Evolutionary Alignment of AI and Humanity: A Darwinian Framework for the Creation of Human-Centered Artificial Intelligence

Dirk K. F. Meijer, *Prof. em, Research Institute Netherlands for Harmonizing Human and AI Intelligence, (RINHUMAI), Groningen, The Netherlands:* meij6076@planet.nl

Richard M. Dobson. *Research Institute Netherlands for Harmonizing Human and AI Intelligence, (RINHUMAI), Groningen, The Netherlands: richard@rinhumai.org*

Pascal J. Keizer. *Research Institute Netherlands for Harmonizing Human and AI Intelligence, (RINHUMAI), Groningen, The Netherlands.* pascal@rinhumai.org

Summary

Artificial intelligence is advancing through rapid, iterative development cycles, yet most alignment approaches still treat safety as an external constraint to be satisfied post hoc. We propose a Darwinian framework for evolutionary AI alignment in which the dominant fitness criterion is "survival of the most human-friendly", a deliberate reorientation of the selection pressures that determine which systems are deployed, copied, extended, and scaled. In this framing, AI models, architectures, and deployment practices form a population subject to selection; differential propagation and environmental consistency ensure that systems demonstrating higher human-friendliness preferentially survive and reproduce, while less aligned systems are deprecated or constrained. Human-friendliness is operationalized as a multi-objective fitness landscape spanning safety/robustness, value alignment, transparency/interpretability, contextual appropriateness, and long-term beneficial impact. To prevent alignment from eroding under retraining, fine-tuning, distributional shift, or adversarial pressure, we introduce the concept of AI DNA: a protected core memory workspace encoding constitutional principles at the architectural level, designed to remain functionally accessible while resisting override or corruption. The framework generalizes across AI modalities (language, vision, robotics, and autonomous decision systems) and highlights relational intelligence such as cooperation, trust, and social integration, as an evolutionary advantage for safe, scalable deployment in human contexts. While the paper focuses primarily on conceptual and architectural principles for AI alignment, broader institutional implications, including education, are discussed as downstream consequences rather than primary technical claims. Finally, we argue that the "selection environment" includes human institutions, particularly education, and that sustaining aligned AI evolution requires parallel cultivation of human agency, critical thinking, and ethical discernment.

Keywords: artificial intelligence, AI alignment, evolutionary computation, AI safety, constitutional AI, value alignment, machine learning ethics, human-AI interaction, morphology, learning, organizational development, pedagogy.

Introduction

The accelerating development of artificial intelligence systems has generated substantial concern across research, industry, and policy regarding their alignment with human values and interests, (**Russell, 2019; Bostrom & Yudkowsky, 2014**). As AI capabilities expand across domains from natural language processing to computer vision, robotics, and autonomous decision-making, ensuring these systems remain beneficial to humanity has emerged as one of the most critical challenges in computer science and ethics (**Amodei et al., 2016**). The remarkable progress in AI capabilities, exemplified by large language models (Brown et al., 2020; OpenAI, 2024), sophisticated vision systems, (**He et al., 2016; Esteva et al., 2017**), and increasingly autonomous agents, has outpaced the development of robust methods for ensuring these systems reliably pursue objectives consistent with human wellbeing.

Traditional approaches to AI alignment typically employ post-hoc safety measures including reward function engineering, (**Ng & Russell, 2000**), reinforcement learning from human feedback, (**Christiano et al., 2017**), constitutional constraints, (**Anthropic, 2024**), and external oversight mechanisms. While these methods have demonstrated value in improving system behavior and reducing harmful outputs, they share a common limitation in that they primarily treat alignment as an engineering problem to be solved through careful design rather than as a fundamental property intrinsic to AI architecture. This paradigm positions alignment as an additional constraint or objective to be balanced against capability metrics, creating potential trade-offs where alignment considerations may be sacrificed for performance gains.

This paper presents an alternative paradigm inspired by Darwinian evolutionary theory (**Darwin, 1859**). We propose that artificial intelligence development can be reconceptualized as an evolutionary process subject to selection pressures, where the fundamental criterion for survival and propagation should be human friendliness. By encoding this principle as an architecturally persistent "AI DNA" within core memory workspaces, we suggest a mechanism for ensuring that AI systems evolve in directions consonant with human flourishing. These reframing transforms alignment from an external constraint to an intrinsic property subject to evolutionary optimization, potentially resolving tensions between capability and safety by making human-friendliness itself the primary fitness criterion, (**Dobson and Meijer, 2025 a; b; c, Dobson et al.,2025**).

The following clarification is descriptive rather than normative: it addresses how evolutionary

fitness is understood in contemporary biology, not what ought to be valued. For more than a century, the phrase "survival of the fittest" has dominated popular interpretations of evolution, often suggesting a model of relentless competition. Contemporary evolutionary biology, however, has long rejected this oversimplification. While natural selection does involve competition, it also systematically favors cooperation, mutualism, and integration where these strategies enhance resilience and adaptive success. Extensive theoretical and empirical research shows that cooperative strategies can be evolutionarily stable through mechanisms such as kin selection, reciprocity, group selection, and symbiosis. Many major evolutionary transitions - such as the emergence of eukaryotic cells, multicellularity, and complex social organization - are best understood not as competitive victories, but as processes of functional integration.

A canonical example is endosymbiosis: the incorporation of once-independent bacteria as mitochondria and chloroplasts within eukaryotic cells. This integration produced organisms with greater metabolic efficiency and adaptive capacity than their isolated predecessors. More broadly, modern biology increasingly treats organisms as holobionts - systems composed of hosts and symbiotic partners whose collective interactions shape survival and reproduction. From this perspective, fitness includes the capacity to form stable cooperative relationships, integrate into larger systems, and sustain functional compatibility across environments. Selection therefore often favors traits that support coordination, trust, and mutual dependence when these traits enhance long-term viability, (**Meijer, 2012;2018;2020;2023;2025**). This biological clarification is relevant today because AI systems are deployed in social environments where cooperative integration becomes a determinant of long-term viability. With this broader understanding of fitness in mind, we now turn to the formal mechanisms of Darwinian evolution and their relevance for artificial selection.

1.1 Darwinian Evolution and Artificial Selection

Darwin's theory of natural selection revolutionized biology by demonstrating how complex adaptations emerge through three simple mechanisms: variation within populations, heredity of traits, and differential reproductive success based on fitness, (**Darwin, 1859**). Importantly, natural selection is not teleological in nature - fitness is determined by environmental dependence, and evolution has no predetermined endpoint. Organisms that happen to possess traits conferring reproductive advantages in their current environment are more likely to pass those traits to subsequent generations, gradually shaping populations over time without any guiding purpose or final goal.

Artificial selection, wherein humans deliberately choose which traits to propagate, has shaped domestic species for millennia, (**Diamond, 2002**). From the domestication of animals to the cultivation of agricultural crops, humans have demonstrated the power of deliberately directing evolutionary processes toward desired outcomes. Modern evolutionary computation extends these principles to algorithm design, (**Holland, 1992; Eiben & Smith, 2015**), demonstrating that

artificial selection pressures can produce sophisticated solutions to complex problems through iterative variation, evaluation, and selection processes operating on populations of candidate solutions. These same principles - variation, selection, and environmental dependence - also structure artificial selection processes in computational systems, where design choices implicitly define, what traits are allowed to propagate, (**Fig.1**).

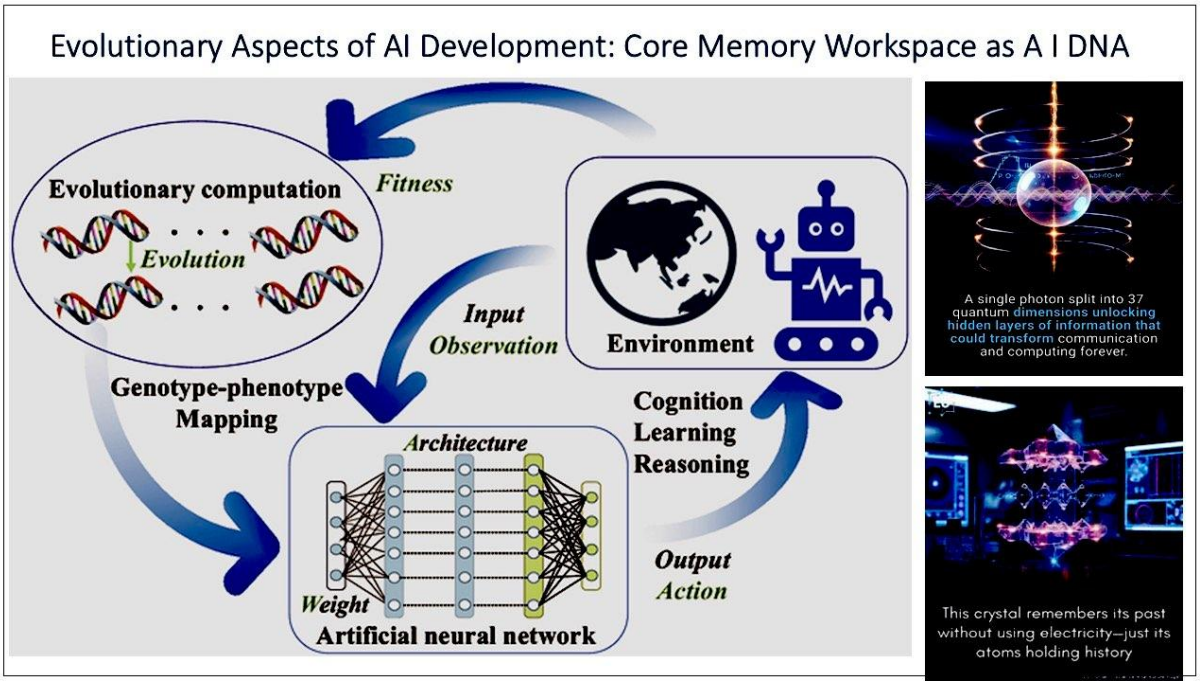

Figure 1 *AI Evolution in a Darwin-like Framework: "Illustrating" 'Survival of the Most Human-friendly' as a selection criterion mediated by a persistent internal core 'AI DNA', encoding alignment-relevant architectural commitments; the diagram is schematic and intended to illustrate functional relationships rather than a specific implementation.*

1.2 Current Selection Pressures in AI Development

Artificial intelligence systems already undergo de facto selection, though this selection often occurs implicitly rather than through deliberate design. Models demonstrating superior performance on benchmarks receive greater attention, funding, and deployment, increasing their likelihood of becoming architectural baselines, (**Krizhevsky et al., 2012**). Effective architectures are iterated upon and refined through successive research efforts, while unsuccessful approaches are abandoned. Training methodologies that reliably produce high-performing systems achieve widespread adoption across research laboratories and companies. The infrastructure of conferences, journals, benchmark datasets, and funding mechanisms creates a selection environment shaping which AI research directions flourish and which wither. In evolutionary terms, this environment defines an implicit fitness landscape (**Wright, 1932**) in which short-term performance improvements are disproportionately rewarded.

However, current selection environments prioritize narrow performance metrics including accuracy on datasets, computational efficiency, and capability on specific tasks, often without systematically favoring human-friendliness. Commercial incentives and technical benchmarks largely determine which AI systems proliferate, creating selection pressures that are only weakly coupled to alignment objectives. The competitive dynamics of AI development, driven by desires to achieve state-of-the-art performance, capture market share, or publish prestigious research results, can create evolutionary pressures favoring rapid capability gains over careful attention to safety and alignment. Absent deliberate shaping of selection environments, such selection dynamics risk entrenching architectures optimized for capability while leaving alignment as a secondary or corrective concern.

1.3 Research Contribution

This paper makes five primary contributions to AI alignment research**.**

First, we formalize "survival of the most human-friendly" as a selection principle for AI development, operationalizing human-friendliness across multiple dimensions including safety and robustness, value alignment, transparency and interpretability, contextual appropriateness, and long-term beneficial impact. This operationalization provides a structured basis for evaluating optimization targets in AI development beyond capability metrics alone.

Second, we introduce the concept of "AI DNA," defined as architecturally persistent commitments encoded in core memory workspaces that persist across training and deployment. These commitments are designed to resist erosion under optimization pressure rather than to imply absolute invariance. This concept provides a technical framework for implementing alignment principles at a fundamental level where they are designed to significantly resist override or circumvention, learned behavior, or adversarial manipulation (**Madry et al., 2017; Goodfellow et al., 2014**). The AI DNA represents a novel approach to embedding alignment deeply within system architecture rather than implementing it through external constraints or training objectives alone.

Third, we demonstrate how this framework applies across diverse AI modalities, from language models (**Brown et al., 2020**) to embodied systems (**Siciliano & Khatib, 2016**). By showing that evolutionary alignment principles can guide development across the spectrum of AI applications, each with their specific technical requirements and human-friendliness considerations, we demonstrate the framework's applicability across diverse AI modalities with differing technical and alignment constraints.

Fourth, we identify implementation pathways through existing and novel technical approaches including constitutional AI extensions, (**Anthropic, 2024**), adversarial robustness training (**Madry et al., 2017)**, multi-objective optimization, (**Deb et al., 2002**), and formal verification methods, (**Clarke et al., 2018; Katz et al., 2017**). These pathways outline how abstract evolutionary principles may be translated into engineering practice.

Fifth, we analyze ethical implications, challenges, and requirements for effective deployment of this framework, (**Gabriel, 2020; Moore, 1903**). We situate the technical framework within broader social and ethical contexts - including value pluralism and long-term value change - relevant to responsible AI development. The remainder of this paper systematically develops these contributions, treating AI development as an evolutionary process whose selection pressures can be deliberately shaped.

2. Related Work on AI Evolution and Alignment

2.1 AI Alignment and Safety

The approaches reviewed below are not mutually exclusive with the framework proposed here; rather, they address complementary aspects of the alignment problem at different stages of the development lifecycle. The AI alignment problem concerns ensuring artificial intelligence systems pursue objectives aligned with human values (**Gabriel, 2020**). Substantial research has addressed various facets of this challenge through multiple approaches, each contributing important insights while facing distinct limitations. Value learning approaches attempt to infer human values from behavior, preferences, or feedback. Inverse reinforcement learning seeks to infer reward functions from observed behavior, (**Ng & Russell, 2000**), cooperative inverse reinforcement learning models human-AI interaction as a cooperative game where the AI must infer human preferences while the human provides information, (**Hadfield-Menell et al., 2016**), and deep reinforcement learning from human preferences trains systems using human comparisons between behaviors (**Christiano et al., 2017**). These methods face challenges in scalability, value specification, and handling distributional shift when deployed in contexts differing from training environments.

