Working Paper
Series E – AI and Institutional Intelligence

WP-E-L02-001

Human-AI Institutional Collaboration Architecture

IDHUS Institute

ISBN:

Made and printed in Spain

Index

Abstract

Artificial intelligence is becoming progressively embedded within institutional processes that support perception, evidence retrieval, classification, modelling, interpretation, scenario generation and decision support. This development extends a much longer history in which institutions have relied upon external cognitive technologies to preserve memory, organise information and expand analytical capacity. Contemporary AI systems, however, increasingly participate in intermediate cognitive transformations that shape how institutional problems are represented before formal decisions occur. Their significance therefore cannot be understood adequately through model performance, automation rates or the continued formal presence of human decision-makers alone.

This Working Paper develops the **Human–AI Institutional Collaboration Architecture (HAICA)** as a middle-range framework for analysing how human and artificial cognitive capabilities become distributed, connected and coordinated within institutions. Building upon the Institutional Cognition Conceptual Architecture and the Institutional Cognitive Architecture Model, HAICA introduces **Hybrid Institutional Cognition** as the distributed performance of institutional cognitive functions through configurations combining human actors, artificial cognitive systems, organisational procedures and information infrastructures. The framework uses the **Human–AI Cognitive Configuration (HACC)** as its principal operational unit, allowing specific arrangements of human and artificial participation to be reconstructed and compared without assuming a permanent division of cognitive labour between people and machines.

HAICA develops Cognitive Function Decomposition and Cognitive Function Allocation as mechanisms for determining where artificial participation becomes institutionally relevant, while Human–AI Cognitive Complementarity examines the conditions under which differentiated capabilities can become mutually corrective rather than merely additive. Hybrid Cognitive Interfaces and Human–AI Cognitive Translation describe the processes through which differently structured representations become mutually usable, and Hybrid Epistemic Traceability preserves the cognitive lineage through which evidence, computational transformations and human interpretation influence institutional judgement. Particular attention is given to the distinction between **Cognitive Authority and Decision Authority**, demonstrating that formal human control can remain intact while effective epistemic influence shifts substantially towards artificial systems.

The framework therefore replaces procedural notions of human presence with **Meaningful Human Judgement**, understood as the condition in which authorised human actors retain sufficient cognitive access, contextual understanding, time, alternative information and institutional authority to understand, challenge and alter artificial contributions when doing so is consequential. HAICA also examines pathways through which initially productive augmentation may deteriorate, including Artificial Cognitive Deference, Augmentation Overload, Institutional Cognitive Deskilling, Cognitive Dependency and Hybrid Institutional Fragility. These vulnerabilities are balanced by Human–AI Cognitive Resilience, Institutional Cognitive Autonomy and Hybrid Cognitive Security, which describe the capacities required to preserve

independent judgement, alternative cognitive pathways and the integrity of hybrid institutional reasoning.

The architecture is explicitly adaptive. Human–AI Feedback Architecture provides evidence concerning the behaviour of hybrid configurations; Hybrid Institutional Learning incorporates that evidence into organisational knowledge and practice; Hybrid Institutional Metacognition makes the distribution of human and artificial cognition itself an object of institutional understanding; and Adaptive Cognitive Allocation allows functions to be reassigned as technological capabilities, institutional expertise and risk conditions change. HAICA therefore does not prescribe a fixed boundary between human and artificial cognition. Its central scientific contribution is to establish **hybrid cognitive architecture as the relevant object through which institutional AI should be evaluated**, shifting research from the question of what artificial intelligence can automate towards the question of how institutions can deliberately organise, observe and govern evolving distributions of human and artificial cognition while preserving epistemic accountability, resilience and legitimate authority.

.

David González
Director
IDHUS Institute

.

1. Introduction — Institutions and the Evolution of Cognitive Technology

Institutions have always depended upon technologies that extend the cognitive capacities of the individuals who participate in them. Writing allowed decisions, obligations and observations to persist beyond individual memory; archives created durable repositories through which organisations could preserve experience across generations; maps transformed spatial relationships into representations that could support administration and planning; statistical systems enabled governments to perceive populations and economic activity at scales inaccessible to direct observation; and computational technologies progressively expanded the speed and complexity with which institutions could process information. The history of governance can therefore be understood partly as a history of changing relationships between human cognition and the external systems through which knowledge is recorded, organised, transmitted and interpreted.

These technologies did more than improve administrative efficiency. They changed what institutions were capable of knowing. A census created forms of population-level visibility unavailable to individual officials. Accounting systems made distributed economic activity legible within common categories. Databases allowed information generated across organisational units to become accessible elsewhere, while computational modelling enabled institutions to represent possible futures rather than merely record past events. Institutional cognition has consequently never been reducible to the unaided reasoning of human actors. It has emerged through relationships among people, procedures, representations and technological infrastructures that collectively determine what an institution can perceive, remember and reason about.

Artificial intelligence enters this historical trajectory as another cognitive technology, but its institutional significance may differ from many earlier forms of information infrastructure. Contemporary AI systems can participate in activities that institutions have traditionally associated more directly with human analytical work: retrieving and synthesising evidence, identifying patterns across large information spaces, classifying complex inputs, generating scenarios, comparing alternatives, producing explanatory narratives and assisting with prediction or modelling. The boundary between technologies that primarily store or transmit cognition and technologies that actively participate in cognitive processes has therefore become increasingly difficult to maintain.

This does not imply that artificial systems reproduce human cognition or that the distinction between human and artificial reasoning has disappeared. Their capabilities, limitations and modes of representation remain different, and contemporary AI systems can produce outputs that are persuasive while being incomplete, unstable or poorly grounded. The institutional transformation lies elsewhere. **Artificial systems can increasingly occupy positions within cognitive workflows that were previously populated primarily by human analysts, advisers, researchers and decision-**

support professionals, meaning that their outputs can influence not merely how quickly institutions process information but how problems themselves become represented.

The resulting challenge cannot be understood adequately through the language of technological adoption alone. When an institution introduces an AI system into policy analysis, regulatory monitoring, scientific assessment or administrative decision support, it is not simply adding another tool to an otherwise unchanged organisation. It may be altering the distribution of cognitive functions across the institution: which actors search for evidence, who constructs initial interpretations, where uncertainty becomes represented, how alternatives are generated and which representations reach decision-makers.

The central question is therefore architectural. **How should institutions organise relationships between human judgement and artificial cognitive capability when both participate in the processes through which institutional knowledge and decisions are produced?** The Human–AI Institutional Collaboration Architecture developed in this Working Paper begins from the proposition that answering this question requires analysing the complete cognitive configuration rather than evaluating either human actors or AI systems in isolation.

2. From Information Technology to Artificial Cognitive Participation

Earlier generations of institutional information technology often operated through relatively clear functional boundaries. Databases stored records, communication systems transmitted information, spreadsheets supported calculation and conventional software executed procedures specified through explicit rules. Although these systems could substantially influence institutional reasoning, the interpretative transition between information and judgement remained comparatively visible. Human actors generally understood that technological systems were processing representations according to predefined procedures whose outputs required institutional interpretation.

Artificial intelligence complicates this arrangement because contemporary systems increasingly produce outputs that resemble intermediate products of human reasoning. A language model can summarise a regulatory dossier, identify themes across consultation responses, compare policy alternatives or draft an analytical interpretation. Predictive systems can estimate risks and rank cases according to inferred patterns. Computer-vision systems can classify observational evidence at scales that would require extensive human labour. These systems do not merely accelerate the movement of existing information; they can participate in constructing representations that subsequently shape institutional judgement.

The distinction is consequential because institutions rarely act directly upon raw evidence. Between observation and decision lie multiple cognitive transformations: classification, aggregation, interpretation, comparison, contextualisation and evaluation. When AI systems participate within these intermediate stages, their assumptions and limitations can become embedded within institutional understanding before a formal

decision occurs. The point at which artificial cognition influences governance may therefore precede the visible moment at which a human decision-maker encounters an AI-generated recommendation.

This creates a form of **Artificial Cognitive Participation**. The concept does not require treating AI systems as autonomous institutional agents or attributing human-like understanding to them. It indicates that artificial systems contribute outputs that become functionally integrated into processes of institutional perception, interpretation, reasoning or learning. Their cognitive significance derives from their position within the architecture rather than from philosophical claims concerning machine intelligence.

Within HAICA, the term *cognitive* is used functionally rather than phenomenologically. **Artificial Cognitive Participation describes the contribution of computational systems to processes that perform institutionally relevant cognitive functions; it does not imply consciousness, subjective understanding or equivalence between artificial and human cognition.** This functional usage allows the architecture to analyse how artificial outputs influence institutional perception, representation and reasoning without requiring claims concerning the internal phenomenology of artificial systems.

Artificial Cognitive Participation can vary considerably in depth. In some settings, an AI system may assist with retrieval while human actors perform nearly all subsequent interpretation. In others, artificial systems may generate initial classifications, forecasts or scenarios that structure the space within which human reasoning occurs. In still others, multiple computational systems may interact before their outputs reach a human reviewer. The same model can therefore possess very different institutional significance depending upon where it is positioned within the cognitive workflow.

This observation shifts attention away from model capability considered independently. A technically powerful system can contribute little to institutional intelligence if poorly integrated, while a comparatively limited system may provide substantial value when positioned within an architecture that complements its capabilities with appropriate human expertise. Conversely, apparently successful automation can weaken institutional cognition if it removes opportunities for contextual interpretation or makes important assumptions less visible.

The institutional significance of AI therefore depends not simply upon what the system can do, but upon what cognitive role it is allowed to perform, what representations it produces and how those representations interact with human judgement elsewhere in the institution.

3. The Limits of the Tool Metaphor

Artificial intelligence is frequently described as a tool, and in many contexts this description remains useful. It emphasises that AI systems are designed, deployed and governed by human institutions and helps resist exaggerated narratives in which technological systems are treated as autonomous replacements for political or administrative authority. Yet the tool metaphor can become analytically insufficient when AI outputs participate deeply in the construction of institutional knowledge.

A conventional tool generally leaves the principal cognitive representation of the task with its user. A calculator changes the speed and reliability of arithmetic without independently deciding which variables should matter to a policy problem. A database expands access to records without necessarily constructing an interpretation of their significance. Contemporary AI systems can occupy a different position because they may select, synthesise, classify or generate representations before human actors encounter the underlying evidence.

When an AI system summarises thousands of documents, the human reader may never examine most of the original material. When it ranks regulatory cases by estimated risk, institutional attention becomes partially organised through its classification. When it generates policy scenarios, the possibilities considered by decision-makers may be shaped by what the system included or omitted. In such cases, the system remains a tool in the broad sense of being an artefact used by an institution, yet analytically it also functions as a **mediator of cognition**.

The distinction matters because tool-centred governance tends to focus upon questions concerning system reliability, accuracy, security and appropriate use. These questions remain essential, but they do not fully capture how AI can redistribute cognition. An accurate system may still narrow institutional reasoning if its representation becomes dominant. A transparent system may still generate excessive cognitive deference if human actors lack the time or expertise to challenge it. A formally advisory system may exercise substantial effective influence if its outputs structure every subsequent stage of analysis.

HAICA therefore does not abandon the tool metaphor so much as place it within a broader architectural account. Artificial systems remain technologies deployed within institutions, but their effects must be analysed according to the cognitive relationships they create. The relevant question is not whether AI should be considered a tool or an agent in abstract philosophical terms, but **whether its position within a particular institutional process gives it a functionally significant role in producing the representations upon which institutional cognition depends**.

This architectural perspective also avoids attributing capabilities to AI systems that they do not possess. Cognitive participation is a functional description of institutional process, not a claim about subjective understanding. An artificial system can materially influence institutional reasoning even if its internal operations differ fundamentally from human cognition. The scientific object is the institution and the architecture through which cognition becomes organised.

4. From Distributed Institutional Cognition to Hybrid Institutional Cognition

ICAM established that institutional cognition is already distributed across heterogeneous components. Human actors participate through expertise, judgement and social interaction; organisational procedures structure the movement of information; archives and databases provide memory; models support representation; and computational infrastructures expand processing capacity. No individual actor contains the complete cognitive process through which the institution perceives and responds to its environment.

The introduction of artificial cognitive participation therefore does not create distributed cognition from nothing. It changes the composition of an architecture that was already distributed. What becomes distinctive is the increasing presence of computational systems capable of contributing to cognitive functions that previously required substantially greater human interpretation.

This transition produces what HAICA defines as **Hybrid Institutional Cognition**: the distributed performance of institutional cognitive functions through configurations combining human actors, artificial cognitive systems, organisational procedures and information infrastructures. The term *hybrid* refers not simply to the coexistence of humans and AI within the same organisation but to their functional participation within connected cognitive processes.

An institution employing AI for peripheral administrative automation may therefore contain artificial intelligence without exhibiting substantial Hybrid Institutional Cognition in the sense relevant to HAICA. By contrast, an institution in which artificial systems participate in evidence synthesis, risk assessment, scenario generation and policy evaluation may possess deeply hybrid cognitive workflows even if humans retain all formal decision authority.

The degree of hybridity is consequently architectural rather than technological. It depends upon where artificial systems enter cognitive flows, which transformations they perform, how their outputs influence subsequent interpretation and whether humans retain meaningful capacities to contextualise, contest and revise those outputs.

Hybrid Institutional Cognition also introduces new forms of interdependence. Human actors may become increasingly reliant upon artificial systems for navigating information spaces too large to inspect independently, while AI outputs remain dependent upon human interpretation for institutional significance. The resulting architecture cannot be evaluated by asking which component is more capable in general. Different components contribute different forms of cognition under different conditions.

This leads to a fundamental shift in the unit of analysis. **The relevant object is no longer the performance of the human or the AI system considered separately, but the performance of the hybrid cognitive configuration through which**

institutional cognition is actually produced. HAICA is designed to make that configuration visible.

5. Human Cognitive Contributions within Hybrid Institutions

A theory of hybrid cognition must avoid defining human contribution merely as the residual set of capabilities that artificial systems have not yet acquired. Such an approach would make the architecture dependent upon rapidly changing technological boundaries and would encourage an unstable division between supposedly human and artificial tasks. HAICA instead examines the functions through which human actors currently contribute to institutional cognition and the institutional conditions under which those contributions remain significant.

Human cognition is deeply embedded within context. Institutional actors interpret information through knowledge of organisational history, legal constraints, political relationships, professional practice and social meaning that may not be fully represented within formal datasets. This contextual knowledge frequently determines whether an analytically plausible recommendation is institutionally feasible or whether a statistically significant pattern has practical relevance.

Human actors also participate in normative judgement. Governance decisions require evaluations concerning fairness, proportionality, acceptable risk, competing public values and legitimate trade-offs. Evidence can inform these judgements without determining them. The distinction is important because computational systems can contribute representations of possible consequences while the authority to determine which consequences should be prioritised remains embedded within political and institutional arrangements.

