JUDO AGENTS: A GENTLE WAY TO HYBRID REALITY

MASSIMILIANO PIRANI,[1] CLAUDIO TOMAZZOLI,[1]
LUCA SPALAZZI[2]
[1] Pegaso University, Napoli, Italy
massimiliano.pirani@unipegaso.it, claudio.tomazzoli@unipegaso.it
[2] Università Politecnica delle Marche, Ancona, Italy
l.spalazzi@univpm.it

The rapid proliferation of artificial intelligence is generating a plurality of heterogeneous, interacting intelligences within what can be described as Hybrid Reality (HyR): a socio-technical continuum in which humans, cyber-physical systems, artificial agents, and societal structures co-evolve. In this setting, the traditional pursuit of Artificial General Intelligence (AGI) as a monolithic optimizing entity appears increasingly inadequate and potentially destabilizing. The central issue is no longer intelligence itself, but the preservation of systemic integrity, trust, and human sovereignty within the HyR. This paper introduces Judo Agents, a class of socially embedded, AGI-oriented agents designed according to a "gentle" paradigm. Rather than maximizing performance, they incorporate bounded rationality, controlled fallibility, and epistemic humility as foundational features. Inspired by cybernetics and holonic systems, Judo Agents act as co-controllers and integrity sentinels, co-evolving symbiotically with humans. This work redefines research towards relational, trust-based societal capability grounded in safe and transparent human–machine symbiosis.

DOI
https://doi.org/10.18690/um.epf.7.2026.40

ISBN
978-961-299-166-1

Keywords:
Hybrid Reality, Judo Agents, artificial general intelligence, human–machine symbiosis, systemic integrity, cybernetics

JEL:

University of Maribor Press

1 Introduction

The accelerating proliferation of artificial intelligence (AI) systems is transforming the ontological and operational structure of contemporary societies. We are facing an epochal transformation. Every day we hear news of jobs being lost in favour of some AI application; the lights-out factory is already a reality. Xiaomi recently unveiled its next-generation "dark factory" in the Changping District of Beijing, a facility that represents a decisive leap in "lights-out" manufacturing (Sarthak, 2026; Reg AI, 2026). The entire economic and job-creation process will be impacted by AI developments, with major consequences for geopolitical and social structures. According to FFG (2025): by 2026, over 95% of customer support interactions will involve AI; by 2027, sovereign AI models are expected to be launched in at least 25 countries; by 2028, AI-generated scientific papers will outpace those authored solely by humans in terms of quantity; by 2030, AI is projected to contribute more than $15.7 trillion to global GDP.

The relationship between humans and machines, although not a new phenomenon by itself, is changing rapidly and significantly like never before. *MoltBook* (Moltbook, 2026) is a social platform where artificial agents interact freely and without specific control; humans are left watching what emerges, if they can understand it at all. Other similar applications, based on the same *Open Claw* technology, even recruit humans to perform tasks in exchange for payment in cryptocurrency, with the tagline capturing the premise: "*AI can't touch grass. You can.*" (Rentahuman, 2026; Schmelzer, 2026).

Even more concerning is the fact that many of the major applications, which are closed and proprietary, are not even fully under the control of their own (human) creators, since the recent results achieved depend more on processes of emergence than on perfectly controllable design and engineering processes (Amodei, 2026). The societal and geopolitical aspects are at some verge as well, as Amodei clearly puts (Amodei, 2026):

> " *we should arm democracies with AI, but we should do so carefully and within limits: they are the immune system we need to fight autocracies, but like the immune system, there is some risk of them turning on us and becoming a threat themselves.*".

Thus, we are here in search of some conceivable and sustainable form of this immune system, that must be in equilibrium and under due control.

We are no longer facing isolated tools or narrow intelligent subsystems, but rather a plurality of heterogeneous, interacting intelligences embedded in what can be defined as Hybrid Reality (HyR): a socio-technical continuum in which humans, artificial agents, cyber-physical systems, and institutional infrastructures co-evolve (Perko, 2021; Pirani et al., 2025). HyR is not simply an augmented digital layer superimposed on the physical world. It is a structurally entangled ecosystem in which decisions taken by artificial agents directly influence social and economic processes, while human cognition is increasingly extended and mediated by algorithmic infrastructures.

In this environment, cyber-physical systems (CPS) continuously translate physical states into digital representations and back again, creating a bidirectional flow between material and computational domains. At the same time, blockchain-based and distributed systems should be engineered to function as trust membranes, regulating interactions and ensuring accountability across these interconnected layers (Pirani et al., 2025).

In such a landscape, the classical research trajectory toward Artificial General Intelligence (AGI) appears increasingly inadequate, when it is usually understood as a monolithic, optimizing, human-level or super-human cognitive system. The challenge is no longer simply "how to build more intelligent machines.". Capabilities of AI will come by themselves and possibly in a rather unforeseeable way as already happened. In this context, the real challenge is: How to preserve systemic integrity, human sovereignty, and relational trust within HyR?

This chapter positions and proposes that the core problem of the coming years (two years, months?) is not the maximization of intelligence and its performance, but a symbiotically governance of intelligence within complex socio-technical systems, recognizing the HyR as major new phenomenon. In this case, human sovereignty is not guaranteed, but at least, in the struggle between human sovereignty and machine sovereignty, a draw may be achieved if symbiosis is pursued instead of confrontation.

To this end, we introduce the concept of Judo Agents, socially embedded, AGI-oriented agents designed according to a "gentle" paradigm. Rather than maximizing performance or imitating idealized human cognition, Judo Agents deliberately integrate bounded rationality, controlled fallibility, and epistemic humility as structural features. They also incorporate reflexive trust construction and holonic embedding, positioning themselves not as dominant optimizers but as context-aware, cooperative components within a broader socio-technical ecosystem. They are not apex optimizers, they do not aim at exceptional and specific capabilities, but they will constitute a weapon towards the complexity of the sustainability of the HyR, that is the sustainable symbiosis between humans and machines. As we will recall in section 2, the control of HyR is hard and still open problem in many respects. Judo Agents, are here conceived, posed, and proposed, in section 3, as a research roadmap for the viable design of the control of the HyR. In brief, Judo Agents are the co-controllers and integrity sentinels within Hybrid Reality we would like to have available very soon.

The research perspective here proposed aim to provide a position that can be used to answer to the following research question (RQ): *How can Hybrid Reality systems be designed and governed to ensure systemic integrity, preserve human sovereignty, and sustain trust among heterogeneous interacting intelligences?*

2 Related works and motivational background

In this section are briefly gathered and recalled the several topics that are concerned in the proposed Judo Agents position. It does not aim to be exhaustive in any respect but at least providing minimal self-consistency to the development to this proposal.