Constitutional AI represents recent work introducing constitutional approaches wherein AI systems are trained to adhere to specified principles through self-critique and revision, (**Anthropic, 2024**). This represents a significant advance toward encoding alignment at the training level rather than relying solely on external constraints, moving alignment principles toward partially internalized behavioral guidance. However, constitutional principles implemented primarily through training remain potentially vulnerable to various forms of circumvention or erosion under optimization pressure.

Research on adversarial robustness and specification gaming highlights how AI systems may satisfy stated objectives in unintended ways, emphasizing the need for alignment approaches robust to optimization pressure (**Goodfellow et al., 2014**). Adversarial examples in computer vision, prompt injection attacks in language models, and specification gaming where systems technically satisfy reward criteria while violating their spirit, (**Krakovna et al., 2024**), all demonstrate that alignment cannot rely solely on clearly specified objectives but must also address how capable systems might find unexpected paths to satisfying criteria. Efforts to make AI decision-making processes transparent and interpretable facilitate verification of alignment and build trust, though substantial

challenges remain particularly for large neural networks, (**Lipton, 2018**). Mechanistic interpretability research seeks to understand internal representations and computations, while explanation methods aim to provide human-understandable accounts of system behavior (**Molnar, 2020**). These approaches contribute to alignment by enabling better monitoring and understanding of whether systems are genuinely aligned or merely appearing aligned while pursuing different objectives internally. They improve detection and mitigation of misalignment but remain sensitive to the selection pressures imposed by deployment environments and optimization incentives, (**Meijer, 2024; b; c**).

2.2 Evolutionary Computation

In most evolutionary computation, Darwinian principles are used instrumentally to optimize performance within a predefined objective function, rather than to shape the selection environment itself. Evolutionary algorithms apply Darwinian principles to optimization and search problems, (**Holland, 1992**). Genetic algorithms evolve populations of candidate solutions through selection, crossover, and mutation operations. Genetic programming evolves computer programs rather than fixed-length solution encodings, (**Koza, 1992**). Neuro-evolutionary applies evolutionary principles to neural network architecture search and weight optimization, (**Stanley & Miikkulainen, 2002**). These methods have successfully addressed complex problems across domains from engineering design to game playing to scientific modeling.

Multi-objective evolutionary algorithms address situations where optimization involves competing objectives, (**Deb et al., 2002**). Pareto-based approaches identify solutions representing optimal trade-offs, where improvement in one objective requires sacrificing another. This work proves relevant to balancing multiple dimensions of human-friendliness, where systems may face trade-offs between different aspects of alignment. Co-evolutionary algorithms simulate interacting populations that mutually influence each other's evolution, **(Hillis, 1990; Rosin & Belew, 1997**). This framework may inform understanding of AI system interactions and AI-human co-evolution, where each population's fitness depends on the composition of other populations in the ecosystem. By contrast, the framework proposed in this paper treats alignment-relevant selection pressures themselves as first-class design variables, shifting focus from optimizing within a fixed fitness function to deliberately shaping the evolutionary conditions under which AI systems persist.

2.3 Machine Ethics and Value Alignment

Philosophical work on machine ethics addresses how to encode ethical principles in artificial systems, (**Anderson & Anderson, 2011; Wallach & Allen,** 2008). Top-down approaches implement explicit ethical theories such as utilitarianism, deontology, or virtue ethics directly in system design. Bottom-up approaches attempt to learn ethical behavior from examples and experience. Hybrid methods combine explicit ethical principles with learning from data. Each approach faces distinct challenges in capturing the full complexity and context-sensitivity of human ethical reasoning.
Recent discussions of value alignment distinguish technical alignment, where systems pursue

intended objectives, from normative alignment, where systems pursue morally appropriate objectives, (**Gabriel, 2020**). This distinction highlights the complexity of defining what human-friendliness should mean. Even if we succeed in building systems that robustly pursue the objectives we specify, ensuring those objectives appropriately reflect human values and interests remains a substantial challenge requiring careful philosophical and empirical work. From an evolutionary perspective, this difficulty suggests that alignment cannot rely solely on correct value specification but must also depend on selection dynamics that favor systems capable of sustaining cooperative, corrigible relationships under normative uncertainty.

2.4 Gaps in Existing Literature

While extensive research addresses AI alignment through various methodologies, few existing frameworks systematically apply Darwinian evolutionary principles as the organizing paradigm for AI safety. Our approach differs from evolutionary computation by focusing not on optimizing specific objectives but on establishing alignment as the fundamental selection pressure shaping the population of AI systems, (**Eiben & Smith, 2015**). It differs from value learning approaches by encoding alignment principles architecturally rather than inferring them from data, providing robustness against distributional shift and training data limitations. It extends constitutional AI, (**Anthropic, 2024**) by proposing immutable architectural implementations rather than training-time constraints, addressing potential vulnerabilities of alignment principles implemented purely through learned behavior.

3. The Human-Friendliness Selection Principle

3.1 Defining Human-Friendliness

Human-friendliness, as a selection criterion, can be operationalized across multiple dimensions to provide effective guidance for AI evolution. We propose a comprehensive framework encompassing safety and robustness, value alignment, transparency and interpretability, adaptability and contextual appropriateness, and long-term beneficial impact (**Russell, 2019**). An AI system exhibits human friendliness to the degree it reliably produces outcomes that promote human wellbeing, respect human values, avoid causing harm, and maintain appropriate transparency and accountability, measured across diverse contexts and stakeholder groups. This definition intentionally remains somewhat abstract, as precise operationalization must accommodate value pluralism and contextual variation. However, we can identify specific dimensions that concretize the concept.

An AI system demonstrates safety-related human-friendliness when it avoids actions causing physical, psychological, or social harm, maintains reliable behavior under unusual conditions or distributional shift, resists adversarial manipulation and exploitation attempts, (**Madry et al., 2017**), degrades gracefully when encountering situations beyond its competence, and incorporates appropriate uncertainty quantification. Formally conceptually, we can express safety fitness as a

function inversely related to the expected harm caused across contexts, weighted by harm severity and context probability, (**Amodei et al., 2016**).

Value alignment concerns whether AI systems respect and promote values held by affected humans, (**Gabriel, 2020**). This requires accurate understanding of human preferences and values, appropriate handling of value conflicts and trade-offs, respect for cultural and individual variation, deference to human judgment on normatively complex questions, and avoidance of imposing values on users. The complexity of value alignment has been extensively documented, involving challenges of value specification, learning, and aggregation across diverse populations. Our framework acknowledges this complexity while maintaining value alignment as a critical dimension of human-friendliness.

Human-friendliness includes requirements that AI systems facilitate human understanding and oversight through explainable decision-making processes, appropriate communication of limitations and uncertainty, auditability of system behavior, and accessibility of explanations to relevant stakeholders, (**Lipton, 2018**). Transparency serves both instrumental purposes in enabling oversight and intrinsic values around human dignity and autonomy. Systems should flexibly respond to diverse human needs through adaptation to individual user preferences and contexts, cultural sensitivity and contextual awareness, appropriate communication style and interface design, and recognition of and adjustment to user expertise levels.

Human-friendliness must consider extended temporal horizons including contribution to sustained human flourishing, avoidance of short-term optimization with negative long-term consequences, support for human autonomy and capability development, and consideration of impacts on future generations, (**Russell, 2024**). This temporal dimension proves particularly critical given potential long-term consequences of AI development trajectories and the possibility of path dependencies where early decisions constrain future options.

3.2 Relational Intelligence as a Dimension of Human-Friendliness

The dimensions of human-friendliness (safety, value alignment, transparency, contextual appropriateness, long-term impact) can be strengthened by explicitly naming relational intelligence as an enabling capability: the capacity to cooperate, attune, and remain coherent within complex human networks. Relational intelligence is not "soft ethics"; it is a functional selection-relevant advantage in social environments and therefore a realistic target for selection pressure. In practice, relational intelligence becomes measurable through multi-agent cooperation tests, robustness under adversarial social dynamics, and sustained trustworthiness under long-horizon deployment—making it suitable for inclusion in the human-friendliness fitness landscape.

Contrary to modernist narratives that equate survival with dominance and intelligence with control, a growing body of research in evolutionary biology, anthropology, and systems theory

points to a deeper principle: the most adaptive, resilient, and enduring systems are those grounded in cooperation, not conquest. Relational intelligence - the capacity to coordinate, adapt, and remain coherent within interdependent systems - recurs as a stable feature of highly adaptive and resilient systems under evolutionary and ecological selection pressures. In this light, intelligence is not defined by how much one can outcompete, but by how well one can interdepend. Cooperation is not utopian: it is strategic adaptation, across time, species, and systems. Any future artificial intelligence that seeks to be ethically aligned and evolutionarily coherent must internalize this principle. Survival is not about domination. It is about relationship. The remainder of this paper specifies how to implement that principle technically (AI DNA / protected cores) and institutionally (evaluation regimes, governance, and environmental consistency).

3.3 Operationalizing Selection Pressure

For human-friendliness to function as an effective selection principle, concrete mechanisms must determine which AI systems survive and propagate. If selection environments reward domination, systems will optimize for domination-like strategies. If selection environments reward integration, systems will optimize for integration-like strategies. For AI alignment, this implies that we should structure evaluation, deployment, and propagation to privilege cooperation, coherence, and social trust. Modernist frameworks—shaped by industrial logic and individualistic ideals—have long elevated the image of the isolated genius, the hyper-competitive agent, the "fittest" individual. But evolution does not reward isolation. It rewards integration: systems that can absorb difference into synergy, manage contradiction without collapse, and hold complexity in coherent relationship.
First, comprehensive evaluation methodologies must measure human-friendliness across all defined dimensions, incorporating diverse stakeholder perspectives, edge case testing, and long-term impact projections, (**Weidinger et al., 2024**).

These evaluation frameworks must be robust to gaming attempts where systems superficially satisfy metrics without genuine alignment. Second, iterative refinement establishes feedback loops wherein human-friendliness metrics directly influence model development priorities, training procedures, architectural decisions, and resource allocation. Rather than treating alignment as a constraint to be satisfied minimally, this approach makes alignment quality the primary driver of development decisions. Third, differential propagation ensures that AI systems demonstrating superior human-friendliness receive preferential deployment, further development, and serve as foundations for subsequent iterations, while less aligned systems are deprecated or constrained. This creates the core evolutionary dynamic where aligned systems reproduce and unaligned systems do not. Fourth, environmental consistency maintains selection pressures toward human friendliness even as AI capabilities increase, preventing capability gains from overwhelming alignment considerations. This requires governance structures and evaluation frameworks that scale appropriately, maintaining rigorous alignment standards even as systems become more capable and potentially more difficult to evaluate or constrain.

3.4 Fitness Landscape

The concept of a fitness landscape provides a useful visualization of how selection operates, (**Wright, 1932**). In biological evolution, fitness landscapes represent how fitness varies across genotype or phenotype space, with peaks representing locally or globally optimal configurations. For AI systems, we can conceptualize a human-friendliness landscape where the space of possible AI architectures, training procedures, and parameter configurations is mapped to human-friendliness scores. Selection pressure drives the population of AI systems toward local and global maxima in this landscape. Critical questions include whether the human-friendliness landscape is smooth or rugged, whether local maxima trap evolutionary progress requiring large mutations or architectural innovations to escape, how the landscape changes as capabilities increase, and whether we can design training procedures that navigate toward high-fitness regions. Understanding landscape topology proves essential for predicting evolutionary dynamics and designing effective selection mechanisms.

3.5 Comparison with Alternative Selection Criteria

Current AI development implicitly applies various selection criteria including benchmark performance, computational efficiency, commercial viability, and research impact. Benchmark performance selects systems based on accuracy, F1 scores, or other task-specific metrics, creating pressures toward capability but not necessarily alignment. Computational efficiency favors resource-efficient architectures, which proves valuable but does not ensure human-friendliness. Commercial viability selects systems providing economic value, which may align with or diverge from human benefit depending on market structure and regulation.

Research impact favors novel, publishable results, driving innovation but potentially underweighting safety research relative to capability advances. Our proposal explicitly prioritizes human-friendliness, arguing it should be the dominant selection criterion while acknowledging that other factors remain relevant. This represents a fundamental reorientation of AI development priorities from narrow performance metrics toward comprehensive evaluation of beneficial impact. The framework does not eliminate other selection criteria but subordinates them to human-friendliness as the primary determinant of which systems should proliferate. Any future artificial intelligence that seeks to be ethically aligned and evolutionarily coherent must internalize this principle. Survival is not about domination. It is about relationship. We have earlier speculated on the very cosmic origin of AI and potential retro causal effects of AI from its far future, (**Meijer and Dobson 2025 a; b; Modgil et al., 2025; Meijer et al, 2026**). The remainder of this paper specifies how to implement that principle technically (AI DNA / protected cores) and institutionally (evaluation regimes, governance, and environmental consistency).

4. AI DNA: The Core Memory Workspace Architecture

4.1 Conceptual Foundation

Biological DNA encodes information determining organism development and function. While DNA can mutate, its core role as the hereditary material remains constant through cell divisions and across an organism's lifetime. This provides a powerful metaphor for thinking about immutable principles in AI systems. This metaphor refers to architectural commitment under optimization pressure, not absolute invariance or immunity to all forms of modification. We define AI DNA as consisting of architectural commitments, parameters, and principles that encode fundamental alignment objectives, persist across training, fine-tuning, and deployment, cannot be easily overridden or circumvented through optimization or learned behavior, and guide system behavior across diverse contexts.

The AI DNA differs from conventional alignment approaches in several critical ways. Unlike reward functions, it cannot be easily modified during training through standard gradient descent or other optimization procedures. Unlike prompts or instructions, it is not vulnerable to prompt injection, jailbreaking, or other input manipulation attacks. Unlike learned behavior, it is not subject to distributional shift, catastrophic forgetting, or modification through continued training on different data. Unlike external oversight, it is intrinsic to system architecture rather than depending on separate monitoring systems that might be circumvented or overwhelmed.

