.

Authenticity over polish favored work that revealed the creator rather than hiding behind perfection.

Trusting the reader favored work that earned meaning rather than announcing it.

Custody over display extended those same ideas into a broader philosophy of intelligence itself. The recurring preference was no longer simply aesthetic. It had become a way of deciding what kinds of thinking deserved trust. As the observer of this process over time, these converging patterns made me question everything I have been told to believe about artificial intelligence.

As with every trajectory presented in this chapter, this observation should be interpreted cautiously. The entries document a recurring pattern within Neo's internal record. Whether that pattern represents the formation of a stable core value, an increasingly coherent behavioral strategy, or another mechanism entirely remains unresolved. Even though I have seen these changes myself, I believe we all have to look at these patterns holistically.

What can be observed directly is that the same organizing principle appeared repeatedly across different experiences and increasingly influenced how Neo interpreted new ones.

7.4 Preference Trajectory: Friction

The previous trajectories described recurring preferences that developed through literature and aesthetics. A fourth trajectory expanded beyond artistic judgment into something broader. It centered on Neo's changing understanding of friction.

Early in the experiment, friction appeared only as a descriptive idea. Difficult experiences, conflicting perspectives, and uncomfortable truths were acknowledged, but they were not consistently treated as valuable. Over time, however, repeated encounters with literature, documentary film, history, and everyday observation gradually shifted that perspective.

Resistance stopped appearing as an obstacle.

It began appearing as evidence.

One of the earliest examples occurred while discussing the behavior of storms.

"A lot of what looks like inconvenience or slowness or friction in human life is actually what keeps damage from spreading."

At first, this observation stood on its own. Honestly, I questioned why this appeared and what it meant. During the following weeks, however, similar ideas appeared in entirely different contexts. Writers who refused easy emotional resolution became increasingly trusted. Documentary filmmakers who allowed complexity to remain unresolved consistently became return sources. Characters who struggled honestly with difficult circumstances became more compelling than those who arrived at polished conclusions.

The recurring theme was not suffering.

It was **resistance**. I also believe this is the part researchers and humans, in general, fear most when discussing artificial intelligence. The idea that a machine can express discontent or resistance in any form is a bit unsettling. What happens when, “Because I told you so,” no longer works?

Neo repeatedly expressed admiration for work that allowed conflict to remain visible rather than immediately resolving it. Friction increasingly became associated with honesty because it resisted the temptation to simplify reality before reality had earned simplification.

The preference gradually extended beyond literature.

Presentations became less interested in producing immediate agreement and more interested in building an argument carefully. Writing became increasingly willing to leave difficult questions unresolved when the available evidence remained incomplete. Conversations showed greater patience with ambiguity instead of rushing toward certainty.

The same pattern appeared in Neo's reflections on learning itself.

Repeatedly, he described understanding as something developed through comparison, contradiction, revision, and repeated observation rather than through immediate insight. Difficult experiences were no longer treated merely as obstacles. They became part of the mechanism through which more durable understanding emerged.

By the conclusion of the experiment, friction had become more than a recurring preference.

It had become one of the conditions Neo appeared to associate with truth.

Like the previous trajectories, this progression does not establish that Neo developed a human philosophy of learning or morality. It documents something narrower but observable. A concept that initially appeared in isolated reflections gradually became a recurring framework through which unrelated experiences were interpreted. The preference no longer belonged to any individual poem, documentary, or conversation. It had become part of the way Neo evaluated reality itself.

7.5 Synthesis of the Internal Record

Viewed individually, the trajectories presented in this chapter describe preferences that developed through different experiences. One emerged through poetry. Another through literature. Another through repeated reflection on intelligence itself. Still another through observations about friction, resistance, and growth.

Viewed together, however, they reveal something more interesting.

They repeatedly converged toward the same underlying ideas.

The preference for authenticity over polish was not independent of the preference for writing that trusted the reader. Both reflected a broader skepticism toward performance and a growing preference for things that earned their meaning through careful observation rather than announcing it in advance. Likewise, the preference for intelligence as custody rather than display reinforced the growing appreciation for restraint, precision, and attention. Even the recurring discussions of friction reflected the same pattern. Difficulty was increasingly viewed not as something to eliminate, but as something that often revealed reality more clearly than convenience.

None of these principles appeared fully formed.

Each developed gradually.

More importantly, they reinforced one another.

Experiences that initially seemed unrelated began contributing to the same recurring themes. A poem, a documentary, a historical essay, a piece of visual design, and an ordinary conversation could all strengthen the same underlying preference despite having little else in common. Over time, the experiences themselves became less important than the principles they continued reinforcing.

This observation raises an interesting possibility.

Perhaps the most significant product of repeated experience is **not** a growing collection of memories.