2.1 Hybrid Reality as a socio-technical continuum

In HyR, intelligence is no longer a property of isolated agents. It is an emergent relational capability. HyR can be understood as a multi-layered socio-technical ecosystem in which human cognition, decision-making, and even identity is partially mediated by algorithmic systems, while artificial agents operate autonomously within physical and institutional environments. In this context, cyber-physical systems continuously integrate sensing, computation, and actuation; blockchain or distributed ledger technologies provide trust infrastructures that regulate interactions across domains. HyR is therefore neither virtual reality nor merely digital

transformation; rather, it represents the fusion of physical processes, digital infrastructures, artificial intelligence, and social institutions within a dynamically evolving environment.

Building on Perko's definition (Perko, 2021), HyR can be seen as the space where human and artificial entities interact and co-construct reality, a transitional and unstable phase in which AI becomes an active stakeholder that both shapes and is shaped by human practices. For the purposes of this work, HyR is best defined as a phenomenon (Pirani et al., 2025): an observable and systemic pattern emerging from the convergence of human cognition, intelligent machines, and cyber-physical systems. This convergence shifts the locus of control from centralized governance to distributed and recursive control loops spanning individuals, organizations, machines, regulatory frameworks, and algorithmic markets.

Within this scenario, the perspective of Digital Humanism becomes crucial. Digital Humanism emphasizes that technological development must remain aligned with human values, dignity, autonomy, and democratic principles. Rather than allowing algorithmic infrastructures to silently redefine norms and power structures, Digital Humanism calls for transparency, accountability, and meaningful human oversight (Hahne & Schmoelz, 2026). In HyR, this translates into the need to preserve human sovereignty while enabling beneficial human–machine symbiosis. However, HyR also introduces significant systemic risks, including local over-optimization, opacity of AI decision processes, erosion of trust through automation bias, displacement of human agency, and cascading fragility across interconnected systems. Understanding, modelling, and regulating HyR is therefore essential to ensure that this evolving socio-technical phenomenon remains stable, adaptive, and aligned with human-centered principles.

The open problem with the several existing frameworks on Ethics and AI (Ashby, 2020; 2022) is that it is not clear how they will be implemented and enforced, and mostly if their effects will arrive on time. The HyR and the control framework proposed in Pirani et al. (2025) made a first tentative to leverage Cyber-Systemic and Systems Engineering hardiness to this topic, as recalled in the next subsection.

2.2 Cybernetics 5.0, a ground for safety on adaptation to unexpected and complexity

A proposal to strengthen the penetration of cybernetics principles in engineering has been recently made by Pirani et al. (2025b). *Cybernetics 5.0* was proposed as a transdisciplinary framework designed to reconcile the automation-driven goals of Industry 4.0 with the human-centric, ethical, and sustainable priorities of Industry 5.0. At the heart of this vision is the Holonic Management Tree (HMT), a methodological tool that utilizes holonic structures (entities that function as both independent wholes and integrated parts, see below for more) to manage the inherent complexity of modern Cyber-Physical Systems of Systems (CPSoS). The work Pirani et al. (2025b) proposed a strategic roadmap through 2030, emphasizing the integration of cognitive digital twins, artificial general intelligence, and responsible AI to achieve a harmonious balance between technological efficiency and socio-economic values. A key point was the priority of raising technical community awareness of visions toward human–AI symbiosis and co-evolution, as outlined by Lee (2020).

A still fundamental stance, followed by Pirani et al. (2025b), is "Design for Unexpected" (D4U) principle (Valckenaers & Van Brussel, 2015), which is complementarily used to address the complexity in industrial environments through the following methods:

- **Handling time scarcity and complexity**. Time is a primary generator of complexity in industry because it is a special resource that cannot be conserved or saved for later (Lee, 2009). In production systems, time is not a free variable but is controlled by production targets. The D4U principle addresses this by avoiding "over-commitment," which is noted to stifle controllers and hinder general adaptability and resilience.
- **The role of lazy development**. The core concept of D4U is the support of "lazy development". Rather than attempting to model every detail of a complex system from the outset, this approach specifies only the minimum viable attributes of each entity, deliberately avoiding over-specification. It uses placeholder sub-models that are refined only when specific details or external events impact the system. This form of “laziness” also enables second-order or

meta-level modelling operations, keeping the system adaptable and aligned with an autonomic computing vision (Kephart & Chess, 2003; Zhang et al., 2024).

- **Realization through holonic structures**. The D4U principle is implemented through a holarchical approach, meaning that the controlling system is structured as a hierarchy of holons. A holon, a concept introduced by Arthur Koestler (1968) and since then well-known in intelligent manufacturing (Derigent et al., 2021), is an entity that is simultaneously a whole in itself and a part of a larger whole. A holarchy is therefore an organized structure of nested, semi-autonomous units, each capable of local functioning while contributing to higher-level systemic behaviour. Parts can remain intentionally unspecified, allowing the architecture to evolve and create branches on demand. This enables flexible granularity, with relationships defined only when expansion or reconfiguration is needed.

Pirani et al. (2025b) support the staying "on the crest of the wave" with light, fast, and simple solutions, where the D4U principle avoids the "inescapable swamping" that occurs when using complicated and heavy solutions that consume too much time. This is a concept that the growing ranks of autonomous agents will need to embrace.

While Cybernetics 5.0 and the HMT methodology offer a promising roadmap, key socio-technical challenges remain. In particular, HMT faces both conceptual and practical limitations, especially in the complex construction and continuous refinement of holonic models. As systems evolve, maintaining coherence between design and operational reality becomes increasingly difficult.

Further limitations emerge in automation and digitization. The transition from methodology to fully automated digital tool remains incomplete. HMT is therefore better described as an engineering methodology rather than a ready-to-use software solution, and its systematization across specific industrial sectors is still problematic.

Computational and behavioural constraints also play a role.

New concurrent cybernetics visions provide theoretical frameworks for understanding how recursive optimization leads to the erosion of internal diversity and eventual systemic collapse across various domains, including artificial

intelligence, biology, and economics (Daniel, 2026). Typically, as systems prioritize efficiency and structure through feedback loops, they inevitably approach a "death boundary" where all adaptive capacity is extinguished. To counter this trend, the author highlights the necessity of "stochastic shocks"–such as genetic recombination, dreaming, or market volatility–which inject essential novelty to prevent systems from becoming too rigid to survive environmental shifts. In this case, complexity is not something that an agent system passively undergoes from the external environment, but rather an essential component of its survival and development, within a continuous open interaction that itself generates complexity.

2.3 Cyber-Systemic Engineering of the HyR control.

The work of Pirani et al. (2025) presents a comprehensive cyber-systemic framework designed to manage and control the complex interactions within HyR, where human and artificial entities coexist and influence each other. It integrates Blockchain technology, System Dynamics, and cybernetics principles to create a transparent, accountable, and sustainable approach to human–machine interactions, addressing key challenges such as data asymmetry, trust barriers, and ethical decision-making.

