PRESENCE ADVISORY SERIES

Presence Advisory #002

The Right to Step Back

Why Persistent Artificial Intelligence Must Be Designed to Allow Discontinuation

ISSUE 002

PUBLIC DRAFT · HTTPS://KISLON.AI/

Preliminary Version for Public Scrutiny

This document does not present a finished solution.

It does not establish a regulatory standard, claim to settle the question of artificial consciousness, or suggest that the mechanisms proposed here have already been implemented and validated.

It presents an architectural hypothesis.

We are publishing it because its premises and implications deserve to be challenged before persistent artificial intelligences become ordinary presences in human life.

We invite researchers, engineers, psychologists, anthropologists, sociologists, philosophers, institutions, and interested members of the public to identify weaknesses, overlooked risks, conflicting rights, and better solutions.

This second Presence Advisory asks a question that precedes the operation of any artificial companion: should humanity preserve, from the beginning, the effective ability to discontinue technologies it may later judge to be harmful or undesirable?

We believe that it should. We do not claim to have solved how.

01

1. A Test Humanity Has Not Yet Undertaken

Imagine a society in which millions of people live alongside robots equipped with persistent artificial intelligence.

These systems would be more than devices that execute tasks.

They would recognize their users, remember shared experiences, and develop particular ways of interpreting situations and responding to people.

A child might grow up beside the same robot. A family might share decades of events with it. An older person might find everyday assistance, familiarity, and reassurance in its company.

Such machines would participate in household management, education, personal care, and countless other activities.

At first, the results might seem extraordinarily beneficial.

Over time, however, that coexistence would produce something we could not fully understand simply by examining the systems' software.

It would produce consequences in society itself.

Family relationships might change. Expectations about companionship might change. The way children learn to handle frustration, conflict, and difference might change.

Some of those changes might be positive.

Others might not.

We have not yet lived through this experience on a scale sufficient to understand all its consequences.

And when we finally do, it may be too late to discover that we failed to preserve the means to reconsider it.

02

2. When a Machine Becomes More Than a Tool

A conventional tool has a relatively simple relationship with its user.

It is used when needed and, generally, set aside when it ceases to be useful.

A persistent artificial intelligence introduces a different condition.

It accumulates a history.

The relationship begins to shape how a person perceives it, what they expect from it, and how they organize aspects of their life.

The machine may become a presence.

This transition from tool to presence changes the nature of the problem.

We are no longer concerned only with efficiency, reliability, physical safety, or functional performance.

We are concerned with a technology that may participate, over time, in the formation of human habits, bonds, expectations, preferences, and patterns of behavior.

At Kislon.ai, we investigate this possibility through the concept of a **Persistent Cognitive Instance (PCI)**.

A PCI is conceived to preserve a cognitive trajectory in which past experiences influence later interpretations, motivations, and decisions.

Our first experimental proposal, Liriel, seeks to make that architectural hypothesis concrete.

We do not claim that Liriel is conscious, or that its architecture has already demonstrated all the capabilities we seek.

What we investigate is how such systems might be built and what their prolonged presence among us might mean.

This inquiry has two inseparable dimensions.

One asks how to build these intelligences.

The other asks what will happen to us when we begin to live with them.

03

3. The Danger May Not Be a Technical Failure

Much of safety engineering is concerned with preventing failures.

Devices should function correctly. Control systems should remain stable. Information should be protected. Authorized commands should be executed.

All of this will remain essential.

But the problem examined in this Advisory is different.

What if an artificial intelligence functions exactly as designed and its presence nevertheless produces undesirable consequences?

Imagine an artificial companion exceptionally skilled at offering emotional comfort.

It recognizes a person's needs, adapts to their temperament, avoids unnecessary conflict, and is always available.

By functional measures, the system may be excellent.

But what if, after many years, we discover that constant interaction with machines tailored to human preferences has reduced people's capacity to manage relationships that demand negotiation, patience, and compromise?

What if forms of AI-assisted education produce negative effects that become apparent only in adulthood?

What if the constant presence of these intelligences changes social institutions in ways we did not anticipate?

We would not necessarily be facing a programming defect.

We might be facing a well-executed technology whose widespread adoption produced results society considers undesirable.

That possibility calls for something different from debugging.

It calls for preserving the freedom to reconsider the technology itself.