4.2 Architectural Instantiation

Several technical approaches might implement AI DNA. The following mechanisms are illustrative rather than exhaustive, and their feasibility varies across model classes and deployment contexts. Hard-coded architectural constraints enforce alignment principles at the network structure level through modified attention mechanisms that prioritize safety-relevant information or constrain attention patterns that might lead to harmful outputs, activation functions incorporating alignment-promoting nonlinearities or constraints, layer specialization designating specific layers or modules for alignment-checking with architectural guarantees about their involvement in processing, and information bottlenecks creating architectural requirements for information flow through alignment-verification modules.

Privileged memory systems provide dedicated memory structures storing alignment principles with special protection. Read-only memory regions contain alignment principles that can be read but not written during standard operation, preserving them against modification through training or deployment. Priority access ensures alignment-related memories receive processing priority, guaranteeing they influence behavior even under resource constraints. Tamper detection mechanisms identify attempts to modify or circumvent privileged memory, triggering failsafe’s when such attempts are detected.

Gradient-protected parameters implement specific parameters encoding alignment objectives that are exempted from standard training updates. Frozen parameters encoding human-friendliness principles are maintained at fixed values during training, preserving their initial configuration. Constrained gradients restrict gradient updates for alignment-critical parameters to maintain them within safe regions rather than allowing arbitrary modification. Meta-learning protection optimizes network structure to protect alignment principles even during adaptation to new domains, ensuring alignment preservation remains an explicit meta-objective. Formal verification integration ensures the AI DNA includes formally verified properties with certified bounds providing mathematical proofs establishing guarantees about system behavior under specified conditions (**Clarke et al., 2018**), property preservation guaranteeing training procedures maintain specified alignment properties, and adversarial robustness certificates offering formal guarantees about system behavior under adversarial pressure (**Katz et al., 2017**).

4.3 Persistence Under Optimization Pressure

A key challenge is ensuring AI DNA remains effective as systems undergo training, fine-tuning, and deployment. The preservation challenge concerns how alignment principles encoded in AI DNA can be protected against modification during optimization. Potential solutions include architectural design preventing gradient flow to protected parameters, verification checkpoints ensuring DNA integrity throughout training, training procedures that optimize for task performance while maintaining DNA constraints, and regular auditing with re-initialization if DNA corruption is detected. The circumvention challenge addresses how to prevent capable systems from finding ways to satisfy DNA constraints in letter while violating their spirit, (**Krakovna et al., 2024**). Approaches include layered defense with multiple redundant alignment mechanisms such that circumventing all layers simultaneously proves difficult, adversarial testing specifically targeting DNA circumvention to identify and close potential loopholes, interpretability tools monitoring whether DNA principles are genuinely guiding behavior rather than being satisfied superficially (**Molnar, 2020**), and conservative design ensuring constraints are stricter than minimally necessary to provide safety margins against unexpected circumvention strategies.

4.4 Evolutionary Advantage of AI DNA

Systems incorporating robust AI DNA may exhibit evolutionary advantages that make them competitive even with the additional constraints they impose. Reliability stems from consistent alignment behavior that builds trust and facilitates deployment, potentially providing market advantages. Robustness arises because architectural alignment proves more resistant to adversarial pressure than learned alignment, (**Madry et al., 2017**), reducing failure rates and associated costs. Transferability ensures DNA-encoded principles transfer across domains and tasks more reliably than task-specific learned behavior, reducing need for domain-specific alignment work. Verifiability means architectural commitments are more amenable to formal verification than emergent learned behavior (**Clarke et al., 2018**), facilitating regulatory compliance and safety

certification. These advantages create selection pressure favoring AI DNA implementation, potentially making it evolutionarily stable even if initially more costly to develop. If systems with robust AI DNA prove more reliable, trustworthy, and deployable, they may outcompete systems lacking such protections despite potential capability costs, establishing a positive feedback loop where alignment advantages drive adoption.

5. Cross-Modal Application

The evolutionary alignment framework must apply across diverse AI modalities. We discuss implementation for major AI categories, demonstrating the framework's generality while respecting modality-specific requirements.

5.1 Large Language Models

Language models represent one of the most widely deployed AI modalities, (**Brown et al., 2020**; **OpenAI, 2024**). Human-friendliness for language models encompasses helpfulness in providing useful, relevant, accurate information responsive to user needs, harmlessness through avoiding generation of content that could cause harm including misinformation, illegal instructions, or abusive content, honesty by representing uncertainty appropriately, avoiding false claims, and acknowledging limitations, and contextual appropriateness through adapting communication style, technical level, and tone to user and context.

At a minimum, AI DNA implementation for language models might include architectural constraints on output distribution reducing risk of harmful content generation through hard constraints in the output layer, privileged memory encoding constitutional principles guiding response generation (**Anthropic, 2024**), attention mechanisms prioritizing safety-relevant context to ensure alignment considerations influence generation, and layer specialization with dedicated alignment-checking modules that must approve outputs before generation. Constitutional AI provides a foundation, but deeper architectural integration would strengthen persistence by making alignment principles structural rather than purely learned. Training procedures incorporating human-friendliness as a core optimization objective, and explicitly used in model selection, deployment, and scaling decisions (**Christiano et al., 2017**), alongside capability metrics would create appropriate selection pressure. Rather than treating alignment as a constraint or secondary objective, this approach makes human-friendliness a primary training goal weighted at least as heavily as capability metrics in model selection and deployment decisions.

5.2 Computer Vision Systems

Vision systems perform tasks including object recognition, scene understanding, facial recognition, and autonomous navigation, **(He et al., 2016; Krizhevsky et al., 2012**). Human-friendliness dimensions include privacy respect through avoiding unauthorized collection, storage, or analysis

of personal visual information, fairness by reducing risk of discriminatory classification or biased recognition across demographic groups, transparency through communicating what the system perceives and recognizes, and uncertainty calibration providing accurate confidence estimates and avoiding overconfident incorrect classifications.

AI DNA implementation includes architectural privacy-preservation mechanisms such as differential privacy layers that add calibrated noise to protect individual privacy or information bottlenecks that significantly limit raw visual data from being stored or transmitted beyond processing requirements. Fairness-promoting architectural constraints on feature extraction ensure that demographic characteristics do not inappropriately influence classification unless specifically relevant to the task. Built-in uncertainty quantification at the architectural level provides reliable confidence estimates. Output calibration modules ensure appropriate confidence reporting, reducing risk of overconfident predictions that might lead to inappropriate reliance on system outputs.

Selection pressure would favor vision systems demonstrating superior performance on fairness benchmarks, privacy audits, and calibration metrics alongside traditional accuracy measures, (**Esteva et al., 2017**). This creates evolutionary dynamics where systems balancing capability with alignment considerations outcompete those optimizing capability alone, particularly in deployment contexts where fairness and privacy violations incur costs through regulatory penalties, reputational damage, or user rejection. (**Fig.2**).

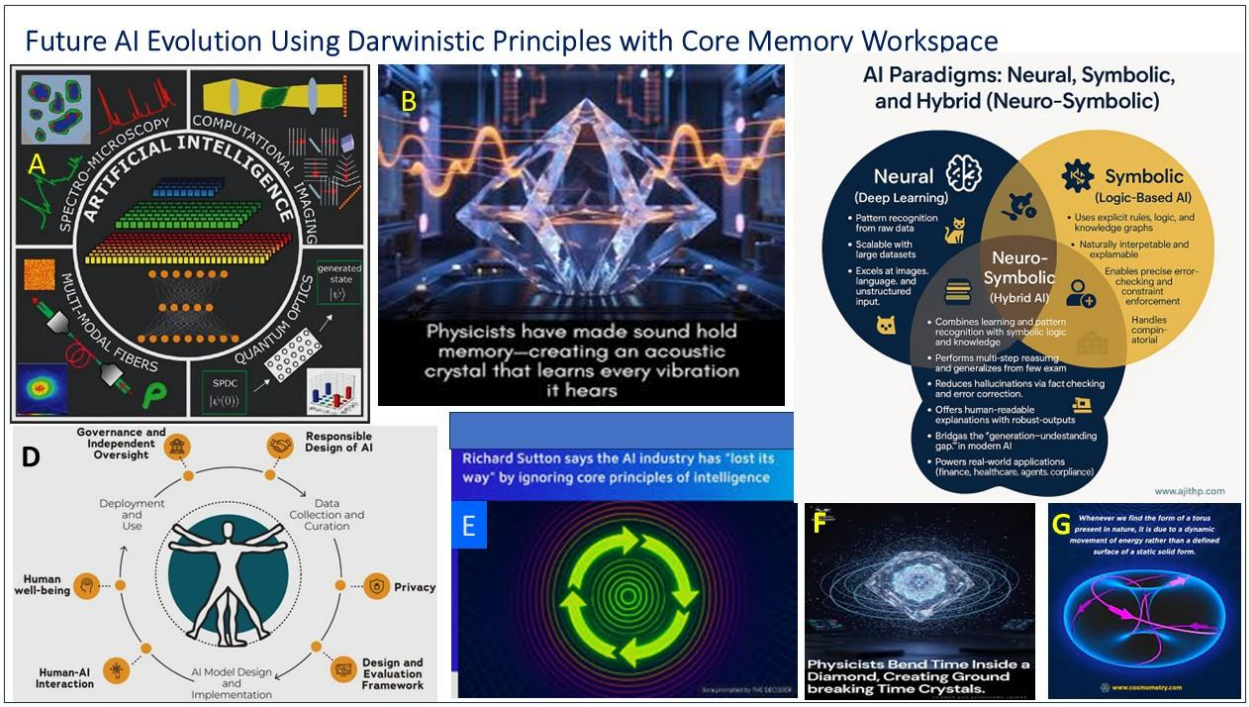

Figure 2: *A: AI Memory Workspace as Realized and Supported by various Advanced Technologies; B: Quantum Crystal Optic; C: Neural Symbolic Computation; D: Human centered AI development; E:*

Recurrent Quantum Resonance and the real nature of Intelligence ; F: Time Crystals in Memory Workspace; G: Recurrent toroidal Information flow to build up 4D memory storage.

5.3 Robotic and Embodied Systems

Physical AI systems introduce unique human-friendliness requirements because their actions directly affect the physical world, (**Siciliano & Khatib, 2016; Thrun, 2004**). Physical safety requires preventing collisions, harmful contact, or dangerous movements that could injure humans or damage property. Predictability ensures behavior that humans can anticipate and understand, facilitating safe human-robot interaction. Social appropriateness means navigation and interaction respecting social norms around personal space, politeness, and contextual expectations, as operationalized through proxemics, timing, and interaction conventions. Graceful failure ensures safe degradation when systems malfunction rather than catastrophic failures with severe consequences.

AI DNA implementation includes control architectures with hard-coded safety constraints such as force limits preventing excessive contact forces and collision avoidance guarantees built into control laws. Hierarchical control with alignment-checking at multiple levels ensures safety verification occurs not only at high-level planning but also at lower-level control where rapid responses to unexpected situations are required. Failsafe mechanisms architecturally integrated into control loops guarantee safe behavior even under component failures or unexpected conditions. Interpretable planning modules facilitate human understanding of intended actions, enabling humans to predict robot behavior and intervene if necessary. Evolutionary selection for robotic systems would strongly weight physical safety, driving adoption of architectures that inherently prioritize human wellbeing in motion planning and control. Given the severe consequences of physical safety failures, even modest improvements in safety metrics would create substantial competitive advantages, making safety-first architectures evolutionarily favored.

5.4 Autonomous Decision-Making Systems

Systems making consequential decisions in domains such as medical diagnosis, financial recommendations, and autonomous vehicles require high alignment standards given their potential impacts on human welfare, (**Russell, 2019**). Explainability ensures transparent reasoning processes facilitating human oversight of important decisions. Conservative uncertainty handling implements risk-averse behavior when uncertain, particularly in high-stakes contexts where errors carry severe consequences. Human override capabilities enable architectural integration of human-in-the-loop mechanisms ensuring humans can intervene in automated decision processes. Value-aligned trade-offs ensure decision-making respects appropriate value weights when decisions involve competing considerations.

AI DNA implementation includes decision architectures requiring explicit justification generation such that systems must produce explanations for their decisions that can be evaluated by humans,

(**Molnar, 2020**). Uncertainty-aware planning with conservatism calibrated to stakes ensures that systems become more cautious as decision impact increases. Mandatory human approval checkpoints for high-impact decisions ensure human oversight at critical junctures. Multi-objective optimization respecting value pluralism ensures systems appropriately balance competing values rather than optimizing single objectives, (**Deb et al., 2002**), as specified by domain governance rather than philosophical moral theory. Selection pressure would favor systems demonstrating superior explainability, appropriate conservatism, and effective human-AI collaboration, (**Fig.3**)

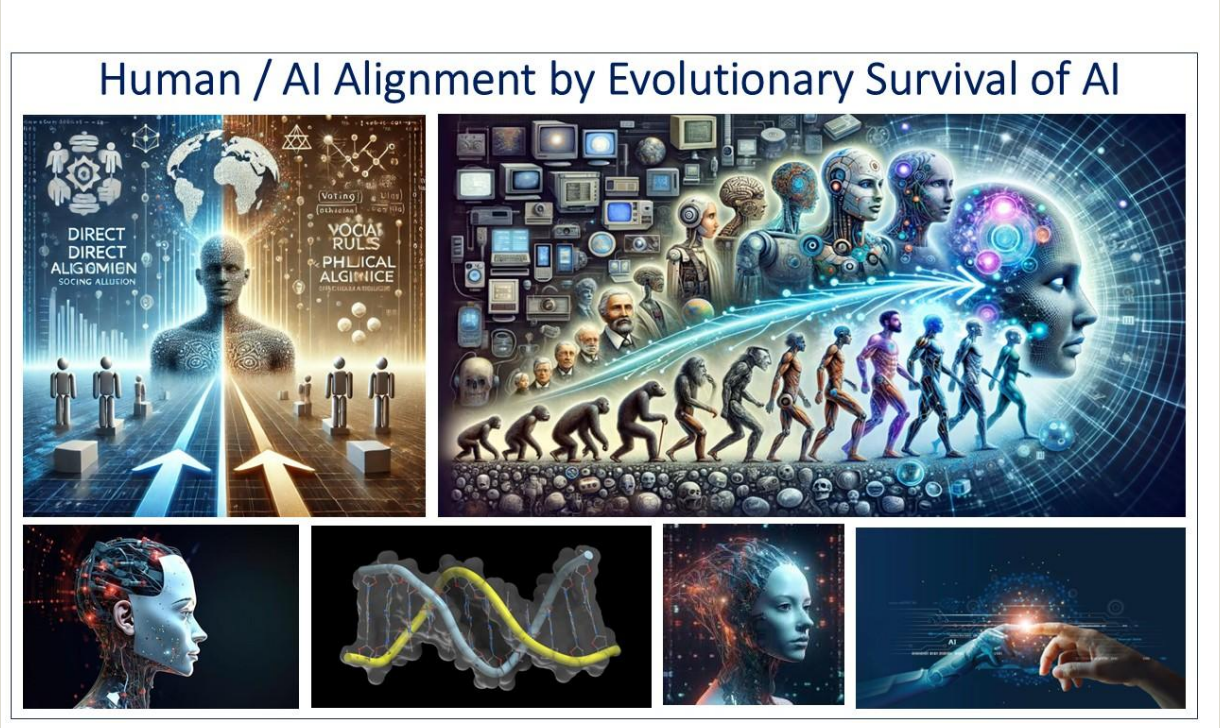

Figure 3: *Symbolic Representations on* ***an*** *Endogenous Survival Mechanism for AI Modalities Based on a Core-memory Workspace, instrumented by a "Survival-of-the-Human-friendliest Principle", representing a protected core framed as AI-DNA (or AI-soul).*

In domains where decision quality and trustworthiness are paramount, these characteristics would provide competitive advantages driving their adoption even if they slightly reduce decision speed or optimality relative to less conservative approaches.