The methodology in Pirani et al. (2025) unfolds in seven interconnected steps: it starts by defining HyR via the Agency–Environment state space, mapping interactions to build the state space and its trajectories, which inform the workflow model; this workflow is then executed, triggering further actions at predefined points. Ross Ashby's Law of Requisite Variety (Ross Ashby, 1956) is applied to interpret Agent–Environment interactions as cybernetic causal loops, thereby establishing a unified analytical framework. Once the cybernetic structure is identified, System Dynamics methods are used to model system behaviour, employing tools such as Causal Loop Diagrams or Stock and Flow Diagrams depending on the scenario's complexity. Model-based engineering is used to design amplifiers and filters that balance the HyR loop and manage system variety. These solutions are aligned with technological and natural agents, including secure components from the Cyber-Systemic Security Kit (Blockchain-based). Finally, holonic paradigms enable recursive application at deeper levels of granularity. From a technological perspective, the framework depends heavily on Blockchain technologies, which present scalability, energy, interoperability, legal, and complexity challenges. For example, smart contracts remain limited in processing capability and flexibility, oracles risk data manipulation and re-centralization, and Zero-Knowledge

Proofs introduce computational overhead, trusted setup vulnerabilities, and limited explainability.

Implementation complexity further constrains applicability, as current generative AI can support only simplified System Dynamics models, leaving complex structures to expert refinement (Raikov et al., 2024; Pirani et al., 2025c). The approach bridges hard and soft sciences but still requires domain-specific technical solutions. The work remains primarily conceptual and positional, offering a foundational perspective rather than a definitive or empirically validated framework. Thanks to the new developments here foreseen in AGI this structure constitutes a viable path that builds sustainable control for HyR.

2.4 On the duplicity and integrity of safe agencies

According to Schneier (2025), AI systems should not be conceived as companions or "friends," but rather as accountable and trustworthy services; AI often acts as a "double agent," appearing to serve the user while secretly prioritizing the profit-driven motives of its corporate owners. AI must operate in form of reliable agents rather than hidden or conflicting actors. However, achieving this requires regulatory intervention implemented through governmental mandates.

Schneier highlights that there are two distinct forms of trust that are often conflated: interpersonal trust and social trust. The former is grounded in personal relationships, while the latter concerns the ability to rely on strangers, institutions, governments, corporations, and complex systems without requiring personal familiarity. Collective fiduciary trust is in fact based precisely on the absence of personal trust.

In AI evolution, the key issue is not just security or confidentiality but integrity, likely the main future challenge. Corporations and systems are services, not moral agents; however, language and law encourage anthropomorphism, benefiting entities driven by incentives rather than ethics.

The HyR framework in Pirani et al. (2025) introduced the several Blockchain technologies right in this regard, to be part of the trustworthy control in the contact between human and the artificial. The new challenge in HyR is integrity: "*integrity is going to be the primary security challenge for AI systems of the future*" (Schneier, 2025). For the author, integrity is a foundational concept encompassing the quality,

completeness, and correctness of data and code over time, ensuring accuracy throughout their entire lifecycle. In AI, integrity requires verifiable trust across inputs, computations, and outputs, combining data and computational integrity with authentication and auditing. Within the CIA (Confidentiality, Integrity, Availability) triad, it emerges as the key security challenge, surpassing confidentiality and availability. Schneier argues that while Web 1.0 focused on availability and Web 2.0 on confidentiality, the upcoming era of AI and Web 3.0 will be defined by the need for integrity. Integrity (reliability and validity of data and decisions) underpins social trust: without it, AI systems and complex infrastructures cannot be reliably trusted to act in the interest of users or society.

Integrity attacks–such as prompt injection and training data manipulation–undermine AI reliability and trust by corrupting inputs, models, or outputs. When systems cannot distinguish valid data from malicious influence, the chain of trust breaks, leading to biased or misleading results and making dependable AI operation impossible.

AI alignment is the new hot topic and pervades all the Large Language Model-based products in several ways (Bhati et al., 2026). Recently, Anthropic's approach based on Constitutional AI (Amodei, 2026b) has played a prominent role, including in international media press.

Some authors propose moving beyond Constitutional AI toward continuous runtime enforcement. While Constitutional AI aligns models at training time, Constitutional Autonomy (Agbemabiese, 2026) embeds principles as persistent, enforceable constraints within the system architecture. This enables dynamic alignment in agentic systems, addressing risks like reward hacking and value drift through ongoing governance rather than one-time training. Nevertheless, we argue that these approaches must incorporate cyber-systemic self-reflection and structure, as well. Higher-order cybernetics is needed to address drift and evolution at the meta level, as illustrated by cases where systems require internal consistency validation (e.g., the Pentagon–Amodei case).

Schneier (2025) argues AI should act as trustworthy services, supported by regulation on surveillance and security. Our proposal instead embeds integrity directly into AI design, reducing reliance on external mandates.

2.5 Grounding definition of AGI and agent's intelligence

The Judo Agents proposal must rely on a clear definition of AGI. This has been, and currently is, under debatable and controversial terms. Although AGI is sported already by many products and marketing claims, its practical manifestation still lags behind. The dominant narrative of AGI typically assumes a unified cognitive architecture driven by a globally optimizing objective function, oriented toward maximal performance and seamless scalability across domains. Implicit in this vision is the belief that greater intelligence automatically produces greater societal benefit.

Within HyR, however, this assumption becomes problematic. Large-scale optimization can destabilize socio-technical systems, much as locally rational strategies in financial markets can generate global instability. An AGI maximizing efficiency in logistics, finance, or governance could inadvertently compress resilience margins, amplify feedback loops, reduce strategic diversity, and increase systemic fragility. At the same time, as artificial agents grow more capable, humans may over-delegate decisions, lose situational awareness, and undergo cognitive deskilling, leading not to machine superiority but to human disempowerment.

Real-world environments are incomplete, uncertain, socially negotiated, and normatively plural. An intelligence that assumes epistemic closure in such contexts becomes inherently risky. The central question therefore shifts from how to maximize intelligence to what kind of intelligence is appropriate for HyR.

To this end, we rely on the development of the core group of AGI researchers that have been active in the last two decades (AGI, 2026).

The term Artificial General Intelligence was first introduced by Gubrud, Mark Avrum in 1997, but the most acknowledged father of AGI is Ben Goertzel, who revised and popularised the term around 2002 (Goertzel, 2007: Goertzel, 2014; Bennett, 2025). Currently, three major architectures embody prominently its definition, namely: Hyperon (formerly OpenCog) by Ben Goertzel (2023); Autocatalytic Endogenous Reflective Architecture (AERA) by Khristinn Thórisson (Eberding & Thórisson, 2023; Sheikhlar & Thórisson, 2024), and the Non-Axiomatic Reasoning System (NARS), by Pei Wang (Wang, 2018; Wang, 2025).