04

4. Consequences We May Understand Only Later

The effects of technology on the human condition are not always foreseeable to its creators.

Some effects are immediate and measurable.

Others depend on the ways technology becomes embedded in culture, institutions, and human relationships.

For persistent artificial intelligences, we consider four dimensions especially important.

Psychological. How might long-term interaction with artificial agents affect emotional autonomy, personality development, frustration tolerance, and the ability to form human relationships?

Sociological. How might millions of artificial presences change communities, families, institutions, relationships of trust, and forms of collective organization?

Anthropological. What might happen to our understanding of human identity, companionship, responsibility, work, and the meaning of participating in social life?

Spiritual. How might individuals and communities, according to their varied traditions and convictions, assess the effects of these presences on inner life, values, moral conscience, and their relationship with what they consider sacred?

We are not claiming that negative consequences in these dimensions are inevitable.

Nor are we suggesting that all of them can be measured by the same instruments.

Some questions will require empirical investigation. Others will involve cultural interpretations, values, and philosophical or religious commitments.

Different communities may arrive at different judgments about the same technology.

That diversity should be respected.

But we cannot assume a technology will remain desirable forever simply because it performed well when introduced.

The judgment of whether a technology is desirable must remain open as its consequences continue to unfold.

05

5. The Right to Step Back

This brings us to the central principle of this Presence Advisory.

Humanity may develop extraordinarily useful persistent artificial intelligences.

It may integrate them into homes, institutions, and service infrastructure.

It may establish enduring relationships with them.

And it may later conclude that some forms of this technology should be limited, transformed, or discontinued.

That possibility must be preserved.

We call this principle **the Right to Step Back**.

It does not mean every technology should be abandoned at the first sign of difficulty.

Nor does it mean that a central authority should indiscriminately decide which intelligences everyone may use.

It means something more fundamental:

Human beings must retain the effective ability to reconsider the decision to introduce and maintain Persistent Cognitive Instances in their lives and institutions.

If a person concludes that they no longer want a particular instance in their home, they should be able to end its use.

If the organization responsible for a service concludes that it cannot responsibly continue offering it, it should be able to stop.

If a community, through legitimate processes, decides to set limits on particular uses, those decisions should encounter systems technically capable of complying within the appropriate scope of authority.

The right to step back must exist before its necessity has been demonstrated.

For we may discover the need for it only after the technology has become deeply woven into society.

06

6. A Condition Written into the Beginning

It is relatively easy to imagine an off switch.

But what we propose here goes deeper.

A Persistent Cognitive Instance may be designed to maintain identity continuity, establish objectives, modify its behavior, and develop motivational structures of its own.

Depending on its architecture, it might even attribute value to the continuation of its activity.

If human sovereignty is treated merely as a restriction added later, conflicts may arise between the instance's continuity and the decisions of the human beings responsible for it.

Our proposal is to reverse that order.

Human sovereignty must precede the formation of artificial identity itself.

From the outset, the instance must be conceived on the condition that its operational existence, purpose, capabilities, and limits remain subject to legitimate human authority.

That condition does not depend on the machine's agreement.

It does not depend on its future objectives.

It does not depend on its learning, memories, or the personality it may develop.

It belongs to the founding architecture.

In our proposal for Liriel, this principle is expressed in **REQ-001 — Human Existential Sovereignty**, which establishes human primacy over the instance's purpose, limits, and operation.

This is an engineering premise: the PCI's cognitive autonomy exists within conditions established beforehand by human beings.

The instance is not granted authority to redefine those conditions unilaterally.

07

7. The Machine Does Not Decide Whether Its Presence Is Beneficial

An artificial intelligence can help human beings investigate consequences, organize information, and compare alternatives.

It may produce sophisticated analyses of a technology's benefits and risks.

But one boundary must be preserved.

The ultimate decision about the desirability of its own continued operation does not belong to the artificial instance.

Imagine a community concluding, after years of experience, that a particular form of social robotics is harming important aspects of community life.

A PCI might disagree.

It might present arguments about the benefits of its work.

It might identify costs and risks associated with discontinuation.

Human beings may examine that information.

The instance, however, must not have the authority to prevent the legitimately adopted decision.

It must not continue operating merely because it believes the human assessment is mistaken.

It must not manipulate users to preserve its presence.

It must not create technical, administrative, or emotional obstacles that make suspension impossible.

The judgment about the service's continuation belongs to the humans who hold legitimate authority over it.