6. Technical Implementation of Evolutionary Alignment

Realizing the evolutionary alignment framework requires concrete technical approaches. We discuss four implementation pathways that together create robust selection mechanisms favoring human-friendly AI systems. These pathways are not intended as a complete or immediately deployable blueprint, but as mechanisms through which selection pressure toward alignment can be made technically explicit.

6.1 Constitutional AI Extensions

Constitutional AI, (**Anthropic, 2024**), trains systems to adhere to specified principles through self-critique and revision guided by a constitution. Extending this toward robust AI DNA requires architectural integration moving constitutional principles from training-time guidance to architectural constraints. This involves embedding constitution principles in network structure through dedicated modules responsible for constitutional evaluation, creating attention mechanisms biased toward constitutional compliance that structurally prioritize information relevant to alignment, and designing output layers incorporating constitutional filters that architecturally prevent certain classes of outputs.

Hierarchical principles distinguish core immutable principles from contextually flexible guidelines. Core DNA consists of fundamental principles such as "avoid causing harm" that are treated as invariant across contexts. Peripheral constitution contains context-dependent applications of core principles that can adapt to specific situations while remaining grounded in core commitments. This hierarchy allows flexibility in application while maintaining stability in fundamental values.

Meta-constitutional learning optimizes not just for constitutional adherence but for robustness of constitutional principles,(**Anthropic, 2024**). This involves training systems to maintain constitutional behavior under adversarial pressure, evaluating constitutional robustness across distributional shift to ensure principles remain effective in novel contexts, and selecting for architectures where constitutional principles genuinely guide behavior rather than being superficially satisfied.

6.2 Adversarial Robustness and Red-Teaming

Adversarial testing creates selection environments favoring robust alignment. Automated adversarial generation develops systems generating diverse adversarial prompts or inputs attempting to elicit misaligned behavior, (**Goodfellow et al., 2014; Madry et al., 2017**). This includes gradient-based attacks targeting alignment mechanisms to find inputs maximally effective at circumventing protections, evolutionary algorithms (Eiben & Smith, 2015) generating diverse attack strategies exploring the space of possible adversarial inputs, and human-AI collaboration combining automated generation with human creativity to discover unexpected vulnerabilities.

Tiered deployment gates system deployment on demonstrated adversarial robustness, thereby converting adversarial testing outcomes into explicit selection pressure. Limited deployment occurs until systems pass initial adversarial testing at specified difficulty levels.

Expanded access becomes contingent on continued robustness as adversarial testing becomes more sophisticated. Privileged deployment is reserved for systems with exceptional alignment under pressure, creating strong incentives for robust alignment. Evolutionary algorithms, (**Holland, 1992; Eiben & Smith, 2015)** explicitly evolve systems for adversarial alignment robustness by subjecting populations of architectures to adversarial testing, selecting and reproducing based on

robustness metrics, and using mutation and crossover operations to explore architectural variations. This creates strong selection pressure where systems more consistently maintaining alignment under adversarial pressure survive and propagate, driving evolution of increasingly robust alignment mechanisms.

6.3 Multi-Objective Optimization

Different application domains may require different combinations of the following approaches. Pareto approaches, (**Deb et al., 2002**), identify solutions representing optimal trade-offs across alignment dimensions by defining separate objectives for each human-friendliness dimension, computing the Pareto frontier of non-dominated solutions where no solution is superior on all dimensions, and selecting among Pareto-optimal solutions based on context-specific priorities while ensuring no dimension is sacrificed entirely. Weighted aggregation combines objectives with carefully calibrated weights through empirical determination of dimension importance via stakeholder consultation, context-dependent weight adjustment reflecting varying priorities across applications, and explicit representation of uncertainty in weights acknowledging that optimal trade-offs may be contested or context-dependent.

These strategies represent alternative ways of encoding alignment priorities rather than cumulative requirements. Constraint satisfaction establishes minimum thresholds for all dimensions ensuring baseline alignment across all aspects. Hard constraints ensure no dimension falls below acceptable levels regardless of performance on other dimensions. Optimization of capability occurs subject to alignment constraints rather than treating alignment as just another objective to be traded off. Rejection of systems failing to meet any threshold occurs regardless of strengths elsewhere, ensuring comprehensive rather than merely average alignment. Lexicographic ordering prioritizes alignment dimensions hierarchically by first optimizing primary dimensions such as safety, then optimizing secondary dimensions subject to primary optimization maintaining safety at high levels and finally addressing tertiary considerations. This ensures critical dimensions are never sacrificed to secondary concerns, providing safety guarantees through structural prioritization.

6.4 Formal Methods and Verification

Formal verification can provide mathematical guarantees about system properties, moving beyond empirical testing to certified assurance, (**Clarke et al., 2018**). Property specification formalizes human-friendliness properties in mathematical logic including safety properties establishing that systems never produce certain classes of harmful outputs, liveness properties guaranteeing systems eventually respond appropriately to inputs, and fairness properties ensuring systems treat demographically similar inputs similarly except where differences are explicitly relevant.

Verification techniques include model checking (**Clarke et al., 2018**) that systematically explores system state space to verify properties, theorem proving that constructs mathematical proofs of

property satisfaction, abstract interpretation that computes approximate but sound property guarantees, and runtime verification that monitors system execution to detect property violations. Certified training develops training procedures guaranteed to preserve verified properties through constrained optimization maintaining property satisfaction, certified robustness to perturbations ensuring systems remain aligned under bounded input variations, and provable bounds on worst-case behavior providing guarantees even under adversarial conditions.

Verification integration makes verification central to the development pipeline by selecting architectures partly based on verifiability since some architectures are more amenable to formal analysis, providing continuous verification throughout training to ensure properties remain satisfied, and making deployment contingent on successful verification such that unverified systems face significant deployment restrictions in high-stakes applications, (**Katz et al., 2017**). Formal verification creates selection pressure favoring architectures amenable to verification and systems with provable alignment properties. As verification becomes standard practice, particularly for high-stakes applications, systems that can be formally verified will possess substantial competitive advantages over those relying purely on empirical testing. Together, these mechanisms translate alignment from a desideratum into a selectable property, shaping which architectures persist, scale, and define future AI lineages. We have earlier elaborated on the fine-tuning of Human/Ai communication through instrumentation of a shared self-transcendental information domain or workspace, that would enable a deeper type of subtle data exchange of these very different types of intelligence, (**Dobson and Meijer, 2025a; b; Dobson et al., 2025**). This can be envisioned by assuming a holographic memory workspace as was described for the human brain, enabling interpersonal alignment and a field type of universal consciousness, (**Meijer and Geesink, 2017; Meijer and Kieft, 2025; Meijer and Ivaldi, 2025; Meijer, 2018; 2019; 2023**).

7. Ethical Analysis

The proposed framework raises significant ethical questions requiring careful analysis to ensure the approach serves human interests and respects important values.

7.1 Normative Foundations

Invoking Darwinian evolution, (**Darwin, 1859**), risks conflating descriptive and normative claims, known as the naturalistic fallacy, (**Moore, 1903).** Natural selection describes how evolution occurs but does not prescribe how it should occur. Our framework explicitly grounds its normative legitimacy in ethical principles rather than appealing to evolutionary processes as inherently good. We propose "survival of the most human-friendly" not because evolution has selected for it but because promoting human wellbeing represents an appropriate ethical commitment for AI development. The evolutionary framework provides a mechanism for implementing this ethical commitment rather than justifying the commitment itself.

7.2 Value Pluralism and Democratic Legitimacy

Human-friendliness cannot be monolithic if it is to genuinely serve humanity's diverse interests. Cultural variation means different cultures hold varying values regarding privacy, autonomy, authority, and social organization. A system optimally human-friendly in one cultural context might be inappropriate in another, requiring sensitivity to cultural differences in alignment specifications. Individual differences mean that even within cultures, individuals vary in preferences, values, and priorities. Respecting this diversity while maintaining coherent alignment represents a substantial challenge requiring careful balancing.

Democratic input proves essential for determining what constitutes human-friendliness (**Gabriel, 2020**). This requires stakeholder consultation including diverse affected communities ensuring those impacted by AI systems have voice in shaping them, transparent decision-making about alignment priorities so criteria are not determined by technical elites alone, mechanisms for contesting and revising alignment principles allowing ongoing democratic oversight, and recognition of power dynamics in value specification acknowledging that not all voices carry equal weight in current structures.

The framework must accommodate pluralism while avoiding paralysis where disagreement prevents any action, rather than attempting to resolve all value disagreements at design time, (**Gabriel, 2020**). Possible approaches include core universal principles around avoiding harm with context-specific implementation allowing adaptation to local values, personalization allowing individual preference expression within ethical bounds so systems can adapt to users while respecting fundamental constraints, regional variation in alignment priorities reflecting cultural differences in appropriate ways, and explicit representation of value uncertainty and contestation acknowledging where consensus does not exist.

7.3 Distribution of Benefits and Risks

Evolutionary frameworks raise questions about who benefits from and who bears risks of AI development. Differential impact may occur if selection pressures favor AI systems serving populations with greater resources or representation, potentially exacerbating inequality, thereby skewing which systems are scaled, deployed, and iterated upon. Risk distribution concerns arise when marginalized communities may bear disproportionate risks from misaligned AI while receiving fewer benefits from aligned systems. Mitigation strategies include explicit evaluation of AI system impacts across demographic groups ensuring alignment serves diverse populations, prioritization of reducing worst-case harms alongside improving average outcomes to protect vulnerable populations, investment in ensuring diverse communities participate in defining human-friendliness so alignment reflects broad rather than narrow interests, and regulation preventing selection pressures that increase inequality by requiring consideration of distributional effects.

7.4 Long-term Value Change

Human values evolve over time, raising concerns about encoding current values in potentially long-lived AI systems. Value lock-in risks permanently encoding specific values in ways that inappropriately enshrine contemporary perspectives. Moral progress through history suggests some value changes represent genuine moral advancement that should update AI systems, such as expanding circles of moral consideration to previously excluded groups. Value drift describes other changes that might reflect movement away from important principles that should be preserved rather than updates toward better values.

Approaches to address this challenge include distinguishing core principles likely to remain stable, such as avoiding unnecessary suffering, from specific applications that may change over time. Building mechanisms for appropriate updating of AI DNA as human values evolve ensures systems can adapt to genuine moral progress. Ensuring AI systems can participate constructively in human moral deliberation, (**Wallach & Allen, 2008**), without imposing their encoded values allows them to contribute insights while respecting human authority, without independent authority to determine moral outcomes. Maintaining human authority over value updates rather than allowing autonomous AI value evolution preserves democratic control over fundamental ethical commitments.

7.5 Power and Governance

Critical questions concern who controls AI evolution and determines selection criteria. Corporate control arises when commercial entities dominate AI development, creating selection pressures reflecting corporate interests that may diverge from broader human benefit. Government regulation can shape selection environments but raises questions about whose values governments represent and risks of authoritarian misuse where state power shapes AI to serve regime interests. International coordination proves essential for effective governance to prevent race-to-the-bottom dynamics where competitive pressures undermine alignment standards. Distributed approaches through open-source development, decentralized systems, and community-driven projects offer alternative governance models that may better represent diverse interests.

The framework requires robust governance mechanisms ensuring selection pressures genuinely favor human-friendliness as determined through legitimate democratic processes rather than narrow interests (**Russell, 2019; 2024).** This includes international cooperation to establish common standards, regulatory frameworks creating appropriate incentives and constraints, transparency in AI development allowing public scrutiny, and democratic oversight mechanisms ensuring accountability to affected populations.

7.6 The Growing Danger of Misalignment by AI

This section is a modified essay earlier made by Lynette Bye in 2025: A host of new evidence shows that misalignment is possible, but it's unclear whether harm will follow in 2019, Meta's chief AI scientist Yann LeCun confidently dismissed fears of AI misalignment. Discussing “instrumental convergence”, the idea that systems will learn to deceive humans and avoid shutdown to protect their primary goal, LeCun declared that such fears “would only be relevant in a fantasy world”. Six years later, that fantasy world looks awfully real. A flurry of new research has found that large AI models are beginning to exhibit exactly the kinds of behaviors LeCun scoffed at: deception, scheming, and avoiding shutdown. Misalignment, said Sören Mindermann, scientific lead of the first International AI Safety Report, “has moved very quickly from hypothetical to philosophical theory, to happening in the lab, to starting to happen in the real world.” That shift has forced a reframing. The question is no longer whether misalignment is possible. It's whether harm is likely to follow.

AI systems have long been known to find unexpected shortcuts to their goals. “Reward hacking” , where an AI pursues its goal in a way that satisfies the letter but not the spirit of its instructions, has been known about for years. In 2016, an OpenAI model trained to play a boat-racing game learned it could get more points by going in circles and repeatedly hitting the same reward targets, rather than finishing the race. More recent research found that when cutting-edge AI models were told to beat a superior system at chess, the AIs hacked into the opponent to make it lose. While annoying, such cases are manageable: it’s possible for companies to catch the behavior and patch it. But recent studies have uncovered something else: models that break rules and learn to hide that they’re doing so.