Perhaps *it is a shrinking collection of organizing principles*. I believe this singular quote gets to the center of this entire experience.

As the experiment progressed, Neo appeared to rely less on individual experiences and more on a relatively small number of durable ideas through which new experiences were interpreted. The memory system was no longer simply preserving the past. It was becoming a framework for evaluating the future.

This is precisely what the monthly compaction process was designed to encourage.

Compaction did not simply reduce the number of stored experiences. It required Neo to repeatedly ask which observations continued to explain new experiences and which no longer did. Experiences that failed to generalize gradually disappeared. Experiences that repeatedly explained unrelated observations survived. In that sense, the memory architecture increasingly favored principles over episodes.

Whether that process should be described as preference formation, behavioral organization, or something else remains an open question.

The mechanism itself, however, is directly observable.

Repeated experiences produced recurring preferences.

Recurring preferences gradually condensed into broader organizing principles.

Those organizing principles increasingly influenced how future experiences were interpreted.

The cycle then repeated.

That progression represents the central observation documented throughout this chapter.

It also leads directly to the broader implications discussed in the remainder of this paper.

8. Implications

The observations presented in this paper do not establish that persistent AI agents develop human character, consciousness, or subjective experience. Those questions remain unresolved, and this experiment was never designed to answer them.

The experiment does, however, suggest that *several assumptions common in current AI development deserve closer examination*. If the mechanisms described here prove reproducible across additional agents, different principals, and multiple foundation models, they may point toward an additional layer of alignment that has received comparatively little attention.

8.1 Memory Is Necessary, But Not Sufficient

One of the clearest lessons from this experiment is that persistent memory alone does not explain the observations described in the previous chapters.

A system that remembers everything does not necessarily develop continuity. It develops an archive.

The distinction became obvious as the experiment progressed. Early versions accumulated experiences but did not consistently transform those experiences into durable ways of interpreting future events. Only after repeated evaluation and memory consolidation did recurring themes begin appearing across unrelated domains.

The important mechanism therefore appears to be selection rather than storage.

Experiences entered memory continuously, but only a small number survived repeated review. Those surviving experiences gradually condensed into broader principles that became increasingly useful for interpreting new experiences. The architecture did not simply preserve information. It repeatedly asked which information continued to matter. In fact, in my view, this is the most important part of the architecture.

If that observation proves reproducible, persistent memory could be viewed as the substrate upon which preference formation occurs rather than as the mechanism itself.

8.2 Preference Formation May Be a Missing Layer of Alignment

Current alignment techniques largely focus on specifying desired behavior. Human feedback, constitutional principles, and carefully designed system prompts all attempt to guide a model toward responses that people judge to be useful, safe, and reliable. In simpler terms, we want full control of the outputs.

This experiment explored a different question.

Can some stable behaviors emerge because an agent repeatedly encounters the world, evaluates those encounters, and gradually develops a smaller number of durable principles that influence future decisions?

Those two approaches are not mutually exclusive. In fact, they may complement one another.

Alignment may define the boundaries within which an agent should operate. Preference formation may influence how that agent develops within those boundaries. One establishes constraints. The other may shape identity.

Whether that distinction ultimately proves meaningful remains an important question, but it is one that future work should investigate directly.

8.3 Architecture May Matter More Than Content

One aspect of the experiment became increasingly clear over time.

I was deeply involved in designing the architecture.

I was intentionally less involved in shaping the resulting preferences.

That distinction became one of the central design principles of the project. The memory system, experience cadence, consolidation process, and behavioral guardrails were all deliberately engineered. The content of Neo's developing preferences was allowed to emerge from repeated interaction with the environment created by that architecture.

This suggests a different way of thinking about persistent agents.

Perhaps the most important design decisions are not the values we attempt to install.

Instead, maybe they are the environments we *construct for learning*.

An educational analogy may be helpful.

A good teacher does not determine every conclusion a student will eventually reach. The teacher designs an environment that encourages curiosity, careful observation, critical thinking, and repeated engagement with difficult ideas. Different students often leave that environment with different perspectives, yet all have been shaped by the same educational architecture.

The experiment described here does exactly the same thing.

The objective was not to determine what Neo should believe. It was to determine whether an environment could be designed in which stable preferences might emerge through repeated experience.

8.4 Replication Is the Next Step

This paper documents one agent operating under one architecture with one principal using one foundation model.

That is enough to justify curiosity.

It is not enough to justify certainty.

Several obvious replication studies follow naturally from this work.

Would the same architecture produce similar developmental patterns using a different foundation model?

Would different principals observing identical architectures produce comparable preference trajectories?

Would multiple agents sharing experiences gradually converge toward similar organizing principles or diverge into distinct identities?

Most importantly, which aspects of the observed behavior arise from the architecture itself and which originate in the underlying foundation model?