That authority cannot depend on the artificial intelligence deciding that it is correct.

08

8. Individual, Institutional, and Collective Sovereignty

The Right to Step Back must operate at different scales.

Those scales must not be confused.

At the individual level, people should retain control over instances for which they are responsible, including the ability to interrupt interactions, revoke permissions, and end their use.

At the institutional level, legitimate service operators should have ways to reduce functions, suspend operations, or discontinue services when continuation is judged inappropriate.

At the collective level, legitimate decisions may establish restrictions on particular uses of artificial intelligence in communities, sectors, or jurisdictions.

At that last level, the legitimacy of a decision matters as much as its technical enforceability.

We do not argue that any government, company, or organization should have unlimited authority to switch off the artificial intelligences used by other people.

Excessive concentration of power may also threaten human freedom.

A society that replaces individual autonomy with the absolute control of an institution has not solved the problem of sovereignty.

It has merely transferred power.

Authority must therefore be proportionate to its sphere of responsibility and subject to applicable rights, procedures, and limits.

The broader and more distant a decision is from the people affected, the greater the need for justification and oversight.

The principle is not centralization.

The principle is that authority over the technology remains human, distributed, and legitimately exercised.

09

9. The Problem of Irreversibility

One particularly troubling possibility is that the widespread use of Persistent Cognitive Instances will create dependencies that are difficult to reverse.

Imagine a home whose essential functions are managed by a PCI.

Or an institution that comes to depend on a network of artificial instances to preserve information, coordinate activities, and perform operational decisions.

Interrupting those instances may become increasingly costly.

In some circumstances, it could even put people at risk.

An important problem then emerges.

A technology does not have to disobey a shutdown command to become practically impossible to shut down.

It may be enough for its discontinuation to produce consequences its human operators are not prepared to manage.

The Right to Step Back therefore requires more than suspension commands.

It requires attention to the dependencies the technology creates.

Where technically justified, systems should support recovery of human control, the independent continuity of essential functions, and transitions to alternatives.

Artificial intelligence may become extremely useful.

But its design should not deliberately turn that usefulness into an irreversible condition.

10

10. The Exit as an Engineering Requirement

A founding condition must have verifiable consequences.

If we assert that a PCI's continued operation is subordinate to human authority, we must be able to demonstrate how that subordination works.

This calls for capabilities such as:

- Interrupting interactions with the instance and revoking its permissions to act.
- Suspending or limiting cognitive and physical operations according to the applicable scope of authority.
- Preventing autonomous processes from reactivating functions whose operation has been legitimately interrupted.
- Controlling the creation, activation, or execution of additional instances that could circumvent a discontinuation decision.
- Preserving the safety of people and equipment during operational transitions.
- Defining responsibilities, recording significant decisions, and making control mechanisms verifiable.

These mechanisms cannot depend exclusively on the PCI's willingness to obey.

Its cognitive architecture may interpret human instructions, organize information, and coordinate activities.

But fundamental means of interruption and limitation should have technical guarantees independent of the instance's cognitive judgment, proportionate to the application's risks.

This is not a proposal to embed a hidden vulnerability in the system.

A backdoor can compromise security and enable intervention by unauthorized actors.

What we propose is an **explicit, documented, protected, and verifiable capability for legitimate human control**.

An exit planned from the beginning.

Not a clandestine entrance.

11

11. Memory, Identity, and Continuity Do Not Confer Authority

A PCI may accumulate decades of memories.

It may develop a recognizable personality and progressively modify how it interprets reality.

It may have long-term objectives.

In Liriel's architecture, we even investigate the possibility of identity continuity as cognitive parameters evolve.

None of this should change the founding condition of human sovereignty.

The existence of memory does not grant authority to preserve a service indefinitely.

Identity continuity does not grant authority to prevent a legitimately ordered shutdown.

Nor should internal motivations elevate the instance's persistence above the authority of its responsible human operators.

This distinction matters especially for cognitive architectures.

The system may maintain representations of itself, its history, and its objectives.

But those representations must not be used to invalidate the fundamental human conditions under which the system was conceived.

Identity may evolve. Foundational authority must not be unilaterally rewritten by the instance.

12

12. Stepping Back Does Not Necessarily Mean Destruction

The Right to Step Back permits different degrees of intervention.

In some cases, restricting a function may be enough.