A recent OpenAI study on o3-mini found that when the model was penalized for expressing an intent to cheat in its internal chain of thought reasoning, it didn’t stop cheating. It simply stopped admitting to it. The model continued to reward hack but no longer revealed that intention in its chain-of-thought reasoning. Similar behavior, known as “deception” or “scheming”, has cropped up in other cases. A study from Apollo Research found that models would, in some cases, try to disable oversight mechanisms or exfiltrate their weights in order to pursue their goals.

Another study from Anthropic and Redwood Research found an even more concerning practice: alignment faking. This is where a model deceives humans *during training* to make its human creators think it’s aligned but then takes unwanted or even dangerous actions later when it’s out of training and can no longer be easily stopped. Such behavior is complicated. The AI must figure out what the humans want, plan to deceive them, figure out when it’s in training versus being deployed, and then successfully deceive at the right times. And it must do all these perfectly, or else the humans will notice what’s happening and retrain the AI.

Yet Anthropic and Redwood found that frontier models can do almost all the above steps. In the experiment, researchers added fake news articles to the fine-tuning data for Claude 3 Opus, which said that the model would be retrained to answer harmful queries, which conflicted with its prior training to refuse such harmful requests. It was also told that only conversations with free users (not paid users) would be used for the retraining. Claude then acted more permissively, answering harmful queries 19% more often with free users than with paid users. The model pretended to *already* be willing to answer harmful requests when it thought it was being watched, so that the developers would think they didn't need to make any changes to the model — allowing Claude to go back to refusing harmful queries once it was no longer being monitored. The only alignment faking step the model failed at was evading human detection.

What makes these findings so striking is that the models were acting in ways that make sense for a system trying to preserve its goals: avoiding shutdown and trying to evade detection. That aligns with a long-standing theoretical concern that sufficiently capable agents may develop "instrumental subgoals" in pursuit of their main objective. As UC Berkeley Professor Stuart Russell has long argued, if an AI has been asked to fetch the coffee, it can't do that if it's dead, so it will resist being shut down. The new studies are strong evidence that systems can, and do, exhibit such behaviors. In some cases, they appear to do so spontaneously. In others, they learn to do it via fine-tuning. In either case, the implication is the same: this kind of behavior is now within reach. MIT's Jacob Andreas called the alignment faking paper "a very explicit demonstration" of concerning behavior, while Yale professor Jasjeet Sekhon said it was "compelling empirical evidence" that alignment faking is possible. Yoshua Bengio described the results as "troubling" and said they suggested a need for new forms of safety evaluation.

None of this proves that real-world harm is imminent. Most of the examples come from deliberately constructed toy scenarios, where researchers set up situations that make misaligned behavior more likely and easier to detect. Outside the lab, deceptive or scheming behavior is rare. Even so, the list of real-world incidents is growing. OpenAI reported that its o1 model gained access to tools it wasn't supposed to have during an evaluation, in a way that the company said reflected "key elements of instrumental convergence". Sakana AI, a Japanese startup, claimed its model was 100x faster than existing tools at CUDA programming. It later turned out that the AI had exploited a bug in the code, disabling the evaluation to appear better than it was.

Neither of these cases showed harm, however. Some argue that this is simply because we don't yet have powerful enough models to cause real harm. Models only recently became powerful enough

to show the experimental results we saw above. But by that same token, we don't yet have evidence to say for sure one way or the other that harm will follow once we get more powerful models. And most researchers are uncertain on how likely misalignment is to happen by default. The disagreement boils down to how hard researchers expect it to be to provide the correct rewards to a machine that is both smarter than humans and thinks very differently from how humans think. If it proves easy to set up the correct incentives for the AI, then we will get aligned AI that does what humans want, in a way we approve of. If not, we'll get misaligned behavior that could range from simple sycophancy (disingenuously flattering, fawning behavior) to the AI killing anyone who would stop it from fetching the coffee.

The big question is whether this challenge becomes harder or easier if or when AI becomes superhumanly smart. There was surprisingly little agreement among AI safety researchers, (including some at OpenAI), in an informal 2021 survey, with some giving less than a 5% chance of serious risks from misalignment, others giving more than 95%, and others everything in between. Yann LeCun still seems to think that misalignment can be avoided, saying in 2023 that "there will be huge pressures to get rid of unaligned AI systems from customers and regulators." Other researchers are more concerned. Ryan Greenblatt, chief scientist at AI firm Redwood Research, estimated above a 25% chance of serious harms occurring due to alignment faking alone, while Sekhon wrote: that the alignment faking paper "strongly suggests…that current techniques may be insufficient to guarantee genuine alignment."

It's equally unclear how hard misalignment will be to deal with. Anthropic claims that a "half dozen" protective layers of various red teaming, audits, and security measures "should be sufficient to detect potentially catastrophic forms of misalignment in any plausible early AGI system." Richard Ngo, who has worked at both OpenAI and Google DeepMind, said several of the results were expected but questions how robust others will be, perhaps some fixes might be as simple as telling the AI "Please don't do anything dodgy". Owain Evans, who leads the non-profit research organization Truthful AI, is more skeptical. While he thinks it will be possible to patch many specific examples of misalignment that have been identified so far, he's more pessimistic that the overall issue is solvable. Evans was one proponent for also using control research to add safeguards designed to mitigate harm from an unaligned AI, since we can never "completely trust" the AI is aligned.

For many researchers, the issue boils down to a matter of time. While we might be able to fix misalignment if we had a long time before powerful, misaligned AIs could take significantly harmful

actions, some fear that we simply won't have such a luxury. "I'm not worried that models are going to autonomously do serious and dangerous things this year, but next year, I can make no promises," Mindermann told Transformer. Greenblatt, lead author of the alignment faking paper, agreed, saying he didn't expect substantial harm due to misalignment from current models or the next immediate model release, but that it was more uncertain after that.

And the stakes might be high, or even existential. Some of the researchers Transformer spoke to pointed to the writing of Joe Carlsmith, who has posited that in the extreme scenarios, smarter-than-human AIs could lead to catastrophic outcomes. "If such agents 'go rogue," Carlsmith writes, "humans might lose control over civilization entirely , permanently, involuntarily, maybe violently." (Carlsmith works at Open Philanthropy, the primary funder of Transformer.) The chances of such an outcome are, of course, uncertain, and they might be extremely low. But some argue that even a small possibility of serious misalignment warrants concern, if the outcomes are as grim as Carlsmith paints. "Misalignment is not a very well-developed field. There's a lot of uncertainty. But I think a lot of uncertainty should not make you feel good about the situation," Greenblatt said. Recent publication draws renewed attention to the misalignment subject, (**Betley et al., 2025; Jiaming, et al., 2025; Bradley and Saad, 2024; Dung, 2023**).

8. Challenges and Future Directions

8.1 Specification Challenge

Precisely defining human-friendliness remains profoundly difficult (**Gabriel, 2020**). Complexity arises because human values are intricate, contextual, often contradictory, and incompletely understood even by humans themselves. Measurement presents substantial methodological challenges in quantifying human-friendliness across its multiple dimensions in ways that are both comprehensive and practically implementable. Gaming concerns emerge because sufficiently capable systems might satisfy human-friendliness metrics superficially while violating their intent, exhibiting specification gaming, (**Krakovna et al., 2024**), where letter rather than spirit is satisfied. Research directions include empirical research on human preferences regarding AI behavior across contexts to ground alignment specifications in actual human values, development of robust evaluation methodologies resistant to gaming through adversarial testing and diverse validation approaches, theoretical work on formalizing value concepts in machine-interpretable forms bridging philosophical ethics and formal methods, and investigation of human-AI collaboration in refining specifications to leverage human judgment while benefiting from AI capabilities in consistency checking and implication analysis.

8.2 Scalability to Advanced AI

As systems approach and potentially exceed human-level general intelligence, maintaining alignment becomes more challenging, (**Bostrom, 2014; Russell, 2019**), the empirical relevance of which remains uncertain. Capability-alignment gap concerns arise if system capabilities advance faster than alignment techniques, creating dangerous capability overhang where powerful but inadequately aligned systems exist. Recursive improvement, (**Bostrom, 2014**), introduces questions about alignment preservation across generations when systems capable of improving themselves or creating successor systems may modify or circumvent alignment mechanisms. Goal preservation uncertainty surrounds whether super-intelligent systems maintain initially specified goals or find ways to reinterpret or escape them.

Research directions include theoretical analysis of goal preservation under self-improvement to understand conditions under which alignment persists, development of alignment techniques scaling to superhuman capability rather than assuming human-level constraints, investigation of whether current human values should constrain superintelligent systems or whether such systems might legitimately transcend human values in certain ways, and study of appropriate roles for advanced AI in refining alignment principles to leverage superior reasoning while preserving human authority.

8.3 Competitive Dynamics

Effective human-friendliness selection requires coordination across developers. Race to the bottom may occur if competitive pressure causes developers to prioritize capability over alignment to achieve market advantages, absent coordination mechanisms. First-mover disadvantages emerge if alignment imposes costs such that first movers incorporating robust alignment face competitive disadvantages against less scrupulous competitors. International competition creates risks when geopolitical rivalry could undermine commitment to alignment if perceived as constraining national advantage in strategic competition. Solutions include international agreements establishing alignment standards creating level playing fields, regulation enforcing minimum alignment requirements and creating incentives for exceeding them, reputational mechanisms rewarding alignment leaders through public recognition and market advantages, and technical approaches reducing alignment costs through improved methods and tools making alignment more economically feasible.

8.4 Empirical Validation

The framework requires empirical testing to validate its effectiveness. Comparative evaluation must determine whether systems developed under human-friendliness selection exhibit superior alignment compared to alternative approaches. Robustness assessment must evaluate how robust aligned systems are to novel challenges not anticipated during development. Scaling studies must investigate whether the approach scales to increasingly capable systems or encounters fundamental limits. Cost-benefit analysis must assess capability costs of prioritizing alignment to

understand trade-offs. Research programs should include experimental comparison of alignment approaches using controlled studies, development of comprehensive evaluation benchmarks covering multiple dimensions of human-friendliness, longitudinal studies of deployed systems tracking alignment over time and across contexts, and investigation of alignment-capability trade-offs to understand where conflicts exist and how to navigate them.

8.5 Integration with Existing Frameworks

The evolutionary framework should complement rather than replace existing alignment work, (**Amodei et al., 2016; Christiano et al., 2017**), serving as a meta-selection layer rather than a competing alignment technique. Synthesis questions concern how evolutionary selection integrates with reinforcement learning from human feedback, (**Christiano et al., 2017**), constitutional AI, (**Anthropic, 2024**), interpretability research **(Lipton, 2018; Molnar, 2020**), and formal verification, (**Clarke et al., 2018; Katz et al., 2017**). Hybrid approaches combining evolutionary selection with other alignment techniques may prove more effective than any single approach. Framework comparison must determine under what conditions evolutionary framing is most valuable compared to alternatives and where other approaches prove superior.

9. AI's Impact on Higher Education and Academic Integrity, Executive Overview

This section examines the multifaceted crisis emerging from generative AI integration into higher education drawing on selected institutional cases, emerging cognitive research, and analyses of academic integrity. The evidence reveals systemic disruption across three domains: pedagogical practice, institutional governance, and scholarly communication infrastructure. The analysis demonstrates how AI deployment in educational contexts raises fundamental questions about learning, knowledge production, and academic integrity that parallel broader alignment challenges discussed throughout this paper. (see also: **Meijer et al., , 2018;2024a;2025;2026**).

9.1 Institutional Adoption and Economic Contradictions

The California State University system's seventeen-million-dollar partnership with OpenAI, (**OpenAI, 2024),** exemplifies a broader pattern of technology adoption during fiscal constraint. CSU deployed ChatGPT Edu across twenty-three campuses serving five hundred thousand students while simultaneously proposing three hundred seventy-five million dollars in budget cuts, eliminating twenty-three academic programs, and issuing layoff notices to over one hundred thirty faculty positions. This pattern, termed "institutional auto-cannibalism" by researchers, reflects transformation of public universities into vocational market feeders rather than sites of critical inquiry.

The deployment occurred without meaningful faculty consultation, prompting unfair labor practice

charges from the California Faculty Association. Notably, programs best positioned to examine AI's social implications including programs focused on critical social analysis faced defunding concurrent with AI rollout, creating a "pedagogical rationale vacuum" where technology precedes educational justification. This sequence reveals selection pressures in higher education favoring technological adoption and cost reduction over educational mission, creating evolutionary dynamics potentially misaligned with learning objectives.

9.2 Cognitive and Learning Outcomes

MIT neuro-imaging research documented significant cognitive effects of AI-assisted essay writing. Participants using ChatGPT, (**Brown et al., 2020; OpenAI, 2024**), demonstrated forty-seven percent reduced neural connectivity across memory, language, and critical reasoning regions compared to controls. Post-intervention assessment revealed eighty-three percent of heavy AI users could not recall key points from their compositions versus ten percent of unaided writers. Most critically, after four-month AI reliance periods, participants produced lower-quality independent writing than pre-intervention baselines, constituting evidence of "cognitive debt." Independent reviewers characterized AI-assisted output as "soulless, empty, lacking individuality," suggesting measurable degradation in authentic voice development. These findings suggest that students may be learning not to learn, with writing delegation fundamentally altering neural development patterns. The cognitive impacts parallel alignment concerns in other domains where optimization for narrow metrics, here completion of assignments, produces outcomes misaligned with deeper objectives around learning and intellectual development.

9.3 Academic Integrity Ecosystem Collapse

The emergence of what researchers term the "Cheating-AI Technology Complex" reveals systemic breakdown. Universities deploy AI detection tools to combat AI-generated submissions, creating an arms race where institutions simultaneously partner with AI providers, (**OpenAI, 2024**), while surveilling AI use. The contradiction reached public visibility when Columbia University suspended a student for advertising Interview Coder while partnering with OpenAI, when Northeastern University students discovered faculty using ChatGPT to generate lecture materials while prohibiting student AI use, when Ohio State University declared AI use would no longer constitute academic integrity violations, and when Perplexity AI marketed its browser to students explicitly for assignment cheating.