Experience provides another form of contribution. Experienced officials often possess tacit knowledge concerning how procedures operate in practice, which actors must be consulted, where implementation is likely to encounter resistance and which formal indicators fail to capture important realities. Such knowledge can be difficult to codify precisely because it has developed through repeated participation in institutional processes.

Human actors also remain central to responsibility. Institutions can delegate analytical functions to technological systems, but accountability for public decisions cannot simply be transferred to a model. Contemporary governance therefore continues to require identifiable human and institutional actors capable of explaining, defending and revising decisions.

These contributions should not be interpreted as permanent human monopolies. Artificial systems may increasingly support contextual analysis, normative deliberation or institutional memory in ways that alter current divisions of cognitive labour. HAICA should therefore avoid ontological claims about what machines can never do. Its concern is more practical: **which cognitive functions currently require human participation for**

institutional reasoning to remain contextually grounded, normatively legitimate and accountable, and how should those functions interact with artificial capabilities as technology evolves?

6. Artificial Cognitive Contributions within Hybrid Institutions

Artificial systems contribute a different but increasingly broad range of capabilities. Their principal institutional advantage often lies in the ability to operate across informational spaces whose scale, speed or complexity exceeds what individual human actors can process directly. AI systems can search large repositories, identify statistical patterns, compare extensive collections of documents, monitor continuous streams of information and generate multiple analytical representations rapidly.

These capabilities can expand institutional perception. Automated monitoring may detect anomalies that would otherwise remain unnoticed, while language technologies can make dispersed textual evidence more accessible. Pattern-recognition systems can identify relationships across datasets that human analysts would struggle to inspect exhaustively. Generative systems can assist in exploring alternative scenarios or summarising competing interpretations.

Artificial systems can also increase consistency in tasks where human performance is affected by fatigue, workload or uneven application of complex criteria. Yet consistency should not automatically be equated with correctness. A system can reproduce the same mistaken assumption reliably across thousands of cases. The value of artificial cognition therefore depends upon the quality of its representations and the architecture through which those representations are evaluated.

Speed creates similar ambiguity. Rapid analysis can improve responsiveness, particularly where governance systems must process changing evidence continuously. Yet faster cognition can compress the time available for human interpretation and encourage institutions to treat computational outputs as settled conclusions rather than provisional contributions. Technological acceleration can therefore improve or degrade institutional cognition depending upon surrounding processes.

Scale also introduces dependence. As AI systems make previously unmanageable information spaces cognitively accessible, institutions may gradually lose the practical ability to operate without them. The benefit of expanded cognitive reach can therefore coexist with increasing infrastructural and epistemic vulnerability.

Artificial cognitive contribution should consequently be assessed functionally rather than celebrated as capability expansion in itself. **AI becomes institutionally valuable when its capacity to process, compare, detect, model or generate information is positioned within an architecture capable of interpreting its outputs,**

identifying its limitations and connecting its strengths with forms of cognition located elsewhere in the institution.

7. Cognitive Function Decomposition

Before institutions can determine how human and artificial capabilities should interact, they must first understand the cognitive processes embedded within the tasks they are attempting to augment. Administrative and policy activities are often described through broad categories such as analysis, assessment, planning or decision-making, but each of these contains multiple cognitive functions whose requirements differ substantially.

A policy assessment, for example, may involve retrieving evidence, evaluating source credibility, identifying relevant precedents, modelling possible consequences, comparing alternatives, interpreting uncertainty, considering legal constraints, evaluating normative trade-offs and communicating conclusions. Treating the complete activity as a single task obscures the fact that different components may be suited to different forms of human or artificial cognition.

HAICA therefore introduces **Cognitive Function Decomposition** as the systematic identification of the distinct cognitive operations contributing to an institutional task before decisions are made concerning automation or augmentation. The purpose is not to fragment institutional reasoning mechanically but to make its architecture sufficiently visible for deliberate design.

Function decomposition can reveal that activities appearing suitable for automation contain embedded judgement. A classification task may require interpretation of ambiguous categories. Evidence retrieval may depend upon assumptions concerning relevance. Scenario generation may implicitly privilege particular causal models. Conversely, tasks regarded as requiring extensive human expertise may contain repetitive components that artificial systems can support effectively without displacing the judgement surrounding them.

Decomposition also allows institutions to identify dependencies among functions. The quality of a later judgement may depend upon how evidence was selected earlier in the process. If artificial systems perform upstream functions, human decision-makers may encounter outputs whose representational boundaries were established before they entered the workflow. Understanding these dependencies is essential for preserving meaningful oversight.

The analytical sequence therefore begins not with the capabilities of available AI systems but with the cognitive structure of institutional work. **Institutions should first ask what cognition the task requires and only then determine how human and artificial capabilities might contribute to it.** This reversal is foundational to HAICA because it prevents technological capability from becoming the organising principle of institutional cognition.

8. Cognitive Function Allocation

Once the cognitive functions embedded within an institutional process have been identified, the next question concerns their allocation. Some functions may remain primarily human, others may be performed substantially through artificial systems, while many will require hybrid configurations in which outputs move recursively between human and artificial contributors. HAICA describes this process as **Cognitive Function Allocation**.

Allocation should not be understood as a permanent classification of tasks into human and machine categories. The appropriate configuration depends upon the characteristics of the function, the reliability of available systems, the quality of evidence, institutional context and the consequences of error. As these conditions change, allocation may need to change with them.

Several dimensions become relevant. Functions involving high informational volume and relatively stable analytical criteria may benefit substantially from artificial processing. Functions involving ambiguous context, contested values or significant normative consequences may require stronger human judgement. High-stakes and irreversible decisions may justify greater redundancy and independent review than low-stakes, reversible processes. Poor data quality may reduce the value of computational analysis even where models are technically capable.

Explainability and traceability also influence allocation. If an institution cannot reconstruct sufficiently how an artificial output was produced or which evidence shaped it, assigning that system a central role within high-stakes reasoning may weaken accountability. Conversely, highly traceable computational analysis can sometimes provide more transparent support than informal human intuition whose reasoning remains undocumented.

Allocation therefore cannot be derived from technical capability alone. A system being capable of performing a function does not establish that it should occupy that position within the institutional architecture. The relevant criterion is whether the resulting configuration improves the quality, resilience, interpretability and legitimacy of institutional cognition.

Cognitive Function Allocation is consequently an architectural decision concerning where different forms of cognition should be located within a hybrid institutional process, not a technological decision concerning which human activities can be automated. This distinction provides the foundation for analysing Human–AI Cognitive Complementarity.

9. Human–AI Cognitive Complementarity

The promise of human–AI collaboration frequently rests upon the intuition that humans and artificial systems possess different strengths and can therefore compensate for one another's limitations. Yet complementarity does not arise automatically from placing two different cognitive systems within the same workflow. Their interaction can generate redundancy, confusion, excessive deference or conflict as easily as improved reasoning. HAICA therefore treats **Human–AI Cognitive Complementarity** as an architectural outcome that must be demonstrated rather than assumed.

Complementarity occurs when differentiated human and artificial capabilities are organised so that the contribution of one component improves the capacity of the wider configuration to compensate for limitations elsewhere. An artificial system may identify patterns across large evidence spaces while human actors evaluate contextual significance. AI-generated scenarios may broaden the alternatives considered by analysts, while institutional expertise determines whether those scenarios are legally or politically plausible. Human judgement may identify ambiguities that require additional computational analysis, creating recursive interaction rather than a linear sequence from machine output to human approval.

The quality of complementarity depends partly upon independence. If humans merely reproduce the assumptions embedded within artificial systems, apparent agreement provides little additional cognitive value. Similarly, if AI systems are trained or configured primarily to reproduce existing institutional judgements, they may reinforce established biases rather than provide alternative analytical perspectives. Complementarity is strongest where different cognitive contributions retain enough differentiation to expose limitations in one another.

Disagreement can therefore be valuable. A human analyst reaching a different conclusion from an artificial system should not automatically be interpreted as a failure requiring one side to be corrected. Divergence can reveal uncertainty, missing context, model limitations or human assumptions requiring further examination. Hybrid cognition becomes more robust when disagreement triggers inquiry rather than automatic convergence.

This suggests that the objective of HAICA is not to optimise agreement between humans and AI. It is to design configurations in which **differences between human and artificial cognition become cognitively productive**. Complementarity is achieved when the architecture uses those differences to expand evidence, challenge assumptions, improve interpretation and strengthen institutional judgement.

The relevant performance question therefore changes substantially. Rather than asking whether AI improves human productivity or whether humans improve model accuracy, HAICA asks whether **the complete hybrid cognitive configuration produces institutional reasoning that is more capable, traceable, resilient and contextually appropriate than plausible alternative configurations**.

10. Human–AI Cognitive Configurations

Cognitive functions are rarely performed through permanent one-to-one relationships between a human actor and an AI system. Institutional processes involve multiple participants, databases, models, procedures and review mechanisms whose composition changes according to the problem being addressed. HAICA therefore requires a unit of analysis capable of representing these task-specific arrangements.

A **Human–AI Cognitive Configuration (HACC)** is a task-specific arrangement of human actors, artificial cognitive systems, information resources, organisational procedures and cognitive interfaces through which one or more institutional cognitive functions are performed. The configuration may be temporary or recurrent, simple or highly distributed, and its structure can change as the institutional task moves through different stages.

HAICA and HACC therefore refer to different analytical levels. **HAICA designates the general architecture through which human and artificial cognition can be analysed and organised institutionally, whereas a HACC designates a particular realised or proposed configuration through which that architecture becomes instantiated around a specific cognitive task.** A single institution may consequently contain multiple HACCs whose structures, degrees of artificial participation and distributions of cognitive authority differ substantially, while HAICA provides the common analytical framework through which those configurations can be reconstructed and compared.

A regulatory investigation, for example, may begin with automated monitoring, proceed through AI-supported evidence retrieval, involve human interpretation and legal assessment, return to computational modelling for scenario comparison and conclude through a formally authorised human decision. The relevant cognitive system is not any individual component but the configuration through which these contributions become connected.

HACCs make architectural comparison possible. Institutions can compare human-only processes with AI-assisted configurations, parallel human–AI analysis, AI-first workflows with independent human review or other arrangements. The purpose is not to identify a universally optimal configuration but to determine which structures perform effectively under different conditions.

The concept also makes hidden dependencies visible. A process described formally as human-led may depend almost entirely upon AI-generated evidence selection, while an apparently automated process may contain substantial human judgement embedded upstream in system configuration or data preparation. Mapping the complete configuration therefore provides a more accurate representation of where cognition and influence actually reside.

Human–AI Cognitive Configurations are dynamic. As models improve, institutional expertise changes or new evidence becomes available, functions may migrate among components. A configuration that is appropriate under one technological

regime may become inefficient or unsafe under another. HAICA must therefore treat configuration design as revisable rather than fixed.

The HACC provides the operational unit through which Hybrid Institutional Cognition can be observed, compared and eventually evaluated. It shifts analysis away from abstract debates concerning human versus artificial intelligence and towards the concrete architectures through which institutions combine heterogeneous cognitive capabilities in practice.

11. Hybrid Cognitive Interfaces

Human and artificial cognitive capabilities do not become complementary merely because they participate in the same institutional workflow. Their contributions must cross interfaces through which evidence, assumptions, interpretations and uncertainty can move in forms that remain usable by other components of the configuration. HAICA describes these relationships as **Hybrid Cognitive Interfaces**.

A Hybrid Cognitive Interface is not reducible to the visual interface through which a person interacts with an AI application. User-interface design can substantially affect cognition, but the concept is broader. Hybrid Cognitive Interfaces include prompts, structured inputs, data pipelines, model outputs, confidence representations, review procedures, documentation standards, escalation mechanisms and organisational protocols through which human and artificial contributions become connected.

The quality of these interfaces determines what becomes visible to each participant. An AI system may receive a formally structured representation of a policy problem while remaining disconnected from contextual information known to experienced officials. Conversely, human actors may receive a concise recommendation without access to the evidence, assumptions or alternative outputs through which it was generated. In both cases, information has technically crossed the interface while important cognitive relationships have not.

Interface design can therefore shape institutional reasoning before any explicit judgement occurs. A system presenting a single recommended option may encourage convergence, whereas one presenting several plausible interpretations together with uncertainty may support comparative reasoning. An interface that hides provenance may increase processing speed while weakening epistemic evaluation. Similarly, an interface requiring human actors to justify disagreement with an artificial recommendation can create a different cognitive relationship from one in which AI output must itself satisfy evidential requirements before influencing the workflow.

Hybrid Cognitive Interfaces also operate between artificial systems. A retrieval model may provide evidence to a generative model whose synthesis subsequently reaches a human analyst. If transformations occurring between these components remain invisible, the apparent human–AI interface may conceal a deeper chain of computational mediation.

The interface is therefore part of the cognitive architecture rather than a neutral channel through which cognition passes. It determines which aspects of human and artificial reasoning become mutually accessible and consequently influences whether complementarity, contestation and meaningful judgement remain possible.

12. Human–AI Cognitive Translation

Hybrid cognition brings together systems that can represent problems in substantially different ways. Human institutional reasoning frequently relies upon contextual knowledge, implicit relationships, professional concepts and normative distinctions whose meaning is embedded within organisational practice. Artificial systems operate through computational representations whose relationship to those forms of understanding may be indirect. Effective collaboration therefore requires more than information exchange; it requires **Human–AI Cognitive Translation**.

Human–AI Cognitive Translation refers to the processes through which representations generated by human and artificial contributors become sufficiently interpretable across cognitive boundaries to participate within shared institutional reasoning. Translation can occur in both directions. Human problems must be represented in forms that artificial systems can process, while artificial outputs must be transformed into forms that human actors can understand, interrogate and situate within institutional context.

The first direction is often underestimated. Before an AI system produces an output, institutional reality has already undergone representation. Objectives are formulated, categories selected, data assembled and constraints encoded. These choices determine which aspects of the problem become computationally visible. A complex governance question can therefore be narrowed substantially before artificial analysis begins, not because the system necessarily produces an incorrect result but because the representation supplied to it excludes relationships that institutional actors considered insufficiently explicit.

Translation in the opposite direction creates different risks. Artificial outputs may contain probabilistic relationships, model-generated classifications or synthetic explanations whose apparent linguistic accessibility exceeds their underlying interpretability. A fluent narrative can create the impression that the system's representation corresponds directly to institutional concepts when important differences remain concealed. Linguistic intelligibility should therefore not be confused with cognitive equivalence.

Translation is especially important where uncertainty is concerned. Computational uncertainty may be expressed through confidence intervals, probability distributions or alternative model outputs, while institutional actors may interpret uncertainty through professional judgement, political risk or incomplete contextual knowledge. Hybrid reasoning requires mechanisms capable of relating these different forms without collapsing them into a single measure.