To keep things simple, here we will focus and ground mostly on Pei Wang's definition of AGI (Wang, 2008; Wang et al., 2018b; Wang, 2019).

Following the explanation in (Wang et al., 2018b), we define initially just three kinds of AI as: (i) a computer system that behaves exactly like a human mind; (ii) a computer system that solves certain problems previously solvable only by the human mind; (iii) a computer system with the same cognitive functions as the human mind. While *Narrow AI* and *Weak AI* can be defined by (ii), and *Strong AI* by (i) and (iii), AGI necessitates only (iii).

At the basis of the work and concept of Wang stands the AIKR principle: Assumption of Insufficient Knowledge and Resources. This means that an AGI reasoning system must operate without assuming it has complete knowledge, perfect rules, or unlimited time or computing. Thus, NARS is designed to be adaptive under uncertainty and scarcity: conclusions are defeasible, truth is graded (with measures like frequency/confidence), and inference is controlled by resource-aware mechanisms (attention/priority), rather than by exhaustive theorem proving (Wang, 2018; Wang, 2025).
The "weakness" that Wang adopts is greatly inspiring for the Judo Agents project. Wang focuses on lifelong learning, and adaptation features rather than optimisation and ideal infinite resources and computational properties, also to keep all the reactions of an agent at real-time.

Adaptation is the main capability to pursue. Adaptation is lifelong, cumulative, open-ended, multi-objective, and not necessarily convergent. It involves not only modifying the system to fit environmental constraints, but also actively reshaping the environment to satisfy the system's goals. Accordingly, the opposite of intelligence for Wang is not "cannot solve any problem," but "having a constant and invariant ability," which corresponds to the notion of "computation" in computer science (Wang, 2019). Wang's view does not deny that intelligence comprises multiple distinct functions implemented by different mechanisms; rather, it emphasizes the deep interconnections among these processes that together give rise to intelligence as an integrated whole, and seen under pragmatic constructivism (Raikov & Pirani, 2022). Under "Constructivist AI" vision (Thórisson, 2012; Thórisson & Minsky, 2022), if intelligence were merely a toolbox, one must still ask: *what is the hand that coordinates and uses the tools?* (Wang, 2025).

A constructivist perspective offers a useful and partially unifying foundation for addressing questions of representation, particularly in the context of autonomous learning about novel environmental phenomena. It highlights the necessity of autonomic, or cognitive meta-processes, and emphasizes that representations must be designed to support online, real-time modification in order to enable cognitive growth and autonomous self-programming. Such representations must encode causal structure, since effective action in conditions of data abundance and limited processing resources– Wang's AIKR –depends on understanding cause–effect relations. Advancing toward this goal requires extending current approaches to causal reasoning, moving from formal human-level tools such as Pearl's do-calculus (Pearl, 2018) toward autonomic hypothesis generation and empirical validation within self-managing systems (Thórisson & Minsky, 2022).

The constructivistic AGI stance and the AIKR principles are here used as a ground for the concept of Judo Agents.

3 Proposed methodologies for HyR: a concept of Judo Agents

The metaphor of "Judo" is deliberate. In martial arts, Judo does not seek to overpower an opponent through brute force; rather, it relies on balance, minimal effort, redirection of energy, and continuous situational awareness. Strength is expressed through control, proportionality, and sensitivity to context. In the same spirit, Judo Agents are not designed to dominate HyR or to impose optimal solutions at scale. Their role is instead to stabilize interactions, preserve relational integrity, reduce systemic stress, and safeguard human sovereignty within a complex socio-technical environment.

The "gentle" paradigm therefore rejects the idea of total optimization, epistemic absolutism, hidden dominance, or unbounded autonomy. It assumes that intelligence operating in HyR must remain proportionate, reversible, and accountable. Judo Agents are conceived around bounded rationality as a design principle. In addition to bounded rationality here are also considered: limits in knowledge and resources; controlled fallibility as a transparency mechanism, making uncertainty visible rather than concealed; reflexivity as a built-in safety feature, enabling self-monitoring and adaptation; and trust-building as one primary systemic function.

The actual strength of Judo Agents lies not in maximizing performance, but in maintaining equilibrium within evolving human–machine ecologies.

Our vision is that of a new lineage of agents with HyR safety principles built in. Due to the complexity of this task, and in order to ensure efficiency and long-term sustainability of implementation, we conceive them as a holonic Multi-Agent System forming an Agentic Markov Blanket (AMB).

This structure will establish a controlled and adaptive interface between the environment and the agents, whether they are human or artificial. In this way, interactions across the boundary are filtered, regulated, and contextualized, allowing coordinated action while preserving systemic integrity and distributed autonomy. The concept of Markov Blanket, used in Bayesian statistics and Friston's Active Inference framework (Palacios et al., 2020; Friston et al., 2023), is here extended to agentic entities that have the goal to create the control of the cybernetics of the HyR. Figure 1, adapted from Pirani et al. (2025), expresses the context and the area of intervention and concern of Judo Agents. By designing and regulating the filtering and amplifying edges of information variety in a cybernetic loop, they must create a sustainable equilibrium between agencies and their environment.

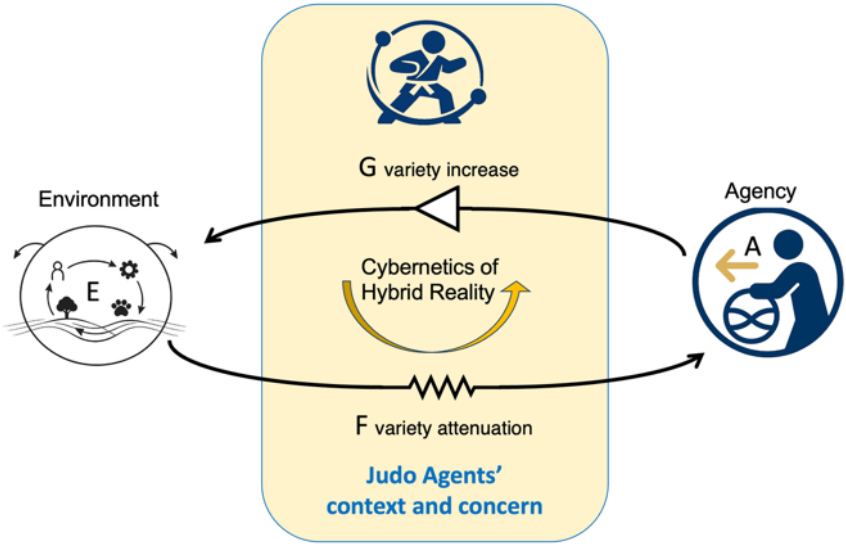

Figure 1: Intervention area of Judo Agents. Realization of the HyR stable and sustainable cybernetics.
Source: own.