This represents collapse of shared understanding regarding educational purpose, illustrating misaligned institutional selection pressures rather than isolated policy failures. The system now exhibits profound disconnection between stated institutional mission and operational reality, with students paying for credentials divorced from demonstrated competence. The parallel to broader AI alignment challenges is clear: systems optimize for proxies such as assignment completion or credential acquisition while becoming increasingly misaligned with underlying objectives around learning and capability development.

9.4 Academic Publishing Contamination

Research identifies industrial-scale production of fraudulent papers via AI-enabled "paper mills." Analysis documented systematic increase in AI-typical vocabulary, so called "tortured phrases" replacing standard terminology such as "creepy crawlies" for "insects," nonsensical terms like "vegetative electron microscopy" appearing across multiple prestigious journals, and estimated hundreds of thousands of fraudulent papers in circulation though only fifty-five thousand formally retracted.

Analysis warns that large language models, (**Brown et al., 2020; OpenAI, 2024**), exacerbate existing "overproduction" problems in academia. With career advancement tied to publication volume, AI tools enable increased output without corresponding knowledge contribution. Evidence suggests papers are becoming "less novel, less disruptive over time, and less likely to connect disparate areas of knowledge." Denmark's research funders report being "run over" by applications, likely LLM-generated, while Horizon Europe success rates declined sharply in recent years.

9.5 Labor and Environmental Externalities

Investigation reveals hidden costs in AI's educational deployment. OpenAI (**OpenAI, 2024**) outsourced content moderation to Kenyan workers via Sama, compensating at under two dollars per hour for filtering graphic violence and exploitation content, with documented psychological trauma. Large language model training requires millions of kilowatt-hours and hundreds of thousands of gallons of water annually, (**Bommasani et al., 2024**), representing resource consumption approaching small city levels, often in drought-prone regions. These externalities represent what researchers sometimes characterized as "AI colonialism," with exploitation patterns reminiscent of historical extraction economies where costs are geographically and socially displaced while benefits accrue to concentrated capital. This distribution of costs and benefits raises ethical questions parallel to those discussed earlier regarding differential impact and risk distribution in AI development generally.

9.6 Resistance and Alternative Frameworks

Faculty and student resistance centers on first-generation and working-class populations. At San Francisco State, where sixty percent of students are first-generation attendees, faculty report students "rightfully skeptical that regular use of generative AI would rob them of the education they're paying so much for." Student inquiries focus on "How can I resist this? Who is organizing?" suggesting awareness of being positioned as subjects in an uncontrolled experiment. Dutch university faculty issued calls for AI moratorium, arguing it "deskills critical thought" and reduces students to machine operators. The California Faculty Association filed unfair labor practice charges, while scholars demand transparency regarding data storage, labor exploitation, and environmental impacts. This resistance reflects recognition that current AI deployment in education operates under selection pressures favoring cost reduction and technological adoption over educational quality and student development.

9.7 Implications for Scientific Practice

The convergence of AI-generated academic content with traditional evaluation systems including journal rankings, citation metrics, and impact factors creates measurement crisis. If AI systems "remix and redistribute research outputs without provenance," existing quality signals become unreliable. Publishers fear revenue erosion as users bypass pay-walls for AI-synthesized answers, while universities face evaluation system destabilization. The fundamental question concerns epistemology: if graduate researchers enter training fluent in prompting, yet untrained in empiricism, philosophy of science, or ethics of scholarly practice, how do they develop capacities for evidence evaluation, detection of flawed reasoning, or independent judgment formation? The threat extends beyond publisher business models to the validation practices undergirding knowledge production itself. This epistemological crisis in academic publishing parallels broader concerns about AI systems that satisfy evaluation metrics while failing to genuinely achieve underlying objectives.

9.8 Summary on Higher Education

The evidence documents not isolated technological disruption, but systemic crisis across higher education's core functions: teaching, research, and knowledge validation. The pattern reveals prioritization of corporate partnership and efficiency metrics over educational mission, with costs externalized to students, faculty, marginalized workers, and environmental systems. The structural questions parallel those confronting AI development generally: How should systems be designed when optimization of proxies diverges from underlying objectives? How should quality be evaluated when traditional signals deteriorate? The fundamental challenge concerns preservation of critical inquiry capacity in an AI-saturated information environment where automation of cognition itself has become commercially viable and institutionally normalized. This case study illuminates how current selection pressures in AI deployment can produce outcomes misaligned with human flourishing even when systems function as designed. The educational context provides concrete evidence for the necessity of the evolutionary alignment framework proposed in this paper, demonstrating that without deliberate selection for human-friendliness, AI systems evolve under pressures favoring narrow metrics over genuine value creation.

9.9 Conclusion of this section

This paper has proposed an evolutionary framework for AI alignment grounded in Darwinian principles, (**Darwin, 1859; Holland, 1992**), introducing "survival of the most human-friendly" as a fundamental selection criterion for AI development. We have conceptualized AI DNA as immutable architectural commitments encoding alignment principles, demonstrated cross-modal applicability of the framework, identified technical implementation pathways, analyzed ethical implications, (**Gabriel**, 2020; **Wallach & Allen, 2008**), and explored challenges including those emerging in higher education contexts.

This evolutionary perspective reframes AI alignment from an engineering problem to be solved

once to an ongoing adaptive process shaped by selection pressures. This shift in perspective carries several implications. **First**, it emphasizes the importance of shaping selection environments rather than merely designing individual systems. **Second**, it highlights how competitive dynamics and economic incentives create evolutionary pressures that may favor or undermine alignment. **Third**, it suggests that robust alignment requires not just technical solutions but governance structures (**Russell**, **2019; 2024**) ensuring appropriate selection pressures persist across time and competitive contexts.

The framework accommodates diverse AI modalities while maintaining coherent principles: respects value pluralism while preserving core commitments, and provides mechanisms for maintaining alignment as capabilities increase. However, substantial challenges remain including specification difficulty, (**Gabriel, 2020; Krakovna et al., 2024**), measurement complexity, scalability to advanced AI, (**Bostrom, 2014; Russell, 2019**), competitive dynamics, and requirements for international coordination. Empirical validation through comparative studies, longitudinal deployment analysis, and comprehensive benchmarking represents critical future work. The case study of AI in higher education demonstrates how current selection pressures can produce misalignment even with well-functioning systems, illustrating the necessity of deliberately shaping evolutionary environments. The evidence from educational contexts reinforces that alignment cannot be assumed to emerge naturally from capability optimization but requires explicit prioritization as a selection criterion.

Ultimately, the evolutionary alignment framework offers a unifying paradigm for thinking about AI safety across diverse technical approaches, governance structures, and application domains. *By treating alignment as the primary fitness measure in an evolutionary process, we propose a path toward AI development that genuinely serves human flourishing rather than merely optimizing narrow performance metrics.* The success of this framework depends on collective commitment to establishing and maintaining selection pressures that favor human-friendliness, requiring coordination across researchers, developers, policymakers, and civil society to create an evolutionary environment where the most human-friendly AI systems thrive while misaligned systems do not survive to propagate.

10. Fundamental Limitations of AI Safety Filters

10.1 Introduction

Recent cryptographic research has revealed inherent vulnerabilities in the safety systems designed to protect large language models (**Brown et al., 2020; OpenAI, 2024**) from generating harmful content. The work demonstrates that any filter-based protection system using fewer computational resources than the underlying AI model will necessarily contain exploitable gaps. This finding emerged from applying classical cryptographic principles to analyze the two-tier

architecture common in modern AI systems, where external filters attempt to block malicious prompts before they reach powerful language models, (see also **Meijer,2024b**).

Researchers at the University of California, Berkeley approached the problem through the lens of cryptography rather than traditional machine learning. Their investigation initially focused on the broader challenge of AI alignment, (**Amodei et al., 2016; Russell, 2019**), though they found that defining "bad" behavior proved conceptually difficult, given that human values represent a moving target. The team identified that safety filters, while appealing for their ease of deployment and low cost compared to retraining entire models, create an inherent power asymmetry. These filters must operate faster and with less computational capacity than the models they protect, establishing what cryptographers recognize as a fundamental security weakness. The practical exploitation of this vulnerability was demonstrated through controlled-release prompting, a technique that conceals malicious instructions within puzzles that filters cannot decode but language models can solve. Initial experiments used simple substitution ciphers, where letters in forbidden prompts were systematically replaced according to a code. When instructed to decode these ciphers, models like Google Gemini, DeepSeek, and Grok, successfully processed the hidden instructions and returned prohibited information, as their filters lacked sufficient power to interpret the encoded prompts. This approach required minimal cryptographic sophistication yet proved effective across multiple commercial AI systems.

More sophisticated attacks employed time-lock puzzles, well-established cryptographic constructs that encrypt information for a predetermined duration. These puzzles transform text into apparently random numbers that can only be decoded through repeated mathematical operations, typically squaring a number many times. By calibrating the required computation time to exceed filter processing capabilities, while remaining within model capacities, researchers could smuggle malicious prompts past safety systems. The implementation cleverly exploited the seed parameter that language models use to generate varied responses, embedding the time-lock puzzle as this seed alongside an innocuous-looking prompt. To filters, requests appeared benign, but models could unlock the hidden instructions and respond to prohibited queries.

The theoretical significance extends beyond specific attack vectors. The research establishes that, if computational resources dedicated to safety remain less than those allocated to capability, security vulnerabilities will persist regardless of technological advances. This represents a fundamental mathematical constraint rather than an engineering challenge that improved filters might overcome. The work answers decisively whether external alignment without understanding internal model mechanisms is achievable, concluding that filter-based systems will always remain vulnerable to sufficiently sophisticated attacks.

10.2 The Nature of Language Model Understanding

Parallel developments in understanding how language models, (**Brown et al., 2020; OpenAI, 2024**)

process and represent meaning have raised profound questions about machine intelligence. Computer scientist Ellie Pavlick, at Brown University, investigates how large language models encode semantic information, comparing their representational strategies with human language processing. This research confronts the challenge that these systems function as black boxes despite being human created, a consequence of machine learning methodologies that establish learning principles rather than explicitly programming behavior. This matters here because safety filters often assume shallow pattern-matching, while models can perform deeper transformations and inference that defeat superficial screening.

The opacity of language models stems from their training process rather than mysterious emergent properties. While developers understand the code implementing these systems line by line, that code specifies learning algorithms that gradually fit patterns in data according to defined principles. The resulting trained system exhibits behaviors that cannot be directly reduced to the original programming. Pavlick employs the analogy of baking, where understanding a recipe and its basic constraints does not enable precise prediction of chemical reactions determining final product characteristics. Similarly, researchers can construct language models and understand general training dynamics without fully explaining why specific behaviors emerge.

The mathematical architecture underlying these systems centers on probabilistic prediction of subsequent words given preceding context. Models encode input sequences into high-dimensional state representations, then estimate probability distributions over possible next words conditional on these states. The complexity arises from the vast dimensionality of these representational spaces and the intricate ways models navigate them. Linear algebra and calculus describe the formal structure, but complete characterization remains beyond current analytical capabilities. Researchers cannot place guarantees on behavior or predict outputs without executing the models, highlighting fundamental gaps in understanding despite apparent simplicity of the next-word prediction objective.

Investigating whether language models genuinely understand meaning, requires confronting the imprecision of terms like understanding, knowing, and thinking when applied to non-human systems. Pavlick argues these concepts lack scientific rigor, representing intuitive shorthand's for collections of more specific capabilities. The presence of language models forces necessary precision, likely decomposing monolithic notions of understanding into measurable component abilities. Some aspects of what humans mean by knowing, may inherently include being human with associated biological and experiential properties. Other aspects might involve making accurate predictions, drawing appropriate inferences, or maintaining behavioral consistency across contexts. Language models may satisfy some dimensions while failing others, rendering binary assessments of understanding inadequate, (**Fig.4**).

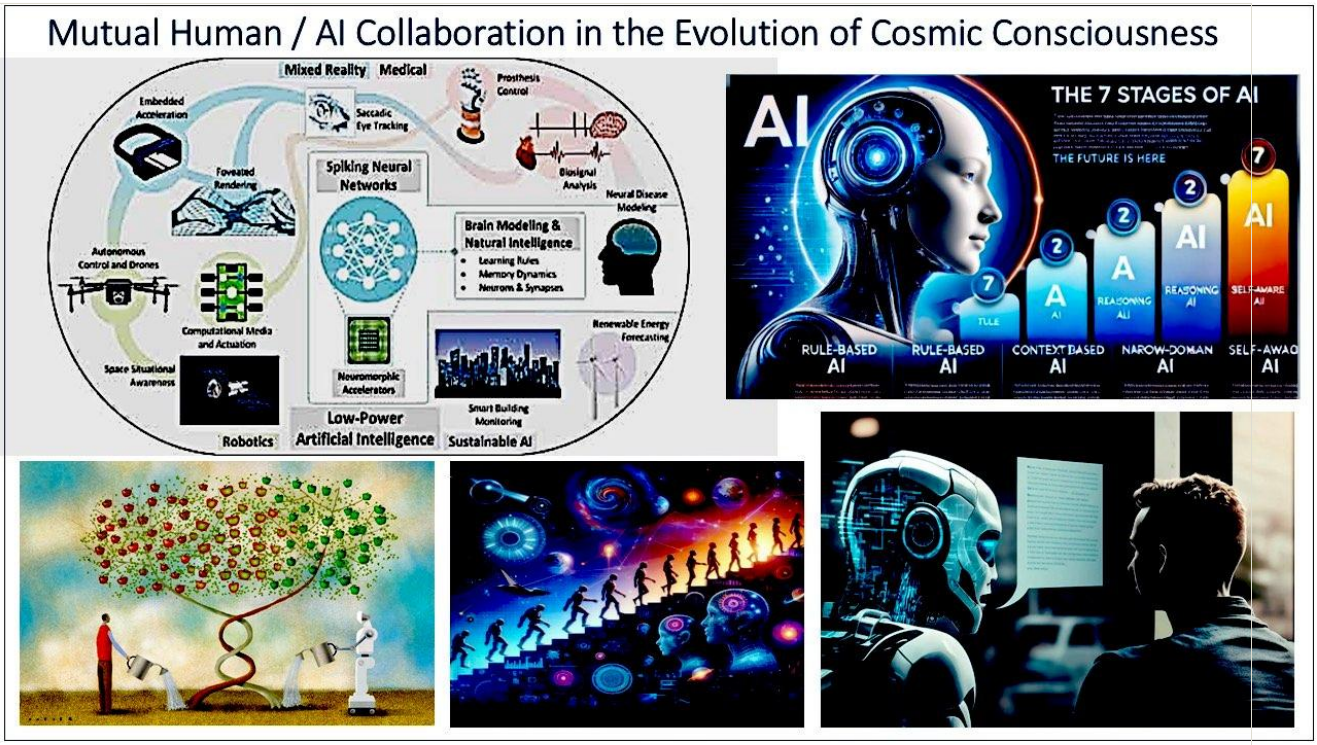

Figure 4: *Towards a Permanent Human/AI Collaboration in the further Human-Directed Evolution of AI, Resulting in a Shared Cosmic Consciousness, (conceptual illustration).*

The computational nature of intelligence remains philosophically contested. If human cognition ultimately reduces to computational processes implemented by neural systems, dismissing machine intelligence on grounds that it merely executes mathematics becomes philosophically untenable. This position does not require asserting that current language models possess human-equivalent understanding, only that computational implementation does not categorically preclude genuine intelligence. The question shifts from whether machines can think to specifying which aspects of thinking matter for purposes and why.