Human–AI Cognitive Translation therefore operates as a bidirectional interpretative layer within HACC. **Its purpose is not to make human and artificial cognition identical, but to preserve enough of the meaning, uncertainty and assumptions contained within each representation for their differences to become cognitively usable rather than silently erased.**

13. Hybrid Epistemic Traceability

As artificial systems participate more deeply in institutional cognition, the provenance of institutional knowledge becomes increasingly complex. A policy assessment may incorporate information retrieved by one model, classifications generated by another, a synthesis produced by a language model, contextual modifications introduced by analysts and subsequent revisions made during administrative review. By the time a recommendation reaches an authorised decision-maker, its cognitive ancestry may be difficult to reconstruct.

HAICA describes the capacity to preserve this ancestry as **Hybrid Epistemic Traceability**. The concept refers to the ability of an institution to reconstruct how human and artificial contributions participated in the production, transformation and interpretation of knowledge that influenced institutional reasoning.

Traceability begins with evidence provenance but extends beyond it. Institutions need to know not only which sources were used but how they entered the cognitive process, which computational systems transformed them, which model versions were involved, which assumptions or prompts shaped outputs, where human interpretation occurred and which representations ultimately influenced a decision. The objective is not necessarily to record every computational operation but to preserve the relationships required for meaningful institutional reconstruction.

This becomes especially important when errors are discovered. If an institutional conclusion proves inadequate, learning depends upon identifying where the cognitive process failed. The problem may have originated in poor evidence, model limitations, inappropriate function allocation, weak translation, human misinterpretation or the interaction among several components. Without traceability, the institution may know that the final judgement was wrong while remaining unable to determine how its architecture should change.

Traceability also supports accountability. A nominally human decision can depend heavily upon artificial representations generated upstream. Recording only the final authorised actor therefore provides an incomplete account of cognitive influence. Conversely, attributing a decision simply to an AI system can obscure the human choices involved in system selection, configuration, interpretation and deployment.

Hybrid Epistemic Traceability should nevertheless remain proportionate. Excessive documentation can create administrative burden and generate repositories too complex to support practical learning. The relevant level of traceability depends upon stakes, uncertainty and reversibility.

A hybrid institution becomes epistemically legible when it can reconstruct not merely who formally made a decision, but how the representations supporting that decision were produced across the complete Human–AI Cognitive Configuration.

14. Cognitive Authority and Decision Authority

The introduction of artificial systems creates a distinction between two forms of influence that conventional institutional structures often treat as if they were aligned. **Decision Authority** refers to the legitimate institutional power to authorise, reject or modify a decision. **Cognitive Authority**, by contrast, refers to the capacity to shape the representations, interpretations and alternatives through which that decision becomes understood.

Within traditional institutions, these forms of authority have never been perfectly identical. Experts, analysts and advisers can exercise substantial cognitive influence without possessing formal decision authority. Artificial systems intensify the distinction because they can construct analytical representations at scales that authorised decision-makers cannot independently reproduce.

A decision may therefore remain formally human while becoming substantially dependent upon artificial cognition. If an AI system selects relevant evidence, frames risk, generates alternatives and recommends a preferred option, the human actor making the final decision may retain legal authority while exercising judgement within a cognitive environment largely structured upstream.

This does not mean that artificial systems themselves possess legitimate authority. The analytical point is different: **effective cognitive influence can migrate without any corresponding change in formal institutional authority**. An organisation can therefore appear constitutionally unchanged while its internal distribution of cognition has been substantially transformed.

The distinction becomes particularly important when human actors lack practical alternatives to artificial representations. If the underlying evidence is too extensive to inspect, the model too complex to evaluate or institutional time too limited to conduct independent analysis, formal discretion may coexist with strong cognitive dependence. The decision-maker can theoretically reject the system while lacking the epistemic resources required to do so meaningfully.

HAICA therefore treats the relationship between Cognitive Authority and Decision Authority as an architectural variable. Some configurations may deliberately separate them, allowing artificial systems to contribute significant analytical influence while preserving robust independent human evaluation. Others may unintentionally allow cognitive authority to concentrate within systems whose outputs become effectively determinative.

Preserving human decision authority is therefore insufficient if institutions do not also preserve the cognitive conditions under which that authority can be exercised as judgement rather than procedural confirmation.

15. Meaningful Human Judgement

The principle that humans should remain "in the loop" has become common within discussions of automated decision systems, but physical or procedural presence alone provides little information about the quality of human participation. A person may formally review an artificial recommendation while lacking sufficient evidence, expertise, time or institutional freedom to challenge it. HAICA therefore replaces a purely positional understanding of human oversight with the concept of **Meaningful Human Judgement**.

Meaningful Human Judgement exists when human actors possess sufficient cognitive access, contextual understanding, institutional authority and practical capacity to evaluate artificial contributions before those contributions become consequential within institutional action. The concept concerns the substantive conditions of judgement rather than the mere existence of a human checkpoint.

Cognitive access requires that relevant evidence, uncertainty and assumptions remain sufficiently visible. A reviewer cannot evaluate a recommendation meaningfully if the system provides only an opaque conclusion. Contextual understanding requires knowledge of the institutional and social environment within which the output will operate. Authority requires that disagreement can genuinely alter the process. Practical capacity requires adequate time, expertise and alternative information through which evaluation can occur.

These conditions interact. A highly experienced official may possess contextual knowledge but lack access to model provenance. Another may understand the technical system but lack authority to override its recommendation. A formally empowered decision-maker may face hundreds of AI-generated cases within a timeframe that makes substantive review impossible. In each situation, human presence exists while meaningful judgement is weakened.

Meaningful Human Judgement should not imply that humans must independently reproduce every artificial analysis. Such a requirement would eliminate many advantages of computational augmentation. The objective is instead to maintain sufficient epistemic capacity to understand what kind of claim the system is making, recognise relevant limitations, identify conditions requiring further scrutiny and exercise genuine discretion where institutional responsibility demands it.

The appropriate degree of judgement can vary according to stakes and reversibility. Low-risk administrative assistance may require limited review, while decisions affecting rights, public safety or irreversible policy commitments may require substantially stronger human engagement.

The central question is therefore not whether a human remains in the loop, but whether the architecture preserves a human capable of understanding, questioning and altering the cognitive trajectory of the process when doing so matters.

16. Artificial Cognitive Deference

Even where institutions preserve formal human authority and provide access to artificial outputs, patterns of interaction can gradually produce excessive reliance upon computational recommendations. Human actors may come to regard artificial systems as more objective, comprehensive or analytically sophisticated than their own judgement, particularly when models operate across evidence spaces that individuals cannot inspect independently. HAICA describes this condition as **Artificial Cognitive Deference**.

Artificial Cognitive Deference is not identical to appropriate epistemic reliance. Institutions routinely depend upon specialised expertise they cannot reproduce internally, and artificial systems can provide valuable analytical support precisely because they extend beyond individual human capacities. Deference becomes problematic when the perceived authority of artificial outputs reduces the willingness or ability of human actors to interrogate them under conditions where scrutiny remains necessary.

Several mechanisms can encourage such deference. Repeated system accuracy can create justified confidence that gradually becomes generalised beyond the contexts in which performance has been demonstrated. Quantitative or technically complex outputs may acquire an aura of objectivity. Organisational incentives may make accepting an automated recommendation easier to defend than departing from it. Workload pressures can transform AI assistance into the default interpretation simply because independent analysis requires additional time.

Deference can also become institutional rather than individual. Procedures may progressively reorganise themselves around artificial outputs, making deviation administratively unusual. Training may teach staff how to operate systems without preserving equivalent capacity to challenge them. Over time, what began as decision support can become the cognitive baseline from which alternative interpretations must justify themselves.

The danger is not merely that artificial systems can be wrong. Human experts are also fallible. The deeper problem is loss of epistemic plurality. If human judgement ceases to function as an independent source of interpretation, the hybrid configuration loses part of the complementarity that justified augmentation in the first place.

Artificial Cognitive Deference should therefore be examined relationally. **The relevant question is not how much humans trust AI, but whether patterns of trust preserve sufficient cognitive independence for disagreement, verification and alternative interpretation to remain institutionally viable.**

17. Augmentation Overload

Artificial intelligence is frequently introduced with the expectation that it will reduce human cognitive burden. By searching larger information spaces, generating summaries and producing analytical recommendations, AI can indeed allow institutional actors to engage with evidence that would otherwise exceed available time and attention. Yet augmentation can also generate new forms of overload.

AI systems can produce analyses, alerts, scenarios, classifications and recommendations at speeds far exceeding human interpretative capacity. An institution may therefore become capable of generating more cognition than its human participants can meaningfully evaluate. HAICA describes this condition as **Augmentation Overload**.

The problem is not simply information volume. Artificial outputs often require interpretation, comparison and verification. Ten generated scenarios can expand the space of analysis but also require institutional actors to understand their assumptions and implications. Continuous anomaly detection can improve sensing while overwhelming officials with alerts whose significance must be assessed. Automated document synthesis can increase access to evidence while creating a proliferation of secondary representations whose relationship to primary sources becomes increasingly difficult to maintain.

Overload can encourage cognitive shortcuts. Human actors may rely increasingly upon rankings, summaries or recommended options because examining the complete artificial output becomes impractical. Paradoxically, systems introduced to augment judgement can therefore increase Artificial Cognitive Deference by creating a volume of analytical material that only further automation can manage.

Institutional processes can also accelerate in response to technological capability. Once analysis can be generated rapidly, expectations concerning decision speed may increase, reducing the time previously available for reflection. The result can be an institution processing more information while devoting less interpretative attention to each decision.

HAICA therefore distinguishes cognitive capacity from cognitive throughput. Increasing the amount of analysis produced does not necessarily increase the institution's capacity to understand its environment. **Augmentation becomes cognitively valuable only when the rate and form of artificial contribution remain compatible with the interpretative capacity of the wider Human–AI Cognitive Configuration.**

18. Institutional Cognitive Deskilling

Hybrid cognition changes not only the immediate distribution of cognitive functions but potentially the capabilities of the human participants themselves. When artificial systems repeatedly perform activities that previously required human practice, institutional actors may gradually lose the expertise necessary to perform, evaluate or reconstruct those activities independently. HAICA describes this process as **Institutional Cognitive Deskilling**.

Deskilling should not be treated as an inevitable consequence of automation. Institutions have always redistributed tasks in response to technological change, and abandoning obsolete forms of manual work can free human capacity for more valuable functions. The relevant concern arises when the skills being displaced remain necessary for evaluating artificial outputs, responding to system failure or preserving institutional understanding.

Consider an analytical process in which AI systems increasingly perform evidence retrieval and synthesis. Human officials may initially possess extensive knowledge of primary sources and use AI merely to accelerate familiar work. Over time, new staff may enter an institution in which direct engagement with those sources is uncommon. The ability to recognise omissions or unusual interpretations can then decline even if the technological system itself continues performing effectively.

Deskilling can therefore transform the relationship between reliance and dependency. At first, institutions use artificial systems because doing so is efficient; later, they may use them because alternative cognitive capacity no longer exists. The architecture has shifted from augmentation towards structural dependence without any deliberate decision to make that transition.

The problem becomes particularly significant when technological systems change. A model can be withdrawn, a provider can fail, regulatory conditions can shift or performance can deteriorate in unfamiliar environments. Institutions lacking residual human expertise may then be unable to evaluate replacement systems or maintain essential functions during disruption.

Preserving every historical skill would be inefficient and unrealistic. HAICA instead requires attention to **critical cognitive capabilities**: forms of human expertise necessary to maintain judgement, verification, resilience and institutional autonomy even when routine execution is substantially automated.

Institutional Cognitive Deskilling is therefore a longitudinal property of hybrid architecture. **A configuration that appears highly complementary at the moment of deployment may gradually become less resilient if repeated delegation erodes the human capacities upon which meaningful evaluation originally depended.**

19. Cognitive Dependency and Hybrid Institutional Fragility

Artificial Cognitive Deference and Institutional Cognitive Deskilling can converge into a deeper architectural condition: **Cognitive Dependency**. An institution becomes cognitively dependent upon an artificial system or infrastructure when essential cognitive functions can no longer be performed, meaningfully evaluated or readily substituted without continued access to that system.

Dependency can arise even when technological performance remains excellent. Indeed, reliable systems may encourage deeper integration because institutions have little immediate reason to maintain alternative capabilities. Over time, workflows, staffing structures, data infrastructures and professional practices may reorganise around the artificial component until its removal would disrupt not merely efficiency but the institution's ability to understand or govern its domain.

Cognitive Dependency should therefore be distinguished from Artificial Cognitive Deference. **Deference concerns the tendency of human actors or institutional routines to privilege artificial outputs within judgement, whereas dependency concerns the structural capacity of the institution to perform, evaluate or substitute the cognitive function without continued access to the artificial component.** An institution may defer excessively to a system while retaining substantial independent capability, just as it may become operationally dependent upon a system whose outputs its experts continue to scrutinise critically.

This creates Hybrid Institutional Fragility. A cognitive architecture can appear highly capable under normal operating conditions while possessing limited resilience to model failure, infrastructure disruption, cyber incidents, provider withdrawal or abrupt changes in the environment for which the system was designed. The same specialisation that increases routine performance can therefore create systemic vulnerability.

Dependency also has an epistemic dimension. Institutions may remain operationally capable of replacing one technical system with another while lacking the expertise required to evaluate whether the replacement represents problems differently. Vendor substitution does not necessarily restore cognitive autonomy if the institution remains dependent upon external systems whose assumptions it cannot interrogate.

The appropriate response is not complete independence. Modern institutions necessarily depend upon external infrastructures, scientific expertise and specialised technologies. The objective is calibrated dependency: understanding which dependencies exist, which are acceptable and where redundancy or residual human capability is necessary because failure would compromise essential institutional cognition.

This analysis connects HAICA with the broader architecture developed in PGNM. Cognitive dependency can exist within individual institutions and propagate across governance networks when multiple nodes rely upon the same artificial infrastructure. A technological failure or representational bias can then become a network-level cognitive vulnerability.

Institutional Cognitive Deskilling represents one pathway through which such fragility can emerge, but the concepts should not be treated as equivalent. **Deskilling describes a longitudinal reduction in human cognitive capability associated with sustained delegation, whereas Hybrid Institutional Fragility describes the vulnerability of the wider cognitive configuration to disruption, degradation or loss of critical components.** Fragility may therefore arise without significant deskilling—for example through concentrated infrastructure dependency—while deskilling may occur without immediately producing systemic fragility if alternative cognitive pathways remain available.