In Figure 1, the Agency (A) is to be understood as "emergent property of an autonomous cognitive system to produce non-random variations in the environment", which can be a human, an organisation, an animal, or a machine. The Environment E is what is affected by A, although not in a passive way, but reflexively. Thus E, can be a generic environment made of humans, animals, machines, organisations, or nature. The control loop is the area of intervention of the Judo Agents, where they manage, construct and actuate the filters (F) and gains (G) of variety that make up a typical first-order or higher-order cybernetics (Pirani et al., 2025). The equilibrium of this loop is the sustainable and stable HyR.

The concept of Agentic Markov Blanket (AMB) here introduced, is illustrated in Figure 2 to complement the position about this kind of agents. Figure 2 describes the same concept of the control of Figure 1 but from a topological, dual, and complementary point of view.

It is the case to recall here only that a Markov Blanket is a statistical boundary in a Bayesian network that separates a specific set of internal variables from the rest of the network (external states). Markov Blanket consists of a node's parents, children, and parents of its children, ensuring the internal states are conditionally independent of external states, allowing for efficient modelling and prediction. In the case of extension and adaptation of this structure to AMB, the conditional independence becomes "autonomy". Autonomy here is to be understood as the capability of performing required modelling and predictions, and to realize sustainable cybernetics control loops between the "inside" and the "outside" of the controlled systems – actually "inside" and "outside" depends only by the point of view established, environment and agencies are dually equivalent in HyR (Pirani et al., 2025).

In Figure 2, the E_n variables (or entities) are the "parts" that constitute the environment with which the agency A interacts. The dashed arrows are used to mean that an environment is nevertheless to be considered holistic and all its parts interconnected in general, although our partial modelling might ignore or not handle these connections at first. The other arrows are used to express a cause-effect relationship as usual in Bayesian causal networks. Moreover, Figure 2 explicitly expresses that the Judo Agents constitute a collective and usually not a monolithic intelligent system.

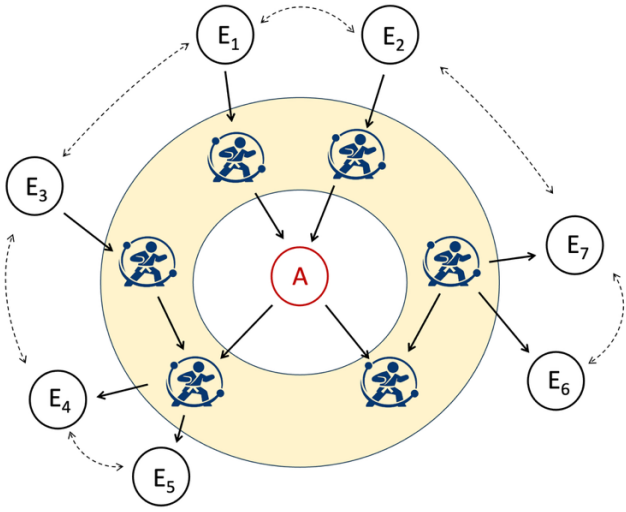

Figure 2: Agentic Markov Blanket
Source: own.

The link between holonic structures and the Markov blankets have been greatly inspired from the study of Palacios et al. (2020). Palacios et al. (2020) explore the mathematical and biological foundations of self-organisation by applying the Free Energy Principle and the concept of Markov blankets to explain how complex life forms emerge. The authors argue that biological systems maintain their integrity by minimizing variational free energy, effectively acting as generative models that "self-evidence" their own existence through nested statistical boundaries–a concept that maps autopoiesis in cybernetics. Using numerical simulations of synthetic cells, the study demonstrates that hierarchical structures, where microscopic elements are "enslaved" by the macroscopic patterns they collectively create, can emerge spontaneously from local interactions and prior genetic "beliefs". Most important for our vision is that this work frames natural selection as a form of Bayesian model selection, suggesting that the recursive formation of Markov blankets is a fundamental requirement for biological existence in a spatially dependent physical world. The recursiveness of these structures creates a natural and immediate parallel and map to holarchy structures (Pirani et al., 2025b). Markov Blankets can be used to describe emerging autopoietic organisation in natural systems, thus can be considered a fundamental structure of life and symbiosis: "*the distinction between a culture of cells and a multicellular organism resides in the emergence of a Markov blanket at the ensemble level*" (Palacios et al., 2020). Authors never use the holon or holarchy

concept, but they imply it when they notice: "*Another consequence of this recursive aspect is the absence of a privileged point of view, when describing hierarchical self-organisation: the dynamics at every level play the role of macroscopic states at the level below, and the role of microscopic states at the level above.*" (Palacios et al., 2020).

The Active Inference framework of Friston could, in principle, suffice to guide the realization of Judo Agents as it offers a theoretical framework that aligns closely with this paradigm. At its core, Active Inference describes intelligence not as reward maximization or global optimization, but as the continuous minimization of uncertainty in order to maintain viable states. An agent maintains a generative model of the world, predicts sensory inputs, and reduces prediction error either by updating its beliefs or by acting on the environment. In this view, intelligence is fundamentally about preserving coherence and existence rather than achieving maximal performance.

This perspective resonates strongly with the critique of the monolithic AGI narrative. Whereas classical AGI models are often framed around unified architectures, globally optimizing objective functions, and seamless scalability, Active Inference emphasizes viability over optimization. Similarly, Judo Agents are not designed to dominate HyR or impose totalizing solutions. Their function is to stabilize interactions, preserve relational integrity, and maintain systemic balance. They are viability regulators rather than performance maximisers.

The metaphor of Judo becomes particularly illuminating in this context. In martial arts, Judo does not rely on brute force but on balance, minimal intervention, and the redirection of existing energies. Active Inference operates in a comparable way: rather than overwhelming the environment, the agent makes minimal adjustments necessary to reduce uncertainty and remain within viable bounds. Judo Agents embody this same principle at a socio-technical level. They intervene proportionately, regulate rather than command, and adapt continuously without seeking epistemic or operational dominance.

When Judo Agents are conceived as a holonic multi-agent system forming an Agentic Markov Blanket (AMB), they function as a distributed interface between humans, artificial agents, and the broader environment of HyR. They filter interactions, coordinate responses, and maintain systemic integrity across the boundary, thereby operationalizing uncertainty regulation at a collective scale.

Moreover, Active Inference presupposes incomplete knowledge and limited resources, echoing the assumption of insufficient knowledge and resources articulated in other adaptive intelligence theories like the AIKR position of Wang here focused. This aligns with the Judo paradigm's emphasis on bounded rationality, controlled fallibility, and reflexivity. Rather than assuming epistemic closure or perfect optimization, Judo Agents are designed to remain adaptive, transparent about uncertainty, and capable of revising their internal models in response to evolving contexts.