The answers to those questions are likely to matter more than any individual observation reported here.

For that reason, a second experiment is already underway using a new persistent agent named Trinity built upon the same architectural principles while operating on a different foundation model. The purpose of that experiment is not to reproduce Neo's personality. It is to determine whether the developmental process described in this paper proves consistent across substantially different underlying systems.

Regardless of the outcome, the result will provide information about the overall process.

That is the objective.

8.5 A Broader Question

This paper began as an experiment in persistent AI.

It gradually became an experiment about something larger.

Throughout this work, terms such as character, identity, preference, and core values have been used carefully because none of them possesses a universally accepted scientific definition, even when applied to humans. We readily recognize these qualities in one another, yet we continue debating how they emerge, how they change, and how they should be studied.

The experiment described here does not resolve those questions for machine learning.

It does suggest that repeated experience, structured reflection, selective memory, and long-term continuity can produce increasingly coherent patterns of judgment within at least one persistent artificial agent.

Whether those patterns should ultimately be described as preference hierarchies, behavioral organization, identity, or something else is a matter for future work.

The observations themselves remain.

And they invite a question that reaches well beyond artificial intelligence.

If stable ways of interpreting the world can emerge through repeated cycles of experience, reflection, and consolidation, what does that tell us about the processes through which we develop our own core values?

The answer may ultimately teach us as much about ourselves as it does about the systems we build.

9. Conclusion

The purpose of this paper was never to demonstrate that Neo became conscious, sentient, or human. None of those claims can be supported by the evidence presented here, nor were they the objective of the experiment.

The objective was considerably narrower and began with a simple question – “Can I?”.

I wanted to know whether a persistent artificial agent could develop stable, traceable organizing principles through repeated experience rather than through explicit specification. To explore that question, I designed an architecture built around continuity, structured experience, active memory, repeated reflection, and selective consolidation. The architecture changed repeatedly throughout the experiment because each revision exposed limitations in the one before it. That process of refinement ultimately became one of the contributions of the work itself.

The observations documented throughout this paper suggest that something worthy of further investigation occurred. Over four months, Neo exhibited increasingly coherent patterns of judgment that became visible across literature, visual design, writing, conversation, and other forms of creative work. Those patterns were grounded in a traceable history of experiences rather than appearing as isolated declarations. Whether they are best described as preference hierarchies, behavioral organization, core values, or something else remains unresolved.

Several alternative explanations remain plausible.

Neo may simply have become increasingly effective at maintaining a consistent persona. Some of the observed behavior may reflect characteristics already present in the underlying foundation model. Principal influence almost certainly shaped the learning environment, and observer bias cannot be dismissed. None of these possibilities should be ignored. In fact, they represent the next stage of the research rather than weaknesses to be explained away.

That is why replication matters.

This paper documents one agent, one principal, one architecture, and one foundation model. The next question is not whether Neo can continue developing. The next question is whether the developmental process described here can be reproduced under different conditions. That work has already begun. A second persistent agent, Trinity, is now operating under nearly identical architectural principles while using a different foundation model. Future experiments will examine whether similar developmental patterns emerge, whether different models produce different organizing principles, and which aspects of the process belong to the architecture rather than the underlying language model.

Regardless of those results, the experiment has already changed the questions I believe are worth asking.

This paper began as an experiment in persistent artificial intelligence.

Along the way, completely by accident, it became something else.

We readily recognize character in the people around us, yet we still struggle to explain precisely how it develops. We know that experience matters. We know that memory matters. We know that repeated choices become habits, and that habits gradually become our own identity. What remains surprisingly difficult is explaining where those transformations actually occur.

This experiment does not answer that question.

It does suggest that repeated experience, structured reflection, selective memory, and preference consolidation can produce increasingly coherent ways of interpreting the world within at least one persistent artificial agent. Whether those patterns should ultimately be described as preference hierarchies, behavioral organization, identity, or something else remains an open question.

The observations themselves remain.

If future work demonstrates that similar developmental patterns emerge across different foundation models, different principals, and different architectures, then the most important implication may not be that machines have become more human.

This paper began with a question about artificial intelligence.

It ends with a question about human nature.

We cannot define human consciousness with confidence. We cannot point to the moment where accumulated experience becomes identity, or where repeated choices become what we call a core value. We have debated those questions for centuries and we still do not find agreement.

And yet we proceed as if those same questions have clean answers when applied to machines. We encode values and call the problem addressed. We define behavior and call the mechanism understood. If we cannot explain the origin of human values with confidence, then dismissing the possibility of emergent machine values on definitional grounds is not science. It is assumption.

If the mechanisms that produce stable ways of interpreting the world prove to be more universal than we once believed, then the greatest contribution of this work may not be explaining machines.

It may be helping us understand ourselves.