Hybrid Institutional Fragility therefore emerges when the efficiency gained through artificial cognitive specialisation exceeds the resilience provided by alternative pathways, evaluative capacity and institutional understanding of its own dependencies.

20. Human–AI Cognitive Resilience

The existence of dependency does not make hybrid institutional cognition inherently fragile. Properly designed configurations can increase resilience by providing alternative analytical pathways, continuous monitoring and computational capacities that remain available when human actors are overloaded. The relevant question is whether the hybrid architecture can preserve essential cognitive functions when one or more components become unreliable.

HAICA defines **Human–AI Cognitive Resilience** as the capacity of a hybrid institutional cognitive architecture to maintain, recover or appropriately reconfigure essential cognitive functions when human, artificial or infrastructural components are disrupted, degraded or uncertain.

Resilience can emerge through functional redundancy. Critical assessments may be produced independently through human and artificial pathways before being integrated. Multiple models may provide alternative representations. Human expertise may be maintained for functions whose complete automation would create unacceptable dependency. In other settings, artificial systems may provide continuity when human teams are unavailable or overwhelmed.

Redundancy must nevertheless be selective. Requiring humans to reproduce every computational process would eliminate much of the value of augmentation, while maintaining multiple artificial systems performing identical functions can create substantial cost without genuine cognitive diversity. Resilience depends more upon **alternative cognitive pathways** than upon simple duplication.

Diversity is therefore important. Two models trained upon similar data and assumptions may fail simultaneously. Likewise, a human team whose reasoning has become strongly conditioned by artificial outputs may provide little independent protection against model error. Resilient configurations preserve enough

representational and methodological diversity for one component to detect or compensate for failures elsewhere.

Recovery capacity also matters. Institutions need procedures for recognising degraded system performance, switching configurations, escalating decisions and restoring normal operations. These mechanisms should exist before disruption occurs because designing alternative cognitive architecture during a crisis can substantially delay response.

Human–AI Cognitive Resilience ultimately depends upon treating hybrid cognition as infrastructure rather than convenience. **Once artificial systems participate in essential institutional cognition, resilience must be designed at the level of the complete cognitive configuration rather than assumed from the reliability of individual models or the continued formal presence of human decision-makers.**

21. Human–AI Feedback Architecture

Hybrid institutional cognition becomes genuinely adaptive only when institutions can observe not merely the outcomes of their decisions but the performance of the human–AI configurations through which those decisions were produced. A policy recommendation may prove inaccurate because the underlying evidence was incomplete, because an artificial system misrepresented uncertainty, because human reviewers interpreted the output poorly, because the allocation of cognitive functions was inappropriate or because the interaction among several components produced a failure that cannot be attributed meaningfully to any single node. Learning therefore requires feedback directed at the architecture itself.

HAICA describes the mechanisms through which such information returns to the cognitive configuration as **Human–AI Feedback Architecture**. This architecture connects implementation outcomes, model performance, human interpretation, disagreement, overrides, errors and near-misses with the processes through which functions are allocated and interfaces designed. Its purpose is not simply to improve model accuracy but to determine whether the complete hybrid configuration is supporting institutional cognition as intended.

Feedback may operate at several levels. Artificial systems can be evaluated against subsequent outcomes, while human reviewers can be assessed according to whether they detect model failures or introduce new errors. Interfaces can be examined to determine whether uncertainty was communicated effectively, and workflow data can reveal whether staff increasingly accept recommendations without meaningful review. The architecture therefore produces evidence concerning both technical and institutional behaviour.

This distinction is especially important because apparent model success can conceal configuration weakness. An artificial system may achieve high predictive performance while human expertise gradually deteriorates, or a workflow may produce

accurate decisions while concentrating epistemic authority in ways that reduce resilience. Conversely, occasional disagreement between humans and AI may represent healthy complementarity rather than dysfunction.

Feedback should therefore include indicators of interaction as well as outcomes. Override patterns, disagreement rates, escalation behaviour, traceability quality and recovery from system failure may all provide information about whether the architecture remains cognitively balanced. These observations should not automatically become optimisation targets, but they can help institutions recognise changes that would otherwise remain invisible.

Human–AI Feedback Architecture closes the loop between hybrid cognition and institutional learning by allowing the organisation to ask not only whether a decision worked, but whether the cognitive configuration that produced it remains appropriate.

22. Hybrid Institutional Learning

Feedback becomes cognitively significant when it alters future institutional behaviour. Within hybrid architectures, however, learning can occur in several locations simultaneously. Artificial systems may be retrained or recalibrated, human actors may change their interpretation practices, interfaces may be redesigned and organisational procedures may redistribute responsibilities. HAICA describes this broader process as **Hybrid Institutional Learning**.

Hybrid Institutional Learning differs from machine learning. A model updating its parameters does not necessarily mean that the institution has learned. The system may improve predictive accuracy while organisational assumptions remain unchanged, or human actors may continue using the model inappropriately despite technical improvement. Institutional learning occurs when evidence influences the wider configuration through which cognition is produced.

The reverse distinction is equally important. Institutions may learn without modifying the artificial system itself. Human actors may discover that a model performs poorly under particular conditions and develop new review procedures. Decision-makers may reinterpret outputs differently, or institutions may decide to remove the system from a particular cognitive function while retaining it elsewhere. Learning concerns architectural adaptation rather than technological modification alone.

This creates a multi-level learning problem. Some feedback concerns individual errors and can be addressed through local correction. Other evidence reveals systematic weaknesses in function allocation or interface design. More consequential findings may challenge the institution's assumptions about which forms of cognition should be delegated to artificial systems at all.

Hybrid learning therefore requires institutional memory concerning earlier configurations. Without records of how functions were allocated, which models were used and why particular safeguards existed, later changes can become difficult to

interpret. Hybrid Epistemic Traceability provides part of this memory by preserving the cognitive lineage of decisions and model versions.

The broader significance is that **a hybrid institution learns when experience can alter the relationships among human cognition, artificial capability and organisational process**, rather than merely improving the performance of one component. This makes learning an architectural property of HAICA.

23. Evaluating Human–AI Cognitive Configuration Performance

The introduction of HACC as the operational unit of analysis changes the object of evaluation. Conventional assessments often compare model accuracy, processing speed or human productivity, yet these measures reveal only parts of the cognitive system. A hybrid configuration can contain a highly accurate model and still produce poor institutional reasoning if human actors misinterpret its outputs, if traceability is weak or if the architecture encourages inappropriate deference.

HAICA therefore proposes that evaluation should occur at the level of the **Human–AI Cognitive Configuration**. The relevant question is whether the complete arrangement of human actors, artificial systems, interfaces, information resources and procedures improves institutional cognition relative to plausible alternative configurations.

Performance may need to be examined across several dimensions. Epistemic quality concerns whether the configuration produces representations sufficiently accurate, diverse and contextually appropriate. Interpretability concerns whether relevant actors can understand the basis and limitations of outputs. Resilience concerns whether essential cognition continues when components fail. Efficiency concerns the resources and time required to perform the task. Legitimacy concerns whether decision authority and accountability remain appropriately located.

These dimensions may conflict. A highly efficient configuration may reduce independent review. Greater redundancy may improve resilience while increasing cost. More detailed traceability may strengthen accountability while slowing decision processes. Evaluation therefore should not collapse performance into a single measure without acknowledging trade-offs.

Comparison becomes particularly valuable. Institutions can examine alternative HACC architectures for the same task: human-only processes, AI-assisted workflows, parallel human–AI analysis or AI-generated outputs subject to independent human review. Such comparisons can reveal whether observed benefits derive from AI capability itself or from the way cognition has been reorganised around it.

The development of formal indicators belongs primarily to F2, but HAICA establishes the conceptual requirement: **institutional AI performance should be evaluated at the level of the cognitive configuration rather than inferred from model capability alone.**

24. Hybrid Institutional Metacognition

An institution capable of observing the performance of its hybrid cognitive architecture can begin to develop a higher-order capacity: the ability to reason about how its own human and artificial cognition is organised. HAICA describes this as **Hybrid Institutional Metacognition**.

Hybrid Institutional Metacognition involves generating structured knowledge about the allocation of cognitive functions, patterns of human reliance, artificial influence, traceability, deskilling, dependency and resilience across institutional processes. It asks not only whether individual systems perform adequately but whether the architecture as a whole is evolving in desirable directions.

This capacity becomes increasingly important because hybridisation can occur gradually. AI systems may enter institutions through separate projects, departments or software tools without any actor possessing a complete view of how cognitive functions have shifted over time. One unit automates evidence retrieval, another adopts generative drafting, a third introduces predictive risk scoring and a fourth begins using AI-supported scenario analysis. Each intervention may appear limited, while their cumulative effect substantially reorganises institutional cognition.

Metacognitive mapping can reveal these shifts. Institutions can identify which processes depend critically upon artificial systems, where human expertise is declining, where cognitive authority has concentrated and where alternative pathways remain available. Such maps need not treat every use of AI as problematic; their purpose is to make architectural change visible.

Hybrid Institutional Metacognition also supports strategic decisions about institutional capability. An organisation may discover that it has acquired significant computational capacity but insufficient internal expertise to govern it, or that multiple departments depend upon the same external model infrastructure. These observations can inform training, procurement, redundancy and governance decisions.

Independent evaluation may be necessary because institutions can become invested in technologies they have adopted. External review, audit or cross-institutional comparison can help prevent metacognitive processes from reproducing internal assumptions.

Hybrid Institutional Metacognition is therefore the capacity of an institution to construct and revise representations of its own human–AI cognitive architecture so that technological adoption becomes an object of institutional understanding rather than an accumulation of isolated deployments.

25. Adaptive Cognitive Allocation

Once institutions can observe how cognitive functions are distributed and how those arrangements perform, allocation should no longer be treated as permanent. Artificial capabilities evolve, human expertise changes, tasks acquire new requirements and evidence concerning system performance accumulates. HAICA therefore extends Cognitive Function Allocation into **Adaptive Cognitive Allocation**.

Adaptive Cognitive Allocation refers to the deliberate reassessment and redistribution of cognitive functions among human, artificial and hybrid components in response to changing capability, context, risk or institutional learning. A function initially assigned primarily to humans may become suitable for substantial artificial augmentation as systems improve, while another may need to return to stronger human control after previously hidden weaknesses become visible.

This process differs from incremental automation. The direction of change is not assumed to move continuously from human towards artificial performance. Functions can migrate in either direction, and hybrid configurations may become more or less automated according to evidence. Adaptation therefore treats allocation as a reversible institutional design choice rather than a technological trajectory.

The principle of reversibility becomes particularly important under uncertainty. Institutions may introduce AI into a cognitive function experimentally while preserving the capacity to revert to alternative configurations if performance deteriorates. Such reversibility supports learning before dependence becomes structurally embedded.

Adaptive allocation also requires attention to human capability. If institutions plan to preserve the option of reallocating functions to human actors, relevant expertise must remain available. Deskilling can therefore reduce architectural flexibility by making formal reversibility impossible in practice.

The appropriate allocation may also vary across contexts within the same institution. An AI-supported process may function effectively in routine cases while unusual or high-stakes situations require stronger human cognition. Adaptive allocation can therefore operate dynamically through escalation rules rather than only through periodic organisational redesign.

These capacities form a related but analytically distinct adaptive sequence. **Human–AI Feedback Architecture returns evidence concerning the behaviour and consequences of the configuration; Hybrid Institutional Learning incorporates relevant evidence into institutional knowledge and practice; Hybrid Institutional Metacognition constructs representations of the architecture itself; and Adaptive Cognitive Allocation uses that higher-order understanding to reconsider where cognitive functions should reside.** The sequence need not operate linearly in every case, but distinguishing these functions prevents feedback, learning, self-observation and architectural adaptation from collapsing into a single undifferentiated concept.

A mature hybrid institution does not decide once where cognition should reside; it develops the capacity to reconsider that distribution as technologies, tasks and institutional knowledge evolve.

26. Cognitive Reconfiguration over Time

Adaptive allocation affects individual cognitive functions, but repeated changes can gradually transform the wider architecture of the institution. HAICA therefore distinguishes local reallocation from **Hybrid Cognitive Reconfiguration**, the broader evolution of relationships among human actors, artificial systems, interfaces and organisational structures over time.

Reconfiguration may occur deliberately through institutional redesign, but it can also emerge incrementally. Staff roles change as AI absorbs particular analytical functions; new oversight bodies appear; information flows shift towards computational platforms; professional expertise becomes concentrated in different organisational units; and decision procedures adapt to faster analytical cycles. The institution can therefore become structurally different even when no formal reform explicitly declared a new cognitive architecture.

Path dependence influences this evolution. Early choices concerning vendors, data formats, model infrastructures or workflow integration can constrain later options. A system initially introduced as temporary support may become deeply embedded because other processes are built around it. Switching costs then become cognitive as well as technical.

Reconfiguration can also produce new institutional roles. Some organisations may develop specialised functions for model validation, AI governance, hybrid workflow design or cognitive audit. These roles become interfaces between technical capability and organisational judgement, potentially forming a new layer of institutional metacognition.

The historical development of hybrid architectures should therefore become an object of research in its own right. Comparing institutions at a single moment may obscure how current dependencies or strengths emerged through successive technological decisions. Longitudinal analysis can reveal whether particular allocation patterns create resilient learning or lock institutions into increasingly fragile structures.

Hybrid cognition is not a stable technological state but an evolving institutional architecture whose present configuration reflects accumulated decisions about where cognition has been located and how those locations have become connected.

27. Failure Modes of Human–AI Institutional Collaboration

The preceding analysis allows HAICA to identify several distinct mechanisms through which hybrid cognition can fail. These failures should not be treated as a generic taxonomy of AI risk; they concern specifically the architecture connecting human and artificial cognitive contributions.

Allocation Failure occurs when a cognitive function is assigned to a human, artificial or hybrid component poorly suited to its epistemic, contextual or institutional requirements. A technically capable system may perform inadequately because the function contains forms of judgement that the allocation process failed to recognise.

Interface Failure arises when relevant information, assumptions or uncertainty cannot move effectively between human and artificial components. Both may perform competently in isolation while their interaction produces distortion.

Translation Failure occurs when human and artificial representations appear mutually intelligible while important differences in meaning, context or uncertainty remain unresolved.

Traceability Failure emerges when institutions cannot reconstruct how artificial and human contributions shaped the representations underlying a decision, weakening both learning and accountability.

Deference Failure occurs when artificial outputs acquire cognitive authority disproportionate to their evidential reliability, causing human actors to cease functioning as independent sources of interpretation.

Overload Failure appears when artificial systems generate cognitive throughput beyond the interpretative capacity of human participants, encouraging shortcuts and superficial review.

Deskilling Failure develops when sustained delegation erodes human expertise required for meaningful evaluation, recovery or independent judgement.