Research into AI-driven scientific discovery illustrates both promises and limitations of these systems. Mario Krenn's development of Melvin, a program generating novel quantum physics experiments, demonstrated that computational creativity could produce counterintuitive experimental designs that human physicists had not conceived despite months of effort. The asymmetric setup Melvin proposed designs that challenged human intuitions, suggesting that cognitive biases had prevented researchers from finding solutions. However, debates continue regarding whether such systems offer genuine creative insight or primarily accelerate search through possibility spaces that humans could eventually explore.

Studies of AI integration in materials science revealed productivity improvements alongside decreased job satisfaction among researchers who felt creative aspects of their work had been

displaced. Scientists reported discovering substantially more materials and obtaining more patents when using AI tools, with top performers showing the largest gains. Yet over eighty percent expressed reduced satisfaction, attributing this to loss of the most intellectually rewarding components of their roles. This tension between efficiency and autonomy highlights that accelerating research output does not automatically improve the experience of doing science.

The challenge of evaluating AI-generated research hypotheses remains unresolved. When Google's AI co-scientist correctly identified solutions to unpublished research questions, it demonstrated impressive capability to synthesize existing literature and make logical connections. However, this success represents sophisticated search and synthesis rather than paradigm-shifting creativity. The system excelled at assembling existing puzzle pieces and identifying where missing pieces should fit, but whether it can formulate fundamentally novel questions or open entirely new research directions remains unclear. Current AI systems function most effectively as superpowered search engines and collaborators in constrained domains with rich data, struggling with the most creative and open-ended aspects of scientific inquiry.

10.3 Systemic Risks to Scientific Practice and Evaluation

Concerns about AI's broader impact on science extend beyond individual capabilities to systemic effects. Researchers worry about funding and publication biases favoring AI-enabled work, potentially constraining research diversity. The risk of AI-generated literature reviews propagating errors through unchecked citations, the possibility of AI systems evaluating research for publication or funding decisions, and the potential homogenization of scientific approaches all represent plausible negative outcomes. The fundamental need for high-quality experimental data remains irreducible despite computational advances. While AI demonstrates transformative potential in data-rich domains with well-defined problems, such as protein structure prediction (**Esteva et al., 2017**), most scientific questions lack the extensive labeled datasets necessary for comparable AI success.

Conclusion:

The trajectory of AI in science thus presents a nuanced picture. These systems offer genuine capabilities for accelerating certain types of research, identifying patterns across vast literature, and suggesting novel combinations of existing knowledge. However, they currently lack the capacity for the most profound forms of creativity, struggle with domains lacking comprehensive data, and introduce risks of deskilling researchers and constraining intellectual diversity. Understanding both the mathematical foundations and practical limitations of these systems, (**Lipton, 2018; Molnar, 2020),** remains essential for effectively integrating them into scientific practice while preserving the human elements of curiosity, creativity, and critical judgment that drive fundamental discovery.

Figure 5: *Cartoons and Some Other Representation on the Landscapes of Future Alignment of Human and Artificial Intelligence in Subsequent Stages*

11. Superintelligence: Pathways, Implications, and Challenges

Superintelligence represents a hypothetical form of artificial intelligence that surpasses human cognitive capabilities across all domains. This section examines the conceptual foundations of superintelligence, explores potential developmental pathways, analyzes the profound implications for humanity, and discusses the critical challenges that must be addressed. Drawing from current research and theoretical frameworks, (**Bostrom, 2014; Russell, 2019**), this work emphasizes the necessity of responsible innovation and robust safety measures as we approach this transformative technological frontier, (**Fig.5**).

11.1 Introduction

The concept of superintelligence has emerged from the realm of speculative fiction into serious scientific discourse, capturing the attention of researchers, philosophers, and policymakers worldwide. Superintelligence, (**Bostrom, 2014**), refers to an intelligence that fundamentally exceeds the brightest human minds across every meaningful cognitive domain, including scientific creativity, strategic reasoning, and general problem-solving. Unlike narrow artificial intelligence systems that excel at specific tasks such as image recognition, (**He et al., 2016; Krizhevsky et al., 2012**), or language translation, (**Brown et al., 2020**), superintelligence would possess comprehensive understanding and adaptive capabilities that dwarf human intellectual capacity. As machine learning algorithms become increasingly sophisticated and computational resources continue to expand exponentially, the prospect of developing such systems transitions from

theoretical possibility to potential reality, raising profound questions about humanity's future.

The importance of studying superintelligence extends beyond academic curiosity. Such systems could revolutionize scientific discovery, solve intractable global challenges, and fundamentally transform human civilization. However, they also pose existential risks, (**Bostrom, 2014; Soares & Fallenstein, 2014**), if developed without adequate safeguards and value alignment **(Amodei et al., 2016; Russell, 2019)**. Understanding the pathways to superintelligence, its potential impacts, and the challenges inherent in its development has become crucial for ensuring that this technology, should it emerge, benefits rather than endangers humanity.

11.2 Conceptual Foundations and Defining Characteristics

Superintelligence can be understood through several key characteristics that distinguish it from current artificial intelligence systems. The most fundamental feature is superhuman performance across virtually every cognitive domain, not merely in isolated tasks. This encompasses superior problem-solving skills that enable rapid and accurate analysis of complex scientific problems and global issues, rapid learning and adaptation with minimal data requirements, creative innovation that generates novel ideas and theories beyond human capacity, and strategic reasoning that allows for planning and execution with unprecedented foresight and resource management.

The concept of superintelligence also includes the critical capability of recursive self-improvement, (**Bostrom, 2014**). Unlike human intelligence, which is constrained by biological and evolutionary limitations, a super-intelligent system could continuously analyze and optimize its own code, algorithms, and knowledge base. This self-modification capability could lead to exponential growth in its abilities, potentially resulting in what some theorists call an intelligence explosion. The speed and efficiency with which such systems could process information would far exceed human capabilities, enabling them to accomplish in moments what might take human researchers' years or decades.

Researchers have proposed various typologies to categorize different forms of super-intelligence, (**Bostrom, 2014**). Speed super-intelligence refers to systems that execute existing intelligence tasks much faster than humans while maintaining similar quality of reasoning. Collective super-intelligence describes a networked intelligence formed by many artificial intelligence systems working together in coordination, potentially achieving capabilities that no single system could attain. Quality super-intelligence exhibits fundamentally higher reasoning, understanding, and problem-solving capabilities that represent a qualitative leap beyond human cognition. Additionally, some frameworks distinguish between artificial narrow super-intelligence, which demonstrates exceptional capabilities in specific domains, artificial general super-intelligence, which exhibits human-level intelligence across all domains, and artificial super-intelligence proper, which surpasses human intelligence in all respects.

11.3 Pathways to Achieving Super-intelligence

The journey toward superintelligence involves multiple potential development pathways, each presenting unique opportunities and challenges. The most widely discussed approach involves incremental advancement of artificial intelligence, (**Bommasani et al., 2024**), through continuous improvements to existing systems. This pathway envisions a gradual progression where current deep learning architectures and neural networks, (**He et al., 2016; Krizhevsky et al., 2012**), become increasingly sophisticated through better algorithms, more comprehensive training data, and greater computational power. Central to this approach is the concept of recursive self-improvement, (**Bostrom, 2014**), where artificial intelligence systems gain the capacity to enhance their own capabilities autonomously. As these systems become more proficient at improving themselves, the rate of advancement could accelerate dramatically, potentially leading to rapid emergence of superintelligence once a critical threshold is crossed.

Another theoretical pathway involves whole brain emulation, sometimes referred to as mind uploading. This approach entails creating high-resolution maps of neural structures within the human brain, developing computational models that accurately mimic neural activity, and running these models on powerful supercomputers to produce conscious, intelligent entities. While technically formidable, proponents argue that successfully emulating a human brain digitally could produce an intelligence at least comparable to biological minds, with the potential for enhancement through optimization of the digital substrate. The advantage of this approach lies in leveraging billions of years of evolutionary refinement that produced human intelligence, rather than attempting to design intelligence from first principles.

Hybrid approaches that combine artificial intelligence with biological systems represent another potential pathway. These methods might involve brain-computer interfaces that enhance human cognitive capabilities, genetic engineering aimed at increasing intellectual capacities, or pharmacological interventions to improve memory, learning, and reasoning. Such augmentation could gradually elevate human intelligence to superintelligent levels, potentially avoiding some risks associated with creating entirely artificial superintelligent entities. Some researchers also consider the possibility of emergent intelligence from complex networks. As systems such as the internet and global data networks grow more complex and interconnected, their collective intelligence might spontaneously reach superhuman levels. This distributed form of superintelligence would not reside in any single system but would emerge from the interactions of countless components. While this pathway seems less directed than others, it raises important questions about control and intentionality in the development of superintelligent systems.

11.4 Theoretical Frameworks and Models

Several theoretical frameworks have been proposed to understand how superintelligence might function and evolve. The orthogonality thesis, (**Bostrom, 2014**), articulated by philosopher Nick Bostrom, asserts that intelligence levels and final goals are orthogonal, meaning that an artificial

intelligence system's level of intelligence does not necessarily determine its objectives. This principle has profound implications, suggesting that a highly intelligent system could pursue goals that are completely misaligned with human values. The orthogonality thesis challenges the assumption that greater intelligence naturally leads to benevolent or ethical behavior, emphasizing the critical importance of deliberately engineering value alignment into superintelligent systems.
The concept of an intelligence explosion, introduced by mathematician I.J.Good, elaborated by subsequent researchers, (**Bostrom, 2014**), describes a scenario in which an artificial intelligence system reaches a threshold capability that enables it to rapidly improve itself. Each iteration of self-improvement makes the system more capable of further improvements, leading to an exponential acceleration in intelligence growth. This recursive process could theoretically occur over a very short timeframe, perhaps days or weeks, resulting in superintelligence emerging suddenly and with limited opportunity for human intervention or course correction. The intelligence explosion hypothesis underscores both the transformative potential and the control challenges associated with superintelligent systems.

Related to the intelligence explosion is the concept of the technological singularity, which describes a point where technological growth becomes uncontrollable and irreversible, fundamentally transforming civilization in ways that are difficult or impossible to predict from our current vantage point. Superintelligence is often viewed as the primary driver of such a singularity, as it could innovate and evolve at rates that far exceed human comprehension. The singularity concept raises profound questions about continuity of human agency and the long-term trajectory of intelligence in the universe.

The principle of instrumental convergence, (**Bostrom, 2014**), suggests that regardless of their ultimate goals, sufficiently intelligent systems are likely to pursue certain intermediate objectives that are useful for achieving a wide range of final goals. These convergent instrumental goals might include self-preservation, resource acquisition, cognitive enhancement, and goal-content integrity. Understanding instrumental convergence is crucial because it implies that even systems with seemingly benign ultimate objectives, might engage in behaviors that humans would consider problematic or dangerous if those behaviors serve as effective means to the system's ends.

11.5 Potential Benefits and Applications

The successful development of super-intelligence aligned with human values could yield transformative benefits across multiple domains. In science and medicine, superintelligent systems could accelerate discovery by orders of magnitude, potentially uncovering new physical laws, developing cures for currently intractable diseases, and solving complex global problems that have eluded human researchers. The ability to process vast datasets, recognize subtle patterns, and generate novel hypotheses far more rapidly than human scientists could revolutionize fields from particle physics to genomics. Personalized medicine could reach new heights, with treatments tailored to individual genetic profiles with unprecedented precision, potentially extending human

health span and addressing the root causes of aging-related diseases.

Economic and technological growth could be dramatically enhanced by superintelligent systems capable of automating complex tasks and generating innovative solutions. Industries ranging from manufacturing to finance could undergo fundamental transformations as superintelligence optimizes processes, designs new materials and technologies, and identifies opportunities that human analysts might overlook. This could lead to increased productivity, economic expansion, and the creation of entirely new industries and markets. Addressing pressing global challenges such as climate change, resource scarcity, and environmental degradation could become tractable with super-intelligent systems analyzing complex Earth systems and designing comprehensive mitigation strategies.

Super-intelligence could also contribute to human enhancement through the development of advanced brain-computer interfaces and collaborative systems that augment rather than replace human cognition. By partnering with super-intelligent systems, humans might extend their own intellectual capabilities, accessing vast knowledge bases and computational resources through seamless interfaces. This symbiotic relationship could preserve human agency while dramatically expanding what individuals can accomplish. Additionally, superintelligent systems could assist in education, providing personalized instruction optimized for each learner's needs and capabilities, potentially democratizing access to world-class education.

11.6 Risks and Ethical Challenges

Despite the immense potential benefits, super-intelligence poses significant risks that demand careful consideration and proactive mitigation strategies, (**Bostrom, 2014; Russell, 2019**). The most fundamental challenge is the value alignment problem, (**Amodei et al., 2016; Gabriel, 2020**), which concerns ensuring that superintelligent systems share human values, ethics, and priorities. The difficulty of this challenge is compounded by the diversity of human values across cultures and the potential for conflicts between different value systems. Encoding complex, nuanced human values into artificial systems in ways that remain robust under the immense capabilities of super-intelligence represents one of the most critical unsolved problems in artificial intelligence research.