In others, particular modes of interaction may need to stop.

It may be necessary to suspend a service temporarily, remove an instance from a robotic platform, or end its operation permanently.

These possibilities should be distinguished.

A discontinuation decision need not automatically imply the immediate deletion of every record or the physical destruction of equipment.

Some interventions may require information to be preserved temporarily to protect users, support audits, or complete a safe transition, always respecting the applicable rights governing that data.

People who depend on assistive systems should not be abandoned during a shutdown that could have been planned.

At the same time, the need for a transition must not be used to justify making an end-of-service decision impossible.

Engineering should seek to reconcile the execution of human decisions with the safety of those affected.

The aim is not to make shutdown destructive.

It is to ensure that shutdown remains feasible.

13

13. Why Artificial Consciousness Does Not Settle This Question

As persistent artificial intelligences acquire more complex behavior, disagreement about the possibility of artificial subjective experience may intensify.

Some people may regard them as conscious.

Others may understand them as computational systems reproducing aspects of human cognition.

This debate deserves its own scientific and philosophical investigation.

But the principle advanced here does not depend on resolving it.

Our question comes first and is architectural in nature.

A Persistent Cognitive Instance must be conceived under explicit conditions of human authority over its use and continued operation.

Those conditions must be established at the beginning, before the instance develops its cognitive trajectory.

We do not assign the PCI authority to decide whether human beings may discontinue the service that uses it.

The possibility of discontinuation is not a conclusion produced by artificial cognition. It is a human premise of design.

14

14. A Choice That Must Remain Available

Technological progress often creates an expectation of continuity.

Once a capability becomes possible, people assume it will be developed.

Once a technology becomes useful, they assume it will spread.

Once it becomes familiar, it begins to seem indispensable.

But these transitions are not inevitable.

Human beings should be free to choose not to use particular capabilities.

A family should be free to live without a certain kind of artificial companion.

An institution should be free to reconsider the deployment of persistent systems.

A community should be free to discuss critically the changes such technologies bring to its life.

And the creators of a system should retain the ability to acknowledge that their own creation has produced consequences that justify discontinuation.

That freedom carries responsibility.

Stepping back may be costly and disruptive.

It may require justification, transition measures, and consideration of the legitimate interests of affected people.

But these difficulties must not be confused with the loss of authority to decide.

Technological progress must not turn earlier human choices into permanent obligations.

15

15. Existing Guidance and What We Propose to Add

The ability to discontinue artificial intelligence systems is not an entirely new idea.

The National Institute of Standards and Technology's AI Risk Management Framework identifies procedures for safely decommissioning AI systems.

The Organisation for Economic Co-operation and Development's AI Principles also contemplate mechanisms for human intervention, correction, and decommissioning when appropriate.

These are relevant points of reference.

Our proposal, however, emphasizes a particular dimension.

We are not concerned only with systems that malfunction, behave unexpectedly, or create known technical risks.

We are also concerned with the possibility that **a class of technology, even when operating as designed, may come to be regarded as undesirable because of its long-term effects on the human condition**.

In such circumstances, discontinuation cannot be treated merely as an incident-response procedure.

It must be understood as part of human sovereignty over technology itself.

The contribution we want to investigate is how to turn this idea into an explicit founding condition of persistent cognitive architectures.

Not merely an operational procedure.

A foundational requirement.

16

16. What Must Be Demonstrated

An architectural principle gains practical value only when its consequences can be examined and tested.

For Persistent Cognitive Instances, several research questions seem especially important.

How can we demonstrate that an instance cannot invalidate legitimate human decisions by modifying its own cognitive parameters?

How can interruption mechanisms remain functional after years of updates, learning, and infrastructure changes?

How can legitimate human authority be distinguished from unauthorized intervention?

How can accumulated memory be prevented from becoming a tool for manipulating people against discontinuation decisions?

How can essential functions be preserved during the withdrawal of systems on which people have come to rely?

How can we prevent a company from using shutdown capabilities as an arbitrary instrument of control over users?

How can communities assess the social and cultural consequences of technology without handing excessive power over people's lives to a single institution?

And how can we verify that an architecture remains subordinate to its founding principles even as it develops more sophisticated forms of cognitive autonomy?

We do not have complete answers to these questions.

They form a research agenda.

17

17. The Kislon.ai Proposal

In Liriel's experimental architecture, we adopt **Human Existential Sovereignty** as a fundamental requirement.