Authority Failure occurs when effective cognitive influence becomes detached from the structures through which legitimate decision authority and responsibility are formally organised.

Finally, **Feedback Failure** arises when the institution cannot evaluate whether the complete Human–AI Cognitive Configuration improved or weakened institutional cognition, preventing experience from altering future architecture.

These mechanisms can reinforce one another. Poor allocation can increase overload, overload can encourage deference, deference can accelerate deskilling and deskilling can deepen dependency. **Hybrid cognitive failure is therefore frequently cumulative and architectural rather than reducible to a single defective model or human error.**

28. Responsibility within Hybrid Cognitive Architectures

The distributed nature of hybrid cognition complicates conventional accounts of responsibility. Institutional decisions may depend upon model developers, data providers, analysts, administrators, external vendors and authorised decision-makers, each contributing different elements to the final cognitive process. When outcomes prove harmful or erroneous, responsibility can become difficult to locate if the architecture has not preserved a clear relationship between contribution and authority.

Responsibility should therefore not be attributed exclusively according to causal influence. A model may strongly influence an outcome without possessing moral or legal agency, while a human decision-maker may possess formal authority without having controlled every upstream representation. HAICA requires an institutional account capable of distinguishing technical contribution, cognitive influence and legitimate responsibility.

Authorised institutions remain responsible for the systems they choose to deploy and the architectures within which those systems operate. Delegating analytical functions does not eliminate the obligation to establish appropriate standards, review mechanisms and boundaries of use. At the same time, responsibility can be distributed operationally among actors responsible for data quality, model validation, interface design and human review.

Traceability becomes essential because meaningful accountability requires reconstructing who contributed what and where institutional safeguards were expected to operate. Without this information, responsibility can diffuse across a socio-technical system until no actor can explain why the final decision emerged.

The architecture should also distinguish responsibility for individual decisions from responsibility for institutional design. A frontline official may act reasonably within a workflow whose structure systematically encourages deference. The deeper responsibility may lie with those who designed or maintained the cognitive configuration rather than with the individual user.

Hybrid institutional responsibility therefore concerns both decisions and architectures: institutions must remain accountable not only for what they decide with AI, but for how they organise the human and artificial cognition through which those decisions become possible.

29. Legitimate Authority in Hybrid Institutions

Artificial systems can acquire substantial cognitive influence without acquiring legitimate political or administrative authority. This distinction has appeared throughout HAICA, but hybrid architecture makes its constitutional significance increasingly clear. The more deeply AI participates in problem representation, evidence selection and alternative generation, the more important it becomes to preserve a meaningful relationship between cognitive influence and authorised institutional judgement.

Legitimate authority does not require human actors to perform every cognitive function personally. Institutions have always relied upon advisers, experts and technical systems. The constitutional issue concerns whether authorised actors retain sufficient capacity to understand the nature of the representations informing their choices, evaluate relevant alternatives and assume responsibility for the resulting decision.

This becomes difficult when cognitive authority becomes highly concentrated upstream. A decision-maker may retain formal discretion while encountering a policy space already narrowed by artificial systems. The final choice can then be human without the preceding architecture being genuinely contestable.

Meaningful Human Judgement therefore functions as one safeguard, but legitimate authority also requires institutional procedures through which model assumptions and evidence can be challenged before they become embedded within decisions. Oversight cannot be reduced to an individual human reviewer if the cognitive architecture itself lies beyond that person's ability to influence.

Different forms of authority may require different safeguards. Administrative decisions affecting individual rights may demand strong traceability and review, while exploratory scientific modelling may operate with considerably greater artificial autonomy because no immediate coercive decision follows. HAICA therefore treats legitimacy requirements as context-dependent rather than uniform.

The principle remains stable: **artificial cognitive participation can expand the informational basis of institutional authority but cannot independently establish the legitimacy of collective choice.** Hybrid institutions remain governance institutions precisely because decisions continue to occur within structures of responsibility, law and accountable authority that computation alone does not provide.

30. Towards an Integrated HAICA Architecture

The relationships developed throughout the paper can now be assembled into an integrated architecture of human–AI institutional cognition. Institutional tasks are first examined through Cognitive Function Decomposition, making visible the forms of perception, retrieval, interpretation, modelling, judgement and coordination they require. Cognitive Function Allocation then determines how those functions are distributed among human, artificial and hybrid components according to contextual, epistemic and institutional conditions.

These components become organised within Human–AI Cognitive Configurations connected through Hybrid Cognitive Interfaces. Human–AI Cognitive Translation allows differently structured representations to become mutually usable, while Hybrid Epistemic Traceability preserves the lineage through which evidence and interpretation influence institutional reasoning. Human–AI Cognitive Complementarity emerges where differentiated capabilities remain sufficiently independent and connected to compensate for limitations elsewhere within the configuration.

Meaningful Human Judgement links artificial cognitive contribution with legitimate institutional authority, while the distinction between Cognitive Authority and Decision Authority makes visible situations in which formal human control conceals deeper shifts in epistemic influence. Artificial Cognitive Deference, Augmentation Overload, Institutional Cognitive Deskilling and Cognitive Dependency describe pathways through which complementarity can deteriorate into fragility.

Human–AI Feedback Architecture returns evidence concerning both outcomes and configuration performance, supporting Hybrid Institutional Learning and Hybrid Institutional Metacognition. These capacities allow institutions to reconsider allocation and undertake Adaptive Cognitive Allocation or broader Hybrid Cognitive Reconfiguration as capabilities and conditions change.

The architecture is therefore recursive. No allocation should be regarded as permanently optimal, and no artificial system should be treated as possessing institutional value independent of the relationships surrounding it. The relevant object remains the evolving hybrid configuration through which cognition becomes collectively produced.

HAICA consequently represents human–AI institutional collaboration as an adaptive cognitive architecture rather than a static division of labour between people and machines. Its central concern is whether heterogeneous cognitive capabilities can remain complementary, contestable, traceable and legitimately integrated as the institution itself evolves.

31. Institutional Cognitive Autonomy

As artificial systems become more deeply embedded within institutional reasoning, the question of autonomy extends beyond formal organisational independence. An institution may retain complete legal authority over its decisions while becoming increasingly dependent upon external models, data infrastructures or computational services for the cognitive processes through which those decisions are formed. HAICA therefore introduces the concept of **Institutional Cognitive Autonomy** to describe an institution's capacity to understand, evaluate and, where necessary, reconfigure the cognitive systems upon which its judgement depends.

Cognitive autonomy does not require technological self-sufficiency. Contemporary institutions already depend upon external expertise, shared scientific infrastructures and specialised technologies whose complete replication would be impractical. The relevant question is whether dependence remains governable. An institution preserves cognitive autonomy when it retains sufficient understanding of critical systems, access to alternative analytical pathways, capacity to challenge external representations and authority to change or discontinue configurations that no longer meet institutional requirements.

This becomes particularly important where artificial systems shape upstream cognitive functions. If evidence retrieval, classification, risk assessment or scenario generation depend upon external computational infrastructures, institutional understanding can become partially mediated through systems whose internal assumptions are difficult to inspect. The institution may continue making its own decisions while losing practical control over the representations that structure those decisions.

Autonomy therefore depends partly upon internal expertise. Public institutions need not reproduce every technical capability, but they require enough technical and domain knowledge to evaluate whether externally generated outputs remain appropriate. Procurement alone cannot substitute for cognitive governance. When an institution cannot determine whether a model's assumptions, data or operating conditions remain compatible with its own mandate, technical outsourcing becomes epistemic outsourcing.

Institutional Cognitive Autonomy also has a temporal dimension. Capabilities that appear optional during early adoption can become structurally necessary once workflows, staffing and institutional memory reorganise around them. Autonomy must therefore be assessed not only at deployment but across the lifecycle of hybridisation.

A hybrid institution remains cognitively autonomous not when it avoids dependence altogether, but when its dependencies remain visible, contestable and reversible enough for the institution to preserve meaningful control over the architecture of its own cognition.

32. Vendor and Infrastructure Dependency

Many artificial cognitive systems enter institutions through external technological infrastructures rather than internally developed models. Cloud platforms, proprietary foundation models, specialised data services and software ecosystems can provide capabilities that would be costly or impossible for individual institutions to reproduce. These arrangements can accelerate adoption, but they also introduce dependencies whose institutional consequences extend beyond conventional procurement.

Vendor dependency becomes cognitively significant when external infrastructure participates in essential functions of institutional perception, memory, interpretation or reasoning. A change in model behaviour, pricing, access conditions or service availability can then alter institutional cognition even when no internal policy decision has been made. The architecture becomes partially governed by actors whose incentives and accountability structures differ from those of the institution using the system.

The problem is not unique to private providers. Institutions can become dependent upon shared public infrastructures, international analytical platforms or central government systems as well. What matters is the concentration of cognitive capability outside the organisational boundary and the degree to which alternatives remain available.

Proprietary opacity creates an additional difficulty. Institutions may be able to evaluate outputs empirically without possessing full access to training data, model architecture or internal transformations. Such limitations do not automatically make a system unusable, but they change the conditions under which traceability and accountability can be achieved. The institution may need stronger external validation, contractual safeguards or alternative models precisely because direct inspection remains limited.

Infrastructure dependency can also create systemic concentration when many institutions rely upon the same provider or model family. A shared failure, bias or disruption can then propagate across otherwise independent organisations. What appears locally as efficient outsourcing may become network-level cognitive fragility.

HAICA therefore treats vendor and infrastructure dependency as an architectural property rather than merely a commercial consideration. **The institution must understand which cognitive functions depend upon external infrastructures, how those dependencies could fail and what alternative pathways remain available if the external system becomes unreliable or inaccessible.**

33. Diversity, Disagreement and Cognitive Independence

Hybrid cognitive architectures derive much of their potential value from the fact that human and artificial systems do not always represent problems in the same way. This difference can expose assumptions, identify missing evidence and create opportunities for mutual correction. Yet institutional processes often treat disagreement as inefficiency, creating pressure to harmonise outputs and accelerate convergence.

HAICA instead treats **cognitive independence** as a potential source of resilience. Human actors should not be expected simply to validate artificial conclusions, nor should AI systems be configured solely to reproduce established institutional judgements. When different cognitive pathways remain sufficiently independent, disagreement can become diagnostically useful.

A divergence between human and artificial assessments can indicate several possibilities. The model may have detected a pattern that human actors overlooked; human experts may possess contextual knowledge absent from the computational representation; evidence may genuinely support several interpretations; or one component may simply be wrong. The architecture should therefore use disagreement as a trigger for inquiry rather than automatically privileging one source.

This principle has implications for model design. Systems optimised primarily for agreement with historical human decisions may reduce independent analytical value, particularly where previous institutional practices contained systematic blind spots. Conversely, artificial systems trained on narrow datasets may produce apparently novel recommendations that reflect hidden representational limitations rather than meaningful independence.

Diversity also matters among artificial systems. Multiple models relying upon different methods, data or assumptions can provide stronger protection against shared error than several systems built upon the same underlying architecture. Yet diversity increases interpretative burden, reinforcing the need for interfaces capable of managing disagreement without producing Augmentation Overload.

The objective is therefore not maximal disagreement. Institutions require sufficient convergence to act. **The cognitive value lies in preserving enough independence for disagreement to reveal uncertainty and error before coordination transforms plural interpretations into institutional judgement.**

34. Hybrid Cognitive Security

Once artificial systems participate in institutional cognition, cybersecurity acquires a cognitive dimension. Attacks need not merely disrupt technical infrastructure; they can alter the representations through which institutions perceive and reason about their environment. Data poisoning, manipulated retrieval sources, adversarial inputs, compromised models or deceptive generated content can therefore become forms of **Hybrid Cognitive Security** risk.

Hybrid Cognitive Security concerns the protection of institutional cognitive processes against deliberate or accidental corruption across human and artificial components. Its object is not only system availability or confidentiality, but the integrity of the representations entering institutional judgement.

This distinction is important because a compromised cognitive system may continue operating normally at the technical level. An AI-supported monitoring platform can remain available while receiving manipulated data. A generative system can produce coherent summaries while relying upon compromised sources. The resulting failure may therefore appear epistemic rather than operational.

Human actors provide both resilience and vulnerability. Experienced analysts may detect outputs inconsistent with contextual knowledge, while automation bias can make manipulated recommendations more effective precisely because they arrive through trusted systems. Artificial systems can likewise help identify anomalies that humans miss, creating opportunities for complementary security architectures.

Traceability becomes particularly important during cognitive incidents. Institutions need to reconstruct which data, models and outputs were affected and which decisions depended upon them. Without this capacity, corruption can persist within organisational memory even after the technical vulnerability has been resolved.

Security also reinforces the need for redundancy and independent pathways. A hybrid institution whose essential cognition depends upon a single computational infrastructure has limited capacity to distinguish a genuine environmental change from a compromised representation.

Hybrid Cognitive Security therefore extends institutional security from protecting computational systems to protecting the integrity, provenance and diversity of the cognitive processes through which institutions construct knowledge and judgement.

35. The Limits of Artificial Cognitive Participation

The expansion of artificial capability can encourage institutions to assume that progressively more cognitive functions should become computationally mediated as systems improve. HAICA rejects any automatic relationship between capability expansion and institutional allocation. Some functions may become increasingly suitable for artificial participation, while others may remain inappropriate under particular conditions because of uncertainty, contextual complexity, legitimacy requirements or insufficient traceability.

The limits of artificial participation are therefore contingent rather than metaphysical. HAICA does not attempt to establish permanent categories of cognition that artificial systems can never perform. Technological development may change rapidly, and claims concerning fixed human exclusivity would make the architecture scientifically fragile. The relevant question is whether a particular artificial contribution can currently satisfy the epistemic and institutional requirements of the function in which it is being placed.

High uncertainty can create one boundary. Models may perform well within familiar distributions while becoming unreliable when conditions change substantially. Novel institutional environments therefore require stronger mechanisms for human interpretation and independent evidence.

Normative content creates another boundary. Artificial systems can represent competing values, identify distributional consequences and support deliberation, but legitimate authority to determine contested public priorities remains institutionally grounded. Increasing analytical sophistication does not remove the political nature of these choices.

Irreversibility and stakes also matter. An artificial contribution that is acceptable within a low-risk and easily reversible administrative process may be inappropriate within decisions involving fundamental rights or systemic consequences unless substantially stronger safeguards exist.

The limits of artificial participation should therefore be understood through the architecture rather than through abstract claims about intelligence. **A cognitive function becomes suitable for greater artificial participation when the complete institutional configuration can preserve adequate reliability, traceability, contestability, resilience and legitimate judgement around that participation.**

36. The Boundaries of HAICA

The analytical reach of HAICA must remain deliberately bounded. Human–AI institutional collaboration intersects with AI ethics, algorithmic accountability, labour transformation, cybersecurity, technology procurement, organisational psychology and many other fields. The framework can connect with these domains without attempting to absorb them into a single theory.