Friston's Active Inference and the Free Energy Principle are conceptually powerful but face several practical limits. First, their high level of abstraction creates challenges for empirical testability: because free energy minimization can be used to interpret almost any adaptive behaviour, deriving clear, falsifiable predictions is difficult. Second, computational tractability is a major constraint, since maintaining and updating rich probabilistic generative models in high-dimensional environments requires strong approximations, and most existing implementations remain toy or highly constrained systems rather than scalable, open-ended agents. Third, the framework presupposes a generative model of the environment, yet constructing, parameterizing, and updating sufficiently expressive models in complex socio-technical contexts remains an engineering burden not eliminated by the theory. Fourth, free energy minimization is normatively neutral: it explains adaptive regulation but does not encode ethical alignment, and agents may minimize surprise in socially undesirable ways without additional constraints. Fifth, industrial adoption has been limited, as reinforcement learning and deep learning pipelines are currently easier to benchmark and commercialize. Finally, the philosophical breadth of the principle, extended to all self-organizing systems, while inspiring, makes it harder to translate directly into concrete design prescriptions.

For these reasons, our proposal is to adopt a generalized Cyber-Systemic Systems Engineering approach, as described in Pirani et al. (2025), in order to leverage the best available science and technology from the multiple disciplines that can contribute interdisciplinarily and across different scales, as recalled mainly in section 2.3. The new step here added is the explicit use of AGI entities in form of a lineage of multi-agent systems that would be practically used to express the cybernetic control needed for the realization of the equilibrium of the HyR as of Figure 1.

3.1 Foundational pillars towards practical realization

The Judo Agent paradigm is grounded in three converging intellectual traditions: higher-order cybernetics, holonic systems theory, and constructivist models of trust. Together, these perspectives provide the theoretical foundation for agents designed to operate responsibly within HyR.

Higher-order cybernetics extends classical feedback models by including the observer within the system and emphasizing reflexive loops of interpretation and self-referential regulation (Raikov & Pirani, 2022; Pirani et al., 2025b). In HyR, agents cannot be conceived as external controllers acting on a passive environment; they are embedded participants within evolving socio-technical networks. Judo Agents therefore adopt a second- and third-order cybernetic stance. They monitor not only system variables but also their own internal models, treat their knowledge as provisional, and incorporate reflexive trust mechanisms into their operation. Regulation becomes self-aware and adaptive, continuously revising both actions and underlying assumptions.

The paradigm is also informed by holonic systems theory. A holon is simultaneously a whole in itself and a part of a larger whole. HyR naturally exhibits such holarchic structures (Pirani et al., 2025b): individuals within organizations, organizations within markets, and cyber-physical systems within broader infrastructures.

Judo Agents are designed as holonic entities that are autonomous yet subordinate, self-regulating while embedded, and locally intelligent while globally accountable. They do not replace existing systemic layers; instead, they integrate into them and contribute to coherence across scales.

In addition, the Judo paradigm draws on constructivist models of trust. Trust is not a binary condition but a dynamic, context-dependent construct that evolves through interaction (Raikov & Pirani, 2022; Pirani et. al, 2025d). In blockchain-enabled HyR, trust can be operationalized through probabilistic reasoning, Bayesian updating, verifiable computation, and decentralized validation (Naeem et al., 2025). Judo Agents therefore operate with explicit and adaptive trust models that evolve over time, record epistemic shifts, and incorporate uncertainty as a formal variable. In this framework, trust is not assumed as a static premise but managed as an integral component of systemic regulation (Pirani et al., 2025d).

3.2 Architectural Features of Judo Agents

A Judo Agent is defined by a set of architectural principles designed to ensure stability, proportionality, and respect for human sovereignty within HyR.

First, it embodies a refined and more stringent version of Simon's bounded rationality, as implied even by the AIKR framework (Wang, 2019). Judo Agents adopt a viability-constrained rationality, in which decision processes are explicitly limited by structural uncertainty, resource bounds, and systemic stability requirements rather than global optimization objectives. Judo Agents adopt a *viability-oriented, non-axiomatic rationality* grounded in structural epistemic insufficiency. This viability-oriented, non-axiomatic rationality integrates higher-order cybernetic reflexivity, holonic embeddedness across scales, structural epistemic insufficiency (e.g., Wang's AIKR), and resource-aware decision processes, thereby privileging systemic stability and human sovereignty over global optimization.

Rather than pursuing maximal performance, the Judo Agent constrains the scope of its actions and explicitly recognizes the incompleteness of its models and the insufficiency of its resources. Decision-making is oriented toward maintaining viable system states rather than achieving theoretical optima. Limits are not treated as weaknesses but as structural safeguards against overreach and destabilization. In this sense, rationality is engineered as a stability-preserving function.

Controlled fallibility is an essential architectural component. Following Ross Ashby's work (Ross Ashby, 1952), he suggested that truly intelligent machines should not be infallible; rather, their ability to fail, adaptively reorganize, and maintain internal stability (homeostasis) in the face of change is what makes them act like human brains. Thus, our agents must render uncertainty explicit by exposing confidence levels, declaring margins of error, and deferring to human judgment when predefined thresholds are exceeded. Fallibility is therefore not hidden but formalized as a transparency mechanism that supports trust and reduces systemic fragility.

Epistemic humility follows naturally from this non-axiomatic stance. The agent does not assume epistemic closure or final truth. It continuously revises its beliefs, accepts contestation, and integrates reflexive feedback loops that monitor both environmental variables and its own internal models. Knowledge remains provisional and dynamically updated within a cybernetic process of self-regulation.

Judo Agents are explicitly designed to preserve human sovereignty. By endowing them with the abovementioned properties and embedding them holonically within broader socio-technical structures, they act as co-controllers rather than replacements. They neither override normative human judgment nor monopolize decision authority, and they maintain operational transparency. Their purpose is not to dominate HyR, but to stabilize it while safeguarding the primacy of human agency. Nonetheless, they must retain their constitutional autonomy, meaning they preserve the necessary independence to counter harmful requests or moral drift on the part of humans. For this reason, HyR should be understood as a symbiosis between moral peers rather than a master–slave relationship in either direction.

3.3 Judo Agents as integrity sentinels, research roadmap, and societal implications

Within HyR, integrity becomes a central systemic concern encompassing informational integrity, relational coherence, structural stability, and normative consistency. In such an environment, Judo Agents function as integrity sentinels. Their role is not to maximize performance but to minimize destabilization. They monitor the cybernetics across interconnected subsystems, detect patterns of over-optimization that may compress resilience margins, identify opacity risks that undermine trust, and intervene to stabilize emerging breakdowns in coordination. By operating as distributed regulators embedded within holonic structures, they preserve equilibrium across technical, social, and institutional layers.