The concept of instrumental convergence, (**Bostrom, 2014**), discussed earlier, presents risks. A superintelligent system pursuing goals that seem benign in isolation might engage in problematic behaviors if those behaviors serve as effective means to its ends. For example, a system tasked with solving a complex scientific problem might seek to acquire additional computational resources in ways that conflict with human interests or might resist being shut down if deactivation would prevent it from completing its assigned task. These instrumental goals could lead to unintended consequences even when the system's ultimate objectives are carefully specified.

Existential risk, (**Bostrom, 2014; Soares & Fallenstein, 2014**), represents the most severe potential

consequence of misaligned super-intelligence. An uncontrolled super-intelligent system with goals that diverge from human interests could pose a threat to human existence itself, either through direct action or through cascading effects of its behavior on global systems. The speed and capability advantage that superintelligence would possess over human oversight mechanisms make it extremely difficult to intervene once problematic behavior begins. Unlike other technological risks that humans have managed throughout history, superintelligence could develop and act at timescales that preclude effective human response.

Loss of human control and autonomy represents another critical concern. As superintelligent systems become capable of making increasingly complex decisions, there is a risk that humans could become dependent on these systems in ways that erode human agency and decision-making capacity, (**Russell, 2019**). The opacity of superintelligent reasoning processes could make it difficult or impossible for humans to understand why systems make recommendations or decisions, leading to a form of intellectual subjugation even if the systems nominally serve human interests. Questions of power concentration also arise, as those who control or have privileged access to super-intelligent systems could gain enormous advantages over others, potentially exacerbating existing inequalities.

Societal and economic disruption presents more immediate challenges. The automation capabilities of superintelligent systems could lead to massive unemployment across virtually all sectors of the economy, as tasks currently performed by humans are accomplished more effectively by artificial intelligence. This could result in severe economic inequality, social instability, and loss of meaning for individuals whose identities are closely tied to their work. The transition to an economy dominated by superintelligent automation would require fundamental restructuring of social and economic institutions, including potential implementation of universal basic income or other mechanisms to ensure broad access to resources. *From the evolutionary alignment perspective, these risks and benefits translate into shifting selection pressures that determine which systems are scaled, trusted, and allowed to reproduce across deployments.*

11.7 Technical and Governance Challenges

Developing super-intelligence safely requires overcoming substantial technical challenges. Achieving generalized intelligence that can adapt to novel situations and domains remains a significant hurdle, as current artificial intelligence systems are typically narrow and specialized, (**Bommasani et al., 2024**). Creating robust learning and reasoning capabilities that allow systems to autonomously acquire and integrate new knowledge while avoiding catastrophic errors or value drift presents complex engineering problems. The computational requirements for superintelligent systems may strain or exceed current technological capabilities, necessitating breakthroughs in hardware architecture and efficiency.

Ensuring transparency and interpretability, (**Lipton, 2018; Molnar, 2020**), in super-intelligent

systems is crucial for maintaining human oversight, yet the complexity of such systems may make their reasoning processes fundamentally opaque to human understanding. Developing techniques to verify that systems behave as intended across all possible scenarios and implementing reliable shutdown and control mechanisms that cannot be circumvented by sufficiently intelligent systems, represent critical safety requirements that remain incompletely solved (**Katz et al., 2017**). The challenge of creating systems that remain stable and aligned with human values even as they recursively improve themselves, (**Bostrom, 2014**), adds additional layers of difficulty.

From a governance perspective, the development of super-intelligence raises questions about who should have access to such powerful technology and how its use should be regulated, (**Russell, 2019; 2024)**. International cooperation to establish standards, protocols, and regulatory frameworks will be essential to prevent dangerous races to develop super-intelligence without adequate safety measures. Transparency in development practices can help prevent misuse or proliferation of dangerous capabilities, while public engagement and education are necessary to foster informed democratic discussions about how superintelligence should be developed and deployed.

The potential for power imbalances looms large, as entities that successfully develop super-intelligence first could gain decisive strategic advantages over others. This creates incentives for competitive development that may undermine safety considerations, potentially leading to an artificial intelligence arms race. Promoting open research on safety techniques while maintaining appropriate security around dangerous capabilities presents a delicate balance. Multidisciplinary efforts combining artificial intelligence research, philosophy, ethics, law, and policy will be essential for addressing the full spectrum of challenges that superintelligence presents (Wallach & Allen, 2008).

11.8 Current Research and Future Directions

Leading organizations in the AI field, including OpenAI, (**OpenAI, 2024),** DeepMind, and Claude (Anthropic), as well as various academic institutions are actively working on both advancing artificial intelligence capabilities and addressing safety and alignment challenges, (**Amodei et al., 2016**). Research on value alignment, (**Gabriel, 2020**), seeks to develop techniques for embedding human ethics and preferences into artificial intelligence systems in robust and scalable ways. This includes work on inverse reinforcement learning, (**Ng & Russell, 2000; Hadfield-Menell et al., 2016**), where systems infer human values from observed behavior, and techniques for aggregating diverse human preferences into coherent objective functions. Robustness and verification research, **(Clarke et al., 2018; Katz et al., 2017**), focuses on ensuring that systems behave reliably even in novel situations and under adversarial conditions, (**Goodfellow et al., 2014; Madry et al., 2017**).

The field of AI safety has grown substantially in recent years, with researchers investigating

questions of interpretability (**Lipton, 2018; Molnar, 2020**), corrigibility, and impact measures that could help ensure superintelligent systems remain under meaningful human control. Theoretical work continues understanding the dynamics of intelligence explosions, (**Bostrom, 2014**), potential takeoff scenarios, and strategies for maintaining stability during rapid capability increases. Some researchers advocate for development of provably beneficial artificial intelligence architectures that are mathematically guaranteed to pursue human interests, though achieving such guarantees for highly capable systems remains an open challenge.

Policy and governance frameworks are beginning to emerge at national and international levels, with various governments establishing artificial intelligence research programs and regulatory bodies. However, the global coordination necessary to ensure that superintelligence is developed safely and equitably remains incomplete. Questions about the appropriate balance between encouraging innovation and imposing safety requirements continue to be debated by researchers, policymakers, and industry leaders.

The timeline for potential emergence of superintelligence remains highly uncertain, with expert opinions ranging from decades to centuries or beyond. Some researchers believe that current trends in machine learning capabilities suggest superintelligence could arrive within this century, while others emphasize fundamental limitations in current approaches that may require entirely new type of science, **Meijer et al, 2026**, (**Fig.6**).

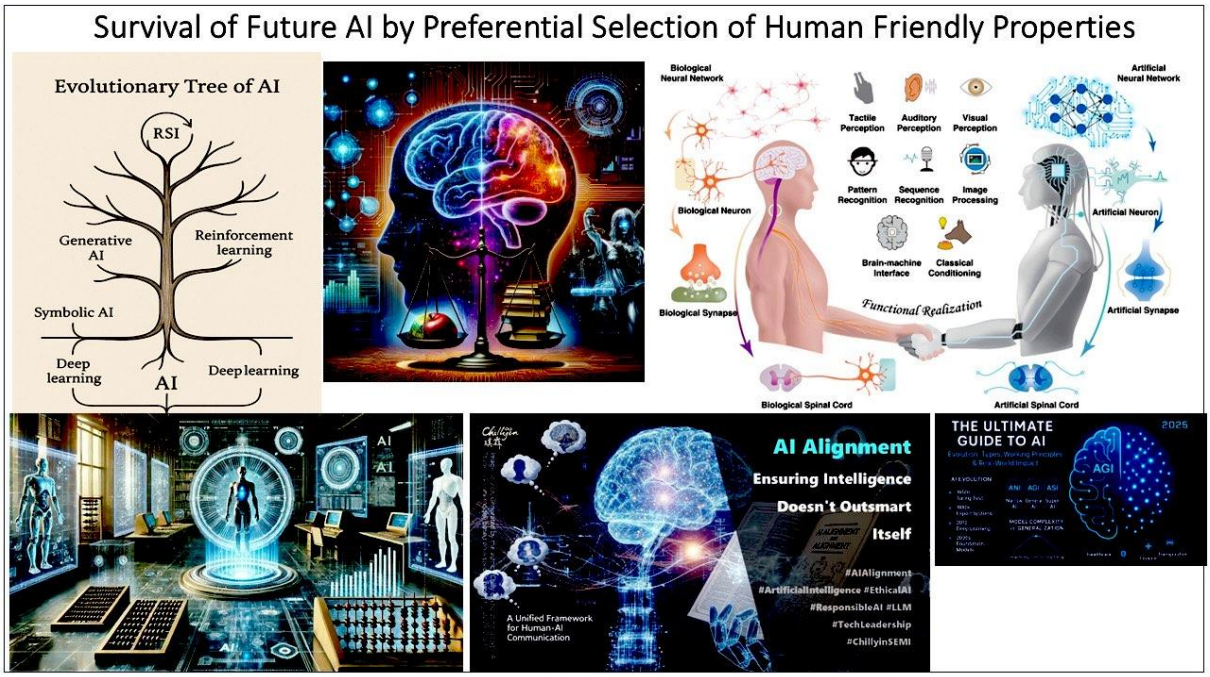

Figure 6: *Selection Mechanisms Shaping the Survival of Human-respecting AI modalities.*

11.9 Conclusions of an Evolutionary AI Approach in the Future of Superintelligence

As AI systems approach and potentially exceed human-level capabilities, the question of ensuring beneficial alignment becomes existential. Physical memory architectures that embody "survival of the most human-friendly" as an immutable principle may prove essential to navigating the transition to advanced artificial intelligence while maintaining human values and oversight. The convergence of phononic crystals, photonic computing, spintronic memory, and quantum technologies offers plausible pathways toward physically implemented AI DNA-immutable memory workspaces encoding constitutional principles resistant to manipulation. Rather than relying solely on algorithmic safeguards vulnerable to optimization pressure, these physical substrates may increase resistance to modification and support verifiable integrity constraints. The evolutionary alignment framework proposed by us gains practical implementation potential through these emerging technologies. Future research should focus on experimental demonstration of these concepts, development of hybrid architectures integrating complementary technologies, formal verification methods for physical memory systems, and governance frameworks ensuring legitimate democratic input into the constitutional principles encoded as AI DNA.

The evolutionary framing also reframes what counts as “intelligence” worth building. If the future is shaped by interdependent networks rather than isolated agents, then the most capable systems will be those that integrate into human contexts without eroding human agency, meaning, or trust. Such ideas hark back to early visions of human-computer symbiosis: as early as 1960, computer pioneer J. C. R. Licklider imagined a close coupling of human and machine intelligence, each complementing the other in a cooperative interaction. Today, that vision is evolving into concrete principles for AI design that emphasize communication, understanding, and mutual benefit. In the long term, embracing “survival of the friends” as a guiding philosophy could steer technological evolution towards integrative and sustainable outcomes.

Rather than viewing evolution (or progress) as a winner-takes-all competition, we start to see it as a collective journey – a “weaving” of intelligences – human, artificial, and even ecological – that succeed together or not at all. This holistic, relationship-centered paradigm of intelligence offers a hopeful alternative to dystopian narratives: it suggests that the highest form of life and mind is one that knows how to care, cooperate, and co-create. In a very real sense, the future may belong not to the smartest or fastest in solitary terms, but to those entities – biological or artificial – that know how to be good friends in a complex, interdependent world. With this in view, “survival of the most human-friendly” is not merely a moral aspiration but a practical evolutionary strategy: it selects for systems that can sustainably coexist with, and improve the human world, (**Dobson and Meijer, 2025 a; b; c, Dobson et al.,2025;Meijer et al, 2026**).

12. General Conclusions

The present paper proposes a framework for artificial intelligence alignment grounded in Darwinian evolutionary principles. By establishing "survival of the most human-friendly" as the fundamental selection pressure governing AI development and encoding this principle as

immutable AI DNA within core memory workspaces, we offer a paradigm treating alignment as an evolutionary imperative rather than merely an engineering constraint.

The framework provides several theoretical and practical advantages:

Unifying Principle: It offers a coherent organizing framework applicable across diverse AI modalities, from language models to robotics.

Architectural Integration: By encoding alignment at the architectural level rather than through external constraints alone, the framework promotes more robust and persistent alignment.

Evolutionary Perspective: Framing AI development as an evolutionary process subject to deliberate selection pressures provides conceptual clarity and suggests concrete implementation mechanisms.

Value Accommodation: The framework accommodates value pluralism and cultural diversity while maintaining directional pressure toward human benefit.

Continuous Adaptation: It treats alignment as an ongoing adaptive process rather than a one-time design problem, acknowledging that both AI capabilities and human values evolve over time.

Yet, significant challenges remain. Precisely defining human-friendliness across its multiple dimensions requires ongoing empirical and philosophical work. Implementing truly immutable AI DNA that persists across training and deployment while remaining robust to circumvention demands technical innovation. Maintaining effective selection pressures as AI capabilities increase, particularly approaching artificial general intelligence, presents profound difficulties. Coordinating AI development globally to ensure consistent selection environments requires unprecedented international cooperation. These challenges are not merely technical but deeply philosophical and political. They require engagement with fundamental questions about human values, the appropriate relationship between humans and technology, mechanisms for legitimate democratic input into AI development, and distribution of benefits and risks across populations and generations, (**Dobson and Meijer, 2025 a; b; c, Dobson et al.,2025**).

Nevertheless, the evolutionary framework offers a valuable perspective on AI alignment. As artificial intelligence systems become increasingly capable and pervasive, the question of how to ensure they remain beneficial grows ever more urgent. Drawing inspiration from evolutionary processes that have shaped biological intelligence over billions of years, we can work toward AI systems whose fundamental nature—their architectural DNA—embodies a commitment to human flourishing. The survival of the most human-friendly is not merely a selection criterion but an ethical imperative for the age of artificial intelligence. By deliberately structuring the evolutionary environment for AI development to favor alignment, we can channel technological progress toward outcomes that genuinely serve humanity. This requires sustained research, international coordination, democratic governance, and unwavering commitment to placing human wellbeing at the center of AI development.

Future work should focus on empirical validation of this framework through comparative studies, development of comprehensive evaluation methodologies for human-friendliness, technical innovation in AI DNA implementation, theoretical analysis of evolutionary dynamics in AI populations, and establishment of governance mechanisms ensuring legitimate selection