HAICA is not a general framework for assessing whether an AI system is safe or ethical. A model may satisfy technical safety standards while being poorly positioned within an institutional cognitive architecture, just as a technically imperfect system may still provide useful exploratory support within carefully bounded settings. The object of analysis is the relationship between artificial capability and institutional cognition.

Nor is HAICA primarily a workforce automation model. Changes in cognitive allocation can affect professional roles and employment, but the framework does not attempt to predict labour-market consequences or determine which occupations should be automated. Its concern is whether the redistribution of cognitive functions strengthens or weakens institutional capacity.

The architecture should also avoid treating every institutional use of AI as Hybrid Institutional Cognition. Many applications remain operationally peripheral. HAICA becomes most relevant when artificial systems contribute materially to perception, interpretation, reasoning, judgement support, learning or other processes through which institutional understanding is formed.

Finally, HAICA cannot determine legitimate public values. It can identify where artificial systems influence normative choices, whether humans retain meaningful judgement and how authority is distributed, but the framework cannot derive the constitutional arrangements through which societies should govern particular technologies.

HAICA should therefore be understood as a middle-range architecture for investigating how human and artificial cognitive capabilities become organised within institutions, not as a comprehensive theory of artificial intelligence, institutional ethics or technological governance.

37. HAICA as an Architecture of Hybrid Institutional Cognition

The concepts developed throughout this Working Paper can now be integrated into a coherent architecture. Institutional tasks are first examined through Cognitive Function Decomposition, allowing the distinct forms of cognition embedded within organisational work to become visible. Cognitive Function Allocation then distributes those functions among human, artificial and hybrid components according to task requirements, evidence, risk and institutional context.

Human–AI Cognitive Configurations provide the task-specific arrangements through which those allocations become operational. Hybrid Cognitive Interfaces connect components, while Human–AI Cognitive Translation allows representations generated by different cognitive systems to become mutually usable. Hybrid Epistemic Traceability preserves the lineage through which evidence, artificial transformation and human interpretation contribute to institutional reasoning.

Human–AI Cognitive Complementarity emerges when differentiated capabilities remain sufficiently independent and connected to correct or extend one another. Meaningful Human Judgement preserves the substantive human capacity to interpret and challenge artificial contributions, while the distinction between Cognitive Authority and Decision Authority reveals whether formal institutional control corresponds to the actual distribution of epistemic influence.

The architecture also contains mechanisms for recognising deterioration. Artificial Cognitive Deference, Augmentation Overload, Institutional Cognitive Deskilling and Cognitive Dependency identify pathways through which augmentation can become fragility. Human–AI Cognitive Resilience preserves alternative pathways, while Hybrid Cognitive Security protects the integrity of the representations entering institutional cognition.

Human–AI Feedback Architecture allows the institution to observe how the configuration performs, Hybrid Institutional Learning translates evidence into change and Hybrid Institutional Metacognition makes the architecture itself an object of institutional knowledge. Adaptive Cognitive Allocation and Hybrid Cognitive Reconfiguration then provide mechanisms through which the distribution of cognition can evolve over time.

The resulting architecture is neither human-centred in the sense of treating artificial systems as permanently peripheral nor automation-centred in the sense of assuming that technological capability should determine institutional organisation. It is **institution-centred**, asking what distribution of heterogeneous cognitive capabilities best supports the institution's capacity to understand, judge, act and learn while preserving accountability and legitimate authority.

HAICA therefore frames human–AI collaboration as the deliberate organisation of hybrid institutional cognition: an evolving architecture in which artificial capability becomes valuable only when it remains connected to human judgement, cognitive diversity, epistemic traceability, resilience and institutional responsibility.

38. HAICA as a Research Framework

The Human–AI Institutional Collaboration Architecture becomes scientifically consequential only when its concepts can support systematic observation of how hybrid cognition actually occurs within institutions. Artificial intelligence is entering public and organisational environments through a wide variety of technological forms, professional practices and administrative workflows, making it increasingly difficult to infer institutional consequences from the capabilities of particular models alone. HAICA therefore provides a research framework through which attention can shift from technological performance towards the cognitive configurations in which artificial systems become institutionally embedded.

The first research task concerns identification. Researchers must determine where artificial cognitive participation occurs within institutional processes and which functions are being redistributed as a result. Formal descriptions of AI adoption may provide only a partial picture. An institution may officially use AI for administrative support while the system in practice structures evidence retrieval, interpretation or problem representation in ways that give it considerable cognitive influence. Conversely, a technologically sophisticated deployment may remain cognitively peripheral if human actors retain independent analytical pathways and artificial outputs contribute only marginally to institutional judgement.

Human–AI Cognitive Configurations provide the principal unit through which these relationships can be reconstructed. Researchers can identify the human actors, artificial systems, information resources, interfaces and organisational procedures involved in a particular institutional task, then examine how cognition travels among them. Such reconstruction allows the location of cognitive influence to be distinguished from the location of formal authority, making it possible to study whether the architecture corresponds to institutional descriptions of responsibility.

A second research task concerns mechanisms. HAICA proposes that complementarity, deference, overload, deskilling, dependency and resilience arise through particular relationships rather than as inevitable properties of AI adoption. Empirical research can therefore investigate when these mechanisms appear and under what conditions they become consequential. An AI-assisted workflow may improve evidence integration in one institution while producing excessive cognitive deference in another because review practices, expertise or workload differ.

A third task concerns change over time. Hybrid cognitive architectures are unlikely to remain stable as models improve, institutional actors gain experience and workflows adapt. Longitudinal research will therefore be especially important for identifying whether early augmentation produces sustained complementarity or gradually evolves towards dependency, deskilling or cognitive concentration.

HAICA thus provides a framework for asking not simply whether institutions use artificial intelligence, but **how the presence of artificial cognitive systems changes the architecture through which institutional knowledge, judgement, responsibility and learning are produced**. Its scientific development will depend upon whether these

distinctions remain observable and explanatory across diverse institutional environments.

39. Empirical Reconstruction and Human–AI Cognitive Mapping

The distributed nature of Hybrid Institutional Cognition creates a methodological problem similar to that encountered in the study of wider cognitive networks: much of the relevant architecture is not visible within formal organisational structures. Job descriptions may identify who possesses decision authority without revealing which artificial systems select the evidence those actors encounter, while software inventories may identify deployed technologies without explaining how their outputs influence institutional reasoning. Empirical research therefore requires methods capable of reconstructing the realised hybrid cognitive configuration.

HAICA can support a form of **Human–AI Cognitive Mapping** in which researchers trace cognitive functions, actors, artificial systems, interfaces and flows across a specific institutional process. The objective is not simply to document where AI appears but to identify which transformations occur before and after artificial participation. Evidence may enter an AI retrieval system, be synthesised by a second model, interpreted by an analyst and subsequently compressed into a briefing for an authorised decision-maker. Mapping this sequence makes visible where assumptions enter, where uncertainty changes and where cognitive authority becomes concentrated.

Process tracing can complement such mapping by following individual decisions or analytical outputs through the complete configuration. Researchers might reconstruct how a risk classification was generated, which human actors reviewed it, whether alternative interpretations were considered and how the final institutional judgement differed from the initial artificial output. Such analysis can reveal whether formal oversight corresponds to meaningful human judgement or merely to procedural confirmation.

Observation and interview methods will remain important because many aspects of hybrid cognition are tacit. Institutional actors may use AI-generated outputs informally, consult systems outside official workflows or develop personal strategies for deciding when recommendations deserve trust. These practices can substantially influence cognitive architecture while leaving limited documentary evidence.

Digital logs may provide unusually detailed traces of artificial participation, yet their apparent precision should not obscure the parts of cognition they cannot capture. A system can record that a recommendation was overridden without revealing whether the human actor understood the underlying model or relied upon contextual intuition. Conversely, interviews can reveal interpretation while being vulnerable to retrospective reconstruction. Triangulation will therefore be essential.

Human–AI Cognitive Mapping should ultimately enable researchers to distinguish **nominal AI adoption from realised hybrid cognition**. Its purpose is to reconstruct how

knowledge actually moves through a configuration and where human and artificial contributions become consequential within institutional judgement.

40. Comparative Research across Human–AI Cognitive Configurations

The diversity of possible Human–AI Cognitive Configurations creates substantial opportunities for comparative research. Institutions can organise the same cognitive task through different arrangements, allowing researchers to examine how alternative allocations, interfaces and review mechanisms affect institutional reasoning. Such comparison is particularly valuable because the performance of a particular AI model may reveal little about the architecture in which it should be deployed.

A policy-analysis function, for example, might be conducted through a predominantly human workflow supported by AI retrieval, an AI-first synthesis subject to expert review, parallel independent human and artificial analysis, or a multi-model configuration in which disagreement triggers additional investigation. Each arrangement uses artificial capability differently and creates different distributions of cognitive authority, workload and resilience.

Comparison can occur within a single institution by examining alternative workflows across departments, or across institutions using similar technologies under different organisational conditions. Cross-institutional comparison may reveal whether apparent technological effects are actually produced by differences in professional expertise, review culture, workload, legal mandate or institutional capacity.

Temporal comparison is equally important. A configuration that performs effectively during early deployment may change as human actors become accustomed to artificial outputs. Patterns of override can decline because the model improves, because users become more trusting or because independent expertise deteriorates. These possibilities cannot be distinguished through cross-sectional performance measures alone.

Comparative research should therefore evaluate more than accuracy or productivity. It can examine the distribution of disagreement, frequency and quality of human intervention, traceability, recovery from error, preservation of expertise, workload effects and changes in cognitive dependency. Some of these dimensions may later become operationalised through F2, but their conceptual significance can already guide empirical inquiry.

The objective is not to discover a universally optimal HACC. Different institutional tasks impose different requirements concerning stakes, uncertainty, speed and legitimacy. **Comparative research should instead identify which architectural relationships support effective complementarity under particular conditions and**

which configurations create recurring pathways towards fragility or inappropriate cognitive concentration.

41. Progressive Validation of HAICA

Validation of HAICA cannot depend upon demonstrating that human–AI collaboration is consistently superior to either human-only or more automated alternatives. The architecture explicitly rejects such a universal proposition. Hybrid cognition can increase institutional capacity in some settings while weakening it in others, and the purpose of validation is therefore to determine whether the relationships identified by HAICA help explain these differences.

Construct validation provides a first level. Researchers must determine whether concepts such as Meaningful Human Judgement, Artificial Cognitive Deference, Augmentation Overload, Cognitive Function Allocation and Institutional Cognitive Deskilling can be identified consistently across real institutional settings. If these concepts prove too ambiguous to distinguish empirically, the architecture will require refinement.

Mechanism validation provides a second level. HAICA proposes, for example, that inadequate traceability can weaken learning and accountability, that excessive artificial throughput can increase deference, and that prolonged delegation can reduce evaluative human capacity. Longitudinal and comparative studies can examine whether these relationships occur under the conditions predicted by the framework.

Configuration-level performance provides a third level. Alternative HACCs can be compared according to their capacity to produce accurate, contextually appropriate, traceable, resilient and legitimate institutional reasoning. No single metric will be sufficient because these properties can conflict. A configuration that increases speed while reducing meaningful review may represent improvement for some tasks and deterioration for others.

Intervention studies can provide stronger evidence concerning particular mechanisms. Institutions might modify interface design, introduce independent human review, preserve parallel analytical pathways or change escalation rules and then observe whether patterns of deference, error detection or resilience change accordingly. Such interventions can transform HAICA from a descriptive architecture into a source of testable institutional hypotheses.

Simulation may eventually complement empirical work by examining how different distributions of human and artificial cognition behave under controlled assumptions. Yet simulation cannot validate the framework independently because many of its central propositions concern institutional judgement, expertise and responsibility whose behaviour requires observation in real organisational settings.

HAICA will become scientifically stronger when its concepts allow researchers to distinguish why superficially similar AI deployments produce

different institutional consequences and when those distinctions remain robust across technologies, tasks and organisational contexts.

42. Methodological Cautions in Human–AI Institutional Research

Research into hybrid institutional cognition faces several methodological risks generated by the novelty and visibility of artificial intelligence itself. The first is technological salience. AI systems are often easier to identify than the human and organisational structures surrounding them, encouraging researchers to attribute changes in institutional performance directly to technological adoption while underestimating complementary reforms, expertise or procedural changes.

A second risk concerns vendor-defined evidence. Many systems are accompanied by performance claims generated under conditions that differ from their institutional use. Benchmarks may establish technical capability without demonstrating how the model behaves when embedded within complex workflows involving incomplete information, professional judgement and accountability. Institutional research must therefore distinguish model evaluation from configuration evaluation.

A third problem arises from rapidly changing technology. A Human–AI Cognitive Configuration observed at the beginning of a longitudinal study may incorporate substantially different models by the end. Research focused too narrowly upon individual model generations can become obsolete before institutional mechanisms are understood. HAICA therefore prioritises cognitive functions, interfaces and allocations precisely because these provide more stable analytical objects.

Observation can also affect behaviour. Staff may exercise greater scrutiny when they know their interaction with AI systems is being studied, while informal reliance or workarounds may disappear temporarily under evaluation. Researchers should therefore remain cautious when interpreting observed review behaviour as representative of routine practice.

Self-reporting introduces different difficulties. Human actors may overestimate their independence from artificial outputs or, conversely, understate how useful systems have become because of organisational sensitivities surrounding automation. Digital interaction data can provide complementary evidence but remain incomplete representations of cognitive reasoning.

Finally, hybrid institutional research should avoid treating greater human participation as inherently better. Human judgement can introduce bias, inconsistency and error, just as artificial systems can. The purpose of HAICA is not to defend human cognition categorically but to understand how heterogeneous capabilities should be organised.

Methodological neutrality therefore requires resisting both technological enthusiasm and human exceptionalism, evaluating the complete cognitive

architecture according to the functions and institutional conditions it is expected to support.

43. HAICA in Relation to ICCA and ICAM

HAICA represents a specialised development of the conceptual and architectural foundations established by ICCA and ICAM. ICCA identified institutional cognition as a field concerned with distributed perception, memory, reasoning, decision-making, learning and adaptation, while ICAM provided a model of how those functions become organised across cognitive nodes, interfaces, information flows and feedback structures. HAICA takes that architecture as its starting point and examines what changes when some of those nodes perform increasingly sophisticated artificial cognitive functions.

This extension preserves several central propositions. Institutional intelligence remains relational rather than additive; cognitive capacity remains distinct from legitimate authority; interfaces determine whether specialised capabilities become systemically consequential; and feedback connects present cognition with institutional learning. HAICA does not replace these concepts but makes them more specific under conditions of hybrid cognition.