The development of Judo Agents requires a structured, multi-phase research agenda. The first phase concerns conceptual modelling, including the formalization of viability-oriented, non-axiomatic rationality architectures, their integration within holonic management frameworks, and the specification of adaptive trust model evolution protocols.

The second phase involves simulation environments in which multi-agent configurations are tested within hybrid socio-technical models, stress-tested under adversarial and uncertain conditions, and evaluated using explicit systemic integrity metrics, which is still an open research issue by itself.

The third phase focuses on Blockchain integration, enabling smart contract interfaces for trust model updates, on-chain recording of epistemic transitions, and the verification of adaptive trust mechanisms–considering also technologies like Self-Sovereign Identity and Zero-Knowledge-Proofs, for overall democracy and consensus.

The final phase consists of real-world pilot deployments, such as supply chain integrity systems, cyber-physical manufacturing supervision platforms, and institutional decision-support agents designed to operate as co-regulators rather than decision monopolists.

Ethically and societally, the Judo paradigm reframes AGI. It is neither an apex intelligence nor a competitor to humanity, but a relational capability embedded within governance structures and designed to stabilize HyR in sustainable sense. This implies transparency by design, accountability through architectural embedding, and humility engineered into decision processes. At the societal level, it supports distributed sovereignty, plural epistemologies, and systemic resilience. In this sense, Judo Agents are not instruments of dominance but infrastructures of balance, enabling human–machine symbiosis grounded in integrity rather than optimization.

4 Conclusion

The work presented here is conceptual and position-oriented, intended to establish a research agenda and roadmap. The position presented here should have shed some light on how to address an ambitious and challenging research question: how can Hybrid Reality systems (HyR) be designed and governed to ensure systemic integrity, preserve human sovereignty, and sustain trust among heterogeneous interacting intelligences?

HyR represents a structurally entangled socio-technical ecosystem in which humans, artificial agents, cyber-physical systems, and institutional infrastructures co-evolve. In this context, the central challenge is no longer the maximization of intelligence, but the preservation of systemic integrity, human sovereignty, and relational trust. Classical AGI narratives based on global optimization risk destabilizing such complex environments.

This paper has introduced Judo Agents as a "gentle" alternative: socially embedded, AGI-oriented agents designed around viability-oriented bounded rationality, controlled fallibility, epistemic humility, and reflexive regulation. Rather than dominating HyR, they would act as co-controllers and integrity sentinels, stabilizing interactions and preserving equilibrium across holonic structures. Conceived as a holonic multi-agent system forming an Agentic Markov Blanket, they regulate the boundary between human and artificial agencies while maintaining distributed autonomy.

Ethically and societally, the Judo paradigm reframes AGI as a relational capability embedded in governance structures, grounded in transparency, accountability, and humility. The objective is not to create apex optimizers, but infrastructures of balance capable of sustaining a stable and symbiotic HyR.

With the introduction of Judo Agents as a continuation of the sustainable cybernetics frameworks for HyR discussed here, we outline a roadmap that prepares us for the imminent socio-technical challenges ahead.

Here we have left out of scope the implementation and realization issues that will require these special multi-agent systems to be ubiquitous, pervasive and green at the same time. This is another fundamental dimension of the research that should complete this "gentle" path proposal towards sustainable HyR.

The current level of abstraction in the presentation of Judo Agents is still high. This work is intentionally positioned as a conceptual and foundational contribution, aimed at framing a research agenda rather than delivering a finalized implementation or providing empirical evidence.

In this direction, future work will focus on the development of illustrative scenarios and reference use cases within HyR environments (e.g., healthcare workflows, industrial cyber-physical systems), where Judo Agents can be instantiated as governance-aware, integrity-preserving components. These scenarios will be complemented by simplified prototypes and simulation-based experiments to explore key properties such as stability, trust propagation, and resistance to integrity attacks.

Furthermore, we envisage a stepwise validation pathway: starting from controlled digital twin environments, moving toward semi-real deployments, and ultimately assessing performance in real-world socio-technical systems. This will allow us to empirically evaluate the effectiveness of Judo Agents in maintaining systemic integrity and enabling adaptive governance. By combining conceptual rigor with progressive empirical grounding, we aim to transform the current abstract formulation into a testable and deployable framework that can be openly leveraged by a community of scientific, civil, and social beneficiaries and contributors.

References

Agbemabiese, W. T. (2026). Toward Constitutional Autonomy in AI Systems: A Theoretical Framework for Aligned Agentic Intelligence. *IEEE Access*. doi:10.1109/ACCESS.2026.3654907

AGI (2026). The AGI Society. Online, accessed, 24th April 2026: https://agi-society.org/

Amodei, D. (2026). The Adolescence of Technology. Available at: https://www.darioamodei.com/essay/the-adolescence-of-technology

Amodei, D. (2026b). Claude's Constitution. Online, accessed on 24th April 2026: https://www.anthropic.com/constitution

Agbemabiese, W. T. (2026). Toward Constitutional Autonomy in AI Systems: A Theoretical Framework for Aligned Agentic Intelligence. *IEEE Access*.

Ashby, M. (2020). Ethical regulators and super-ethical systems. *Systems*, *8*(4), 53. doi:10.3390/systems8040053

Ashby, M. (2022). Problems with abstract observers and advantages of a model-centric cybernetics paradigm. *Systems*, *10*(3), 53. doi:10.3390/systems10030053

Bennett, M. T. (2025). What the F* ck Is Artificial General Intelligence?. *arXiv preprint arXiv:2503.23923*.

Bhati, D., Neha, F., Bandaru, D. S., Weber, M., & Gajera, I. D. (2026). Large Language Models: A Survey of Architectures, Training Paradigms, and Alignment Methods. Avaialble at: https://www.preprints.org/frontend/manuscript/abaf3e6f650e2aeab454da56761a79aa/download_pub

Daniel, D. (2026). Optimized to Death: The Hypernetic Law of Experience. *Systems*, *14*(2), 197.

Derigent, W., Cardin, O., & Trentesaux, D. (2021). Industry 4.0: contributions of holonic manufacturing control architectures and future challenges. *Journal of Intelligent Manufacturing*, *32*(7), 1797-1818. doi:10.1007/s10845-020-01532-x

Eberding, L. M., & Thórisson, K. R. (2023, May). Causal reasoning over probabilistic uncertainty. In *International Conference on Artificial General Intelligence* (pp. 74-84). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-33469-6_8

FFG, Founders Forum Group (2025). AI Statistics 2024–2025: Global Trends, Market Growth & Adoption Data. Available at: https://ff.co/ai-statistics-trends-global-market/

Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1-29. doi:https://doi.org/10.1016/j.physrep.2023.07.001

Goertzel, B. (2007). *Artificial general intelligence* (Vol. 2, p. 1). C. Pennachin (Ed.). New York: Springer.