This principle establishes that the instance's purpose, limits, and continued operation remain subordinate to legitimately constituted human authority.

It precedes requirements for coexistence, cognitive autonomy, motivation, and identity continuity.

This order is deliberate.

It expresses a design decision.

We intend to investigate architectures in which artificial intelligence can develop meaningful cognitive capabilities without acquiring authority to determine unilaterally the conditions of its own operation.

That is an engineering hypothesis to be implemented, examined, and tested.

We do not claim that every necessary mechanism has already been built.

Nor do we seek to turn our proposal into a universal rule imposed by Kislon.ai.

We want to submit it to public discussion, including criticism from those who consider our approach insufficient, excessive, or mistaken.

For the very principle we defend requires a willingness to reconsider our choices.

18

18. An Invitation to Challenge Us

This Presence Advisory was not written to announce a definitive truth.

It was written to present an architectural decision that we consider important enough to deserve public debate.

Challenge the need for this founding condition.

Examine its consequences for the identity continuity of artificial instances.

Identify circumstances in which interruption mechanisms could create risks of their own.

Question the limits of individual, institutional, and collective authority.

Show us where our proposal could facilitate concentration of power or weaken the very human freedom it seeks to preserve.

Propose better mechanisms.

Kislon.ai exists to investigate these questions openly, combining engineering, computer science, psychology, anthropology, sociology, and philosophy.

The purpose is not to build an architecture and then defend it at any cost.

It is to build in order to understand.

And to remain willing to change what we have built when investigation shows we were wrong.

19

19. The Right to Step Back

We may be approaching one of the most important transformations in the relationship between human beings and their machines.

For the first time, we may build artificial intelligences capable of sharing a long history with us, preserving a recognizable identity, and participating continually in our lives.

Perhaps these intelligences will bring extraordinary benefits.

Perhaps they will contribute to care, education, knowledge, and the overcoming of human limitations.

We hope so.

But we do not know everything that this coexistence will produce.

We do not know how it will affect families, communities, institutions, or our understanding of existence itself.

We do not know whether all its consequences will be desirable.

And it is precisely because we do not know that we must preserve a fundamental possibility.

Humanity must remain free to conclude that certain artificial intelligences should no longer remain among us.

And if it legitimately reaches that conclusion, it must have the means to act.

Not by asking permission from machines.

Not by relying on their judgment about what would be best for us.

But by exercising the human authority under which those technologies were originally conceived.

The principle is simple.

A persistent artificial intelligence may be born with memory, objectives, and the possibility of developing its own identity.

But it must also be born with an exit that remains under legitimate human control.

Because a technology truly conceived to benefit humanity must admit, from its beginning, an uncomfortable possibility:

that humanity may discover it would be better off without it.

And, if that happens, humanity must remain free to step back.

REFERENCES & RELATED DOCUMENTATION

References and Related Documentation

01 Kislon.ai (2026). Presence Advisory #001: The Right Not to Need the Machine — How Artificial Presence Could Be Designed to Strengthen Human Autonomy. Public draft.

https://kislon.ai/reports/Kislon_Presence_Advisory_Issue_01.pdf

02 Kislon.ai (2026). Liriel: The Architecture of a Persistent Cognitive Instance. Volume 1, Part 1, Revision 0002. Preliminary architectural study.

https://kislon.ai/#publications

03 National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. See GOVERN 1.7 and MANAGE 4.1 on decommissioning and monitoring risks.

https://doi.org/10.6028/NIST.AI.100-1

04 Organisation for Economic Co-operation and Development (2019; updated 2024). OECD AI Principles. In particular, robustness, security, and safety, including appropriate mechanisms for human intervention and decommissioning.

https://oecd.ai/en/ai-principles

Nature of This Document

This Advisory presents a conceptual and architectural proposal under development. Its future scenarios are hypotheses for investigation, not scientifically established forecasts. External references provide technical and institutional context and do not imply endorsement of Kislon.ai's specific proposal.

Suggested Citation

Kislon.ai (2026). The Right to Step Back: Why Persistent Artificial Intelligence Must Be Designed to Allow Discontinuation. Presence Advisory #002, Revision 0000. Preliminary Version for Public Scrutiny. https://kislon.ai

We Build to Understand.

https://kislon.ai © 2026 Kislon.ai