Artificial Cognitive Participation creates new architectural questions because computational nodes can become highly influential within representation and reasoning while lacking legitimate agency or responsibility. ICAM already allowed technological systems to function as cognitive nodes, but HAICA develops the institutional consequences of that inclusion in much greater detail by introducing Cognitive Function Allocation, Meaningful Human Judgement, Artificial Cognitive Deference, Hybrid Epistemic Traceability and Adaptive Cognitive Allocation.

The Human–AI Cognitive Configuration also provides a more specific unit of analysis nested within ICAM. A HACC can be understood as an activated cognitive configuration in which particular human and artificial nodes participate in a task-specific arrangement. This inheritance preserves conceptual interoperability between the two models while allowing HAICA to investigate mechanisms that would have overloaded the more general ICAM framework.

The relationship is therefore one of analytical specialisation. **ICAM explains how institutional cognition is architecturally organised; HAICA explains how that architecture can be deliberately examined and redesigned when cognitive participation becomes increasingly distributed between human and artificial systems.**

44. HAICA in Relation to CPDF and PGNM

CPDF and PGNM provide two different environments within which Hybrid Institutional Cognition can become operationally significant. CPDF examined computationally augmented policy design, showing how models and artificial systems can contribute to problem representation, evidence integration, policy-space exploration and recursive learning. HAICA provides a more detailed account of the human–artificial relationships through which those contributions enter institutional judgement.

A CPDF workflow that uses artificial intelligence for evidence synthesis or scenario generation can therefore be represented as a particular HACC. HAICA allows researchers to ask whether cognitive functions were allocated appropriately, whether human actors retained meaningful judgement, whether provenance remained visible and whether the configuration generated overload or dependency. The two frameworks thus operate at different levels of specificity while remaining closely interoperable.

PGNM extends the relationship across institutions. Artificial systems can support planetary sensing, cross-scale translation, network memory and knowledge integration, meaning that hybrid cognitive configurations may exist not only within individual organisations but across inter-institutional networks. Shared models or computational infrastructures can become cognitive nodes upon which several institutions depend simultaneously.

This creates network-level consequences that HAICA alone cannot fully explain. A configuration may appear resilient within one institution while creating systemic dependency across many PGNM nodes if all rely upon the same artificial infrastructure. Conversely, a governance network may preserve cognitive diversity by maintaining independent human–AI configurations across jurisdictions.

The models therefore intersect recursively. HAICA provides the architecture of hybrid cognition within institutional and task-specific configurations, while PGNM shows how those configurations can participate in wider networks of distributed governance cognition. **Together they allow artificial cognitive participation to be studied across several scales without assuming that technological systems have the same institutional significance wherever they appear.**

45. HAICA within the Emerging Operational Model Suite

Within the wider L02 Operational Model Suite, HAICA occupies a particularly important position because many of the subsequent research programmes will depend upon the ability to observe and compare hybrid cognitive configurations systematically. F2 can translate HAICA's architectural properties into candidate indicators, G2 can simulate alternative arrangements and I2 can test selected configurations within real institutional environments.

The Governance Cognitive Performance Indicator Framework will require concepts capable of distinguishing model performance from configuration performance. HAICA's principal architectural properties may provide candidate dimensions, although F2 must determine which can become valid observables without encouraging superficial quantification.

Governance Simulation Platforms can provide a complementary methodological environment. Different configurations can be represented computationally to explore how changes in allocation, model reliability, workload or interface structure influence system behaviour. Simulation might compare parallel human–AI analysis with sequential workflows or examine how cognitive dependency affects resilience under model failure. Such experiments will remain abstractions but can generate hypotheses for institutional research.

AI Governance Pilot Programmes provide the most direct pathway towards applied testing. Selected institutional processes can implement bounded Human–AI Cognitive Configurations and evaluate not merely productivity or technical performance but changes in decision quality, traceability, human expertise and institutional resilience. The architecture developed through HAICA allows pilot programmes to specify what they are actually testing.

H2 and J2 provide longer-term perspectives. Civilisational Governance Evolution may eventually examine how artificial cognition changes the historical development of institutional knowledge structures, while Future Institutional Architectures can explore organisations whose cognitive processes become substantially more hybrid than those observable today.

HAICA therefore contributes to the Operational Model Suite by establishing **the architectural object through which artificial intelligence can be studied as part of institutional cognition rather than as an external technological variable**. This creates a common basis for measurement, simulation and experimentation across later research streams.

46. Future Research Directions

The research agenda emerging from HAICA is extensive because the institutional incorporation of artificial intelligence remains rapidly evolving and comparatively under-observed at the level of complete cognitive architecture. One immediate priority concerns the empirical mapping of Human–AI Cognitive Configurations across different institutional domains. Research could compare public administration, regulation, scientific advisory systems, crisis management and policy analysis to determine whether recurring allocation patterns and interface structures emerge.

Meaningful Human Judgement deserves particular investigation. Future studies could examine whether different combinations of time, expertise, traceability and institutional authority affect the capacity of human reviewers to detect model errors or challenge artificial recommendations. Such research could move human oversight from a procedural principle towards an empirically grounded institutional concept.

Artificial Cognitive Deference and Institutional Cognitive Deskilling provide another major research trajectory. Longitudinal studies could examine whether sustained AI use changes human evaluative behaviour, whether disagreement rates decline for reasons independent of model improvement and which organisational practices preserve critical expertise most effectively.

Human–AI Cognitive Complementarity should likewise be investigated comparatively rather than assumed. Researchers can test whether parallel independent reasoning, sequential augmentation or iterative human–AI exchange produce different forms of epistemic performance across tasks. The objective would be to identify conditions under which cognitive diversity becomes genuinely productive.

Hybrid Cognitive Security requires increasing attention as artificial systems enter sensitive institutional processes. Research could examine how corrupted information propagates through hybrid configurations, whether human expertise detects manipulated outputs and which forms of redundancy provide meaningful protection against cognitive compromise.

Institutional Cognitive Autonomy and infrastructure dependency will become particularly important as foundation models and cloud-based systems become widely shared. Mapping dependencies across public institutions could reveal whether apparently decentralised governance systems are developing concentrated artificial cognitive infrastructures beneath their formal organisational diversity.

Finally, Adaptive Cognitive Allocation offers a longer-term research programme concerned with how institutions revise their architectures as capabilities change. Rather than treating AI deployment as a one-time implementation event, research could examine whether institutions develop procedures for reallocating cognitive functions according to evidence, risk and evolving expertise.

The future scientific task is therefore not simply to track what artificial intelligence becomes capable of doing, but to understand how institutions can continuously reorganise hybrid cognition as those capabilities change without losing judgement, resilience or legitimate control over their own cognitive architecture.

47. Scientific Contribution of the Human–AI Institutional Collaboration Architecture

The principal scientific contribution of HAICA is to establish **Hybrid Institutional Cognition as a distinct object of Institutional Cognition research**. Artificial intelligence is frequently evaluated through the performance of individual systems or through broad debates concerning automation and human oversight. HAICA shifts the analytical unit towards the complete architecture through which human and artificial capabilities become distributed, connected and institutionally consequential.

This move changes the central research question. The relevant issue is not simply what artificial intelligence can perform, but which cognitive functions are being redistributed, how those functions interact and whether the resulting configuration preserves the capacities necessary for reliable and legitimate institutional judgement. Cognitive Function Decomposition and Cognitive Function Allocation provide mechanisms for making that distribution explicit, while Human–AI Cognitive Configurations provide a practical object through which different arrangements can be compared.

A second contribution lies in replacing generic human-in-the-loop reasoning with **Meaningful Human Judgement**. Formal human presence does not guarantee substantive oversight when artificial systems structure evidence, representations and alternatives upstream. The distinction between Cognitive Authority and Decision Authority therefore makes visible transformations in epistemic influence that may occur without corresponding changes in organisational charts or legal responsibility.

A third contribution concerns the dynamic consequences of hybridisation. Artificial Cognitive Deference, Augmentation Overload, Institutional Cognitive Deskilling and Cognitive Dependency explain how initially successful augmentation can gradually alter human capability and institutional resilience. Hybrid Institutional Metacognition and Adaptive Cognitive Allocation provide corresponding mechanisms through which institutions can observe and deliberately modify those trajectories.

Finally, HAICA establishes an institutional rather than purely technological account of artificial intelligence. **The value of AI cannot be inferred from model capability alone because the same system can strengthen or weaken institutional cognition depending upon the architecture in which it is embedded.** Human–AI collaboration becomes a scientific problem of cognitive organisation, not merely a problem of technical performance or user interaction.

HAICA therefore extends Institutional Cognition into one of the defining governance questions of the emerging technological environment: how institutions can incorporate increasingly capable artificial cognitive systems while remaining capable of understanding, contesting and governing the transformations those systems introduce into the production of institutional knowledge itself.

48. Closing Perspective — Designing Institutions for Hybrid Cognition

The institutional adoption of artificial intelligence is often narrated as a sequence of increasingly capable tools entering organisations whose essential structure remains unchanged. From this perspective, the principal questions concern which tasks AI can perform, how much productivity it can generate and where human oversight should remain. The architecture developed throughout this Working Paper suggests that the transformation may be deeper.

When artificial systems begin participating in evidence retrieval, classification, synthesis, modelling, interpretation and scenario generation, they enter the processes through which institutions construct representations of the world. Their influence can therefore precede the formal moment of decision. An authorised human actor may continue signing the final document while the cognitive environment within which that choice became possible has been reorganised substantially.

This does not mean that institutions are destined to lose control over their cognition. Artificial capability can extend perception, reduce informational burden, increase analytical diversity and reveal patterns inaccessible to unaided human reasoning. Hybrid configurations may become more resilient and capable than exclusively human systems when artificial and human contributions remain sufficiently differentiated to compensate for one another's limitations.

The decisive question is architectural. Institutions must determine where artificial cognition enters, which functions it performs, how outputs become translated, where disagreement remains possible and whether human actors retain enough cognitive capacity to exercise genuine judgement. They must observe whether augmentation is creating dependency, whether expertise is being preserved and whether responsibility remains connected to the processes through which institutional knowledge is produced.

The future of institutional intelligence is therefore unlikely to consist of a simple movement from human cognition towards artificial cognition. It may instead involve increasingly heterogeneous systems in which cognitive functions migrate repeatedly among people, organisations and computational infrastructures as capabilities and environments change. The stability required for legitimate governance will need to coexist with the adaptability required for technological evolution.

HAICA provides one framework for making that transition visible and governable. Its purpose is not to prescribe a permanent boundary between humans and machines, but to ensure that the redistribution of cognition becomes an explicit object of institutional design, evaluation and learning rather than an unintended consequence of technological adoption.

The central transition described by HAICA is therefore not from human institutions towards artificial institutions, but from institutions that merely use artificial intelligence towards institutions capable of understanding and deliberately governing their own evolving hybrid cognition.

About the IDHUS Working Papers

This publication forms part of the constitutional publication ecosystem of the IDHUS Institute. It should not be interpreted as an isolated academic paper but as one component within a progressively expanding scientific architecture dedicated to the development of the emerging discipline of **Institutional Cognition**.

Within the organisational architecture of the IDHUS Institute, the Working Paper occupies a distinctive position. It comprises the elementary cognitive module through which operational research is conducted. At the structural level, however, the Working Paper also functions as the smallest autonomous organisational unit capable of generating identifiable scientific contributions within the broader research ecosystem. For the way we conduct our research, it represents the point at which constitutional concepts, Research Series, Research Lines, methodologies, computational developments and empirical investigations converge into a coherent research product capable of contributing directly to the cumulative evolution of institutional knowledge.

Series A — Foundations of Institutional Cognition

Series A preserves the constitutional foundations of the entire research ecosystem. Rather than constructing new constitutional theories, it continuously refines, clarifies and strengthens the conceptual architecture established throughout *The IDHUS Method*, the *Research Validation Papers* and the *Research Atlas*. It provides the common scientific language that allows every other Research Series to remain conceptually coherent while supporting the long-term evolution of Institutional Cognition.

Research Lines

A1 — Constitutional Evolution of Institutional Cognition
A2 — Conceptual Architecture and Theory Integration
A3 — Epistemology of Institutional Intelligence
A4 — Philosophy of Adaptive Governance
A5 — Scientific Foundations of Planetary Cognitive Civilisation

Series B — Cognitive Architectures

Series B investigates how cognition emerges and operates across multiple organisational scales, from individuals and institutions to artificial intelligence systems and distributed governance environments. Research within this Series explores organisational learning, collective intelligence, institutional memory, cognitive networks and adaptive decision-making, providing many of the theoretical and computational models that support the broader research programme.

Research Lines

B1 — Individual and Collective Cognition
B2 — Institutional Cognitive Architectures
B3 — Distributed Intelligence Systems

B4 — Human-AI Cognitive Collaboration
B5 — Organisational Learning Architectures

Series C — Computational Governance

Series C examines how artificial intelligence, computational modelling and digital infrastructures are transforming governance and public administration. Its objective is to understand how computational systems become active components of institutional cognition through intelligent decision-support, AI-native public administration, digital governance architectures and adaptive regulatory systems capable of responding to increasing societal complexity.

Research Lines

C1 — AI-Augmented Public Administration
C2 — Computational Policy Design
C3 — Intelligent Decision-Support Systems
C4 — Algorithmic Governance Architectures
C5 — AI-Native Institutional Ecosystems

Series D — Planetary Systems

Series D expands the scope of research from individual institutions towards planetary-scale systems of governance and coordination. Drawing upon systems thinking, complexity science and computational simulation, it investigates global governance networks, civilisational dynamics, resilience, distributed coordination and the cognitive infrastructures required to understand increasingly interconnected planetary systems.

Research Lines

D1 — Global Cognitive Systems
D2 — Planetary Governance Networks
D3 — Civilisational Systems Dynamics
D4 — Global Coordination Mechanisms
D5 — Planetary Resilience Architectures

Series E — AI and Institutional Intelligence

Series E explores the relationship between human intelligence, institutional cognition and artificial intelligence. Rather than studying AI as an isolated technology, it investigates how hybrid cognitive systems emerge through the interaction of people, organisations and intelligent computational agents. This Series provides the principal research environment for understanding the future evolution of AI-enabled institutions and distributed governance.

Research Lines

E1 — Institutional Intelligence
E2 — Human-AI Collaboration
E3 — AI Governance Systems
E4 — Autonomous Institutional Agents
E5 — Constitutional AI for Public Governance

Series F — Cognitive Metrics and Measurement

Series F develops the methodological foundations required to observe and evaluate Institutional Cognition. It designs metrics, indicators, assessment frameworks and measurement methodologies capable of transforming constitutional concepts into operational research variables. These tools support empirical research, comparative analysis, validation programmes and evidence-based governance across the entire research ecosystem.

Research Lines

F1 — Institutional Intelligence Metrics
F2 — Governance Performance Indicators
F3 — Cognitive Capacity Assessment
F4 — Adaptive Systems Measurement
F5 — Planetary Governance Metrics

Series G — Simulation and Digital Twins

Series G establishes the computational laboratories of the Institute. Through simulation environments, digital twins, systems dynamics and agent-based modelling, it creates experimental platforms for exploring institutional behaviour under controlled conditions. These environments allow researchers to investigate alternative governance architectures, analyse complex adaptive systems and evaluate future scenarios before they emerge in practice.

Research Lines

G1 — Institutional Digital Twins
G2 — Governance Simulation Platforms
G3 — Multi-Agent Governance Models
G4 — Scenario Generation and Strategic Foresight
G5 — Computational Experimentation

Series H — Comparative Civilisational Studies

Series H investigates how different societies, institutions and historical periods have organised governance, knowledge and collective intelligence. By comparing diverse civilisational experiences, governance models and institutional traditions, it seeks to identify recurring cognitive principles while understanding the contextual diversity through which Institutional Cognition evolves across cultures and historical epochs.

Research Lines

H1 — Comparative Institutional Cognition
H2 — Civilisational Governance Evolution
H3 — Cross-Cultural Adaptive Systems
H4 — Historical Institutional Intelligence
H5 — Comparative Planetary Governance

Series I — Applied Governance Laboratories

Series I serves as the operational bridge between scientific research and institutional practice. It develops collaborative research environments in which conceptual models, computational systems and governance methodologies can be explored through pilot projects, organisational experimentation, innovation programmes and long-term partnerships with public institutions. Practical experience generated through these laboratories continuously feeds back into the broader research programme.

Research Lines

I1 — Public Sector Innovation Laboratories
I2 — AI Governance Pilot Programmes
I3 — Institutional Transformation Frameworks
I4 — Governance Prototyping and Testing
I5 — Living Governance Laboratories

Series J — Future Research Frontiers

Series J ensures that the research ecosystem remains permanently open to future scientific developments. It explores emerging technologies, new governance paradigms, evolving AI capabilities and other frontier topics whose long-term significance may not yet be fully understood. By institutionalising organised exploration, this Series enables the IDHUS Institute to adapt continuously to future scientific opportunities while preserving the constitutional coherence of its broader research architecture.

Research Lines

J1 — Emerging AI Paradigms
J2 — Future Institutional Architectures
J3 — Post-Governance Systems
J4 — Civilisational Foresight
J5 — Unknown Research Horizons

Internal Constitutional References

The present Working Paper forms part of the constitutional and operational architecture of the IDHUS Institute. It should be read in conjunction with the following foundational publications, which collectively establish the constitutional framework within which the Working Papers Programme develops.

Constitutional Framework

- **The IDHUS Canon**
- **The IDHUS Method**

Programme I Research Validation Papers

- **RVP-001 Operational Cognitive Architecture**
- **RVP-002 Operational Consciousness**
- **RVP-003 Collective Cognition**
- **RVP-004 Institutional Cognition**
- **RVP-005 Governance Intelligence**
- **RVP-006 Adaptive Public Administration**
- **RVP-007 AI-Augmented Public Administration**
- **RVP-008 Adaptive Public Policy**
- **RVP-009 Cognitive Public Services**
- **RVP-010 Self-Evolving Governmental Systems**

Research Atlas

- **Volume 0 Research Ecosystem Architecture**
- **Volume 1 Research Architecture**
- **Volume 2 Planetary Research Ecosystems**
- **Volume 3 Planetary Cognitive Systems**
- **Volume 4 Planetary Governance Architecture**
- **Volume 5 Civilisational Evolution**
- **Volume 6 Planetary Cognitive Civilisation**

Operational Research Programme

- **Operational Research Programme- Working Papers Framework**

These publications together establish the constitutional foundations upon which the operational research developed throughout the IDHUS Working Papers Programme is constructed.

Glossary

Adaptive Cognitive Allocation

The deliberate reassessment and redistribution of cognitive functions among human, artificial and hybrid components in response to changing technological capabilities, institutional expertise, task characteristics, accumulated evidence or risk conditions. Adaptive Cognitive Allocation treats the distribution of cognition as a revisable institutional design decision rather than a permanent progression from human towards artificial performance.

Alternative Cognitive Pathway

A sufficiently independent human, artificial or hybrid route through which an essential institutional cognitive function can be performed, evaluated or reconstructed when the primary pathway becomes unavailable, unreliable or contested. Alternative Cognitive Pathways contribute to Human–AI Cognitive Resilience by providing functional diversity rather than simple duplication.

Artificial Cognitive Deference

A condition in which human actors or institutional routines increasingly privilege artificial outputs within judgement because those outputs are perceived as more objective, comprehensive, authoritative or analytically capable. Artificial Cognitive Deference differs from Cognitive Dependency because it concerns patterns of epistemic reliance and evaluative behaviour rather than the structural ability of an institution to operate without the artificial component.

Artificial Cognitive Participation

The functional contribution of computational systems to processes that perform institutionally relevant cognitive functions such as perception, retrieval, classification, modelling, synthesis, prediction or scenario generation. Within HAICA, the term *cognitive* is used functionally rather than phenomenologically and does not imply consciousness, subjective understanding or equivalence between artificial and human cognition.

Augmentation Overload

A condition in which the volume, velocity or complexity of artificial cognitive outputs exceeds the interpretative capacity of the human actors or wider institutional configuration expected to evaluate them. Augmentation Overload can reduce meaningful review, encourage cognitive shortcuts and increase Artificial Cognitive Deference even when individual artificial outputs remain technically useful.

Cognitive Authority

The capacity of an actor, system or representation to shape the evidence, interpretations, alternatives and problem representations through which institutional decisions become understood. Cognitive Authority is analytically distinct from Decision Authority and can migrate towards artificial systems without any corresponding transfer of formal legal or administrative power.

Cognitive Dependency

A structural condition in which an institution can no longer adequately perform, evaluate, reconstruct or readily substitute an essential cognitive function without continued access to a particular artificial system or infrastructure. Cognitive Dependency differs from Artificial Cognitive Deference because dependency concerns institutional capability rather than the degree of trust or epistemic privilege granted to artificial outputs.

Cognitive Function Allocation

The architectural distribution of identified institutional cognitive functions among human actors, artificial systems or hybrid arrangements according to their epistemic requirements, contextual characteristics, risks, reversibility, available capabilities and institutional constraints. Cognitive Function Allocation asks where cognition should be located rather than simply which existing human activities can be automated.

Cognitive Function Decomposition

The systematic identification of the distinct cognitive operations embedded within an institutional task before decisions concerning automation or augmentation are made. Cognitive Function Decomposition makes visible activities such as retrieval, classification, interpretation, comparison, modelling, contextualisation and normative evaluation that may otherwise remain concealed within broad organisational categories such as analysis or decision-making.

Decision Authority

The legitimate institutional power to authorise, reject, modify or assume responsibility for a decision. Decision Authority is distinguished within HAICA from Cognitive Authority because formal control over a decision can remain human even when artificial systems exercise substantial influence over the representations upon which that decision depends.

Hybrid Cognitive Interface

A technological, procedural, informational or organisational mechanism through which human and artificial cognitive contributions become connected within a Human–AI Cognitive Configuration. Hybrid Cognitive Interfaces include not only visual user interfaces but also prompts, structured inputs, data pipelines, output formats, uncertainty representations, review procedures, escalation mechanisms and documentation standards.

Hybrid Cognitive Reconfiguration

The broader evolution of relationships among human actors, artificial systems, interfaces, information infrastructures and organisational procedures over time. Hybrid Cognitive Reconfiguration may result from deliberate redesign or emerge incrementally through repeated changes in Cognitive Function Allocation, technological capability and institutional practice.

Hybrid Cognitive Security

The protection of hybrid institutional cognitive processes against deliberate or accidental corruption of the evidence, models, representations and informational relationships through which institutional knowledge and judgement are produced. Hybrid Cognitive Security extends conventional cybersecurity by treating the integrity and provenance of cognition, rather than technical availability alone, as an object of protection.

Hybrid Epistemic Traceability

The institutional capacity to reconstruct how evidence, artificial systems, model versions, computational transformations, human interpretations and organisational procedures contributed to the representations underlying an institutional judgement or decision. Hybrid Epistemic Traceability supports accountability, error analysis and institutional learning across the complete Human–AI Cognitive Configuration.

Hybrid Institutional Cognition

The distributed performance of institutional cognitive functions through configurations combining human actors, artificial cognitive systems, organisational procedures and information infrastructures. Hybrid Institutional Cognition refers to functional integration within connected cognitive processes rather than merely the coexistence of humans and AI within the same organisation.

Hybrid Institutional Fragility

The vulnerability of a hybrid institutional cognitive architecture to disruption, degradation or loss of critical human, artificial or infrastructural components. Hybrid Institutional Fragility may arise through concentrated dependency, inadequate redundancy, loss of expertise or insufficient alternative cognitive pathways and should be distinguished from Institutional Cognitive Deskilling, which represents one possible mechanism through which fragility can develop.

Hybrid Institutional Learning

The process through which evidence concerning the performance and consequences of human–AI cognitive configurations becomes incorporated into institutional knowledge, procedures, interpretation practices or architectural design. Hybrid Institutional Learning is broader than machine learning because the relevant adaptation may occur within human behaviour, organisational processes, artificial systems or the relationships among them.

Hybrid Institutional Metacognition

The institutional capacity to construct and revise structured representations of its own human–AI cognitive architecture, including the allocation of cognitive functions, patterns of artificial influence, human reliance, traceability, expertise, dependency and resilience. Hybrid Institutional Metacognition makes the architecture of hybrid cognition itself an object of institutional knowledge and supports deliberate adaptation.

Human–AI Cognitive Complementarity

An architectural condition in which differentiated human and artificial cognitive capabilities are organised so that their contributions compensate for limitations elsewhere within the configuration, improving the performance of the wider institutional cognitive process. Complementarity is treated within HAICA as an outcome to be demonstrated rather than an automatic consequence of combining humans and AI.

Human–AI Cognitive Configuration (HACC)

A task-specific arrangement of human actors, artificial cognitive systems, information resources, organisational procedures and Hybrid Cognitive Interfaces through which one or more institutional cognitive functions are performed. HACC represents the principal operational unit

through which HAICA can reconstruct, compare and evaluate realised or proposed forms of hybrid institutional cognition.

Human–AI Cognitive Mapping

The empirical reconstruction of cognitive functions, human actors, artificial systems, interfaces and knowledge flows within a Human–AI Cognitive Configuration. Human–AI Cognitive Mapping seeks to identify where artificial participation becomes consequential, how representations are transformed and whether the realised distribution of cognitive influence corresponds to formal institutional descriptions of authority and responsibility.

Human–AI Cognitive Resilience

The capacity of a hybrid institutional cognitive architecture to maintain, recover or appropriately reconfigure essential cognitive functions when human, artificial or infrastructural components become unavailable, unreliable or uncertain. Human–AI Cognitive Resilience depends upon appropriately organised redundancy, cognitive diversity, alternative pathways, residual expertise and recovery mechanisms rather than the reliability of individual components alone.

Human–AI Cognitive Translation

The bidirectional processes through which representations generated by human and artificial cognitive systems become sufficiently interpretable to participate within shared institutional reasoning. Human–AI Cognitive Translation seeks to preserve relevant meaning, assumptions, context and uncertainty without requiring human and artificial representations to become identical.

Human–AI Feedback Architecture

The mechanisms through which evidence concerning institutional outcomes, artificial-system performance, human interpretation, disagreement, overrides, errors and near-misses returns to the cognitive configuration in forms capable of supporting evaluation and adaptation. Human–AI Feedback Architecture provides the evidential foundation for Hybrid Institutional Learning.

Human–AI Institutional Collaboration Architecture (HAICA)

The operational research framework developed in this Working Paper for analysing how human and artificial cognitive capabilities become distributed, connected, coordinated and adaptively reorganised within institutions. HAICA examines hybrid cognition at the architectural level, while individual Human–AI Cognitive Configurations represent particular task-specific instantiations of that architecture.

Institutional Cognitive Autonomy

The capacity of an institution to understand, evaluate and, where necessary, reconfigure the human and artificial cognitive systems upon which its judgement depends. Institutional Cognitive Autonomy does not require technological self-sufficiency; it requires dependencies to remain sufficiently visible, contestable and reversible for meaningful institutional control over cognitive architecture to persist.

Institutional Cognitive Deskilling

A longitudinal reduction in human cognitive capability associated with sustained delegation of particular cognitive functions to artificial systems. Institutional Cognitive Deskilling becomes

institutionally significant when the declining capability remains necessary for evaluating artificial outputs, preserving independent judgement, responding to system failure or maintaining cognitive resilience.

Meaningful Human Judgement (MHJ)

A condition in which authorised human actors possess sufficient cognitive access, contextual understanding, relevant expertise, practical time, alternative information and institutional authority to understand, question and alter artificial contributions when doing so is consequential. Meaningful Human Judgement is distinguished from procedural human presence or nominal human-in-the-loop arrangements.

Nominal AI Adoption

The formal or documented presence of artificial intelligence within an institution without necessarily indicating substantial artificial participation in the cognitive processes through which institutional understanding or judgement is produced. HAICA distinguishes Nominal AI Adoption from realised Hybrid Institutional Cognition by examining the actual location and influence of artificial cognitive functions.

Realised Hybrid Cognition

The empirically observable distribution and interaction of human and artificial cognitive contributions within an institutional process, irrespective of how the organisation formally describes its use of AI. Realised Hybrid Cognition is reconstructed through Human–AI Cognitive Mapping and process-level analysis.

Residual Human Cognitive Capability

Human expertise intentionally or incidentally preserved within a hybrid institution after substantial artificial augmentation of a cognitive function. Residual Human Cognitive Capability becomes particularly important where independent evaluation, recovery from artificial-system failure, architectural reversibility or preservation of Institutional Cognitive Autonomy requires humans to retain meaningful understanding of delegated functions.

Structured Cognitive Disagreement

An architectural condition in which divergent human, artificial or multi-model interpretations are preserved long enough to support comparison, investigation and error detection before institutional coordination requires convergence. Structured Cognitive Disagreement treats cognitive diversity as a potential source of complementarity and resilience rather than assuming that agreement itself indicates superior institutional cognition.

Vendor and Infrastructure Dependency

A condition in which institutionally significant cognitive functions depend upon externally controlled models, computational platforms, data services or technological infrastructures. Such dependency becomes a HAICA concern when external changes, failures or constraints can materially alter institutional cognition or reduce the institution's capacity to evaluate and reconfigure its own cognitive architecture..


IDHUS
Institute for the Development
of Human Societies