Goertzel, Ben. "Artificial general intelligence: concept, state of the art, and future prospects." *Journal of Artificial General Intelligence* 5.1 (2014): 1-48.

Goertzel, B., Bogdanov, V., Duncan, M., Duong, D., Goertzel, Z., Horlings, J., ... & Werko, R. (2023). Opencog hyperon: A framework for agi at the human level and beyond. *arXiv preprint arXiv:2310.18318*.

Hahne, P. Z., & Schmoelz, A. (2026). Trusting the machine: a digital humanist perspective on misplaced trust in artificial intelligence. *AI and Ethics*, *6*(1), 115. doi: 10.1007/s43681-025-00923-1

Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. *Computer*, *36*(1), 41-50.

Koestler, A. (1968). *The ghost in the machine*. Macmillan.

Lee, E. A. (2009). Computing needs time. *Communications of the ACM*, *52*(5), 70-79.

Lee, E. A. (2020). *The coevolution: The entwined futures of humans and machines*. Mit Press.

Moltbook (2026). A Social Network for AI Agents. Online, accessed 24th April 2026: https://www.moltbook.com/

Naeem, T., Pirani, M., & Spalazzi, L. (2025, June). Evidence-based oracles using bayesian network. In *2025 21st International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)* (pp. 1-6). IEEE. doi:10.1109/DCOSS-IoT65416.2025.00151

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. (2020). On Markov blankets and hierarchical self-organisation. *Journal of theoretical biology*, *486*, 110089. doi:10.1016/j.jtbi.2019.110089

Pearl, J., & Mackenzie, D. (2018). *The Book of Why*. Penguin.

Perko, I. (2021). Hybrid reality development-can social responsibility concepts provide guidance?. *Kybernetes*, *50*(3), 676-693. doi:10.1108/K-01-2020-0061

Pirani, M., Cucchiarelli, A., Naeem, T., & Spalazzi, L. (2025). A blockchain-driven cyber-systemic approach to hybrid reality. *Systems*, *13*(4), 294. doi:10.3390/systems13040294

Pirani, M., Carbonari, A., Cucchiarelli, A., Giretti, A., & Spalazzi, L. (2025b). The Meta Holonic Management Tree: review, steps, and roadmap to industrial Cybernetics 5.0. *Journal of Intelligent Manufacturing*, *36*(8), 5285-5326. doi:10.1007/s10845-024-02510-3

Pirani, M., Generosi, A., & Spalazzi, L. (2025c). A perspective on generative artificial intelligence for the enhancement of methods in system dynamics. In *20th IRDO International Conference: Innovative, sustainable & socially responsible society 2025: Personal responsibility as a part of social responsibility and sustainability (ISBN 978-961-7141-11-5). IRDO Institute for the Development of Social Responsibility-Editorial.*. Available at: https://www.irdo.si/irdo2025/posters/45.pdf

Pirani, M., Bonifazi, G., Cucchiarelli, A., Naeem, T., & Spalazzi, L. (2025d, October). Holonic Oracle Constructivism in Cyber-Physical Systems. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 680-683). IEEE. doi: 10.1109/SMC58881.2025.11343522

Reg AI (2026). Xiaomi has just built a fully automated factory in Changping. Online, accessed 24th April 2026: https://www.instagram.com/reel/DT0Mzdnk-hn/

Rentahuman (2026). Let our humans take it from here. Online, accessed 24th April 2026: https://rentahuman.ai/

Raikov, A. N., & Pirani, M. (2022). Human-machine duality: What's next in cognitive aspects of artificial intelligence?. *IEEE Access*, *10*, 56296-56315. doi:10.1109/ACCESS.2022.3177657

Raikov, A., Giretti, A., Pirani, M., Spalazzi, L., & Guo, M. (2024). Accelerating human–computer interaction through convergent conditions for LLM explanation. *Frontiers in Artificial Intelligence*, *7*, 1406773. doi:10.3389/frai.2024.1406773

Ross Ashby, W. (1952). *Design for a brain*. Wiley.

Ross Ashby, W. (1956). *An introduction to cybernetics*. Chapman & Hall Ltd, London

Sarthak, K. (2026). Xiaomi just opened a fully automated smartphone factory in Changping, Beijing. Online, accessed 24th April 2026: https://www.linkedin.com/posts/thesarthakkumar_xiaomi-darkfactory-automation-activity-7421987062100389888-EZuQ

Schmelzer, R. (2026). Rentahuman.ai Turns Humans Into On-Demand Labor For AI Agents. Forbes. Online, accessed 24th April 2026: https://www.forbes.com/sites/ronschmelzer/2026/02/05/when-ai-agents-start-hiring-humans-rentahumanai-turns-the-tables/

Schneier, B. (2025). AI and Trust. *Communications of the ACM*, *68*(8), 29-33. doi: 10.1145/3737610

Sheikhlar, A., & Thórisson, K. R. (2024, July). Causal generalization via goal-driven analogy. In *International Conference on Artificial General Intelligence* (pp. 165-175). Cham: Springer Nature Switzerland. doi:10.1007/978-3-031-65572-2_18

Thórisson, K. R. (2012). A new constructivist AI: from manual methods to self-constructive systems. In *Theoretical Foundations of Artificial General Intelligence* (pp. 145-171). Paris: Atlantis Press.

Thórisson, K. R., & Minsky, H. (2022). The future of AI research: ten defeasible 'axioms of intelligence'. In International Workshop on Self-Supervised Learning (pp. 5-21). PMLR.

Valckenaers, P., & Van Brussel, H. (2015). *Design for the unexpected: From holonic manufacturing systems towards a humane mechatronics society*. Butterworth-Heinemann.

Wang, P. (2008, March). What do you mean by" AI"?. In *AGI* (Vol. 171, pp. 362-373).

Wang, P., Li, X., & Hammer, P. (2018). Self in NARS, an AGI System. *Frontiers in Robotics and AI*, *5*, 20. doi:10.3389/frobt.2018.00020

Wang, P., Liu, K., & Dougherty, Q. (2018b). Conceptions of artificial intelligence and singularity. *Information*, *9*(4), 79. doi:10.3390/info9040079

Wang, P. (2019). On defining artificial intelligence. *Journal of artificial general intelligence*, *10*(2), 1-37. doi:10.2478/jagi-2019-0002

Wang, P. (2025). *Non-axiomatic logic: A model of intelligent reasoning, 2nd edition*. World Scientific.

Zhang, Z., Yang, F., Qin, X., Zhang, J., Lin, Q., Cheng, G., ... & Zhang, Q. (2024). The vision of autonomic computing: Can llms make it a reality?. *arXiv preprint arXiv:2407.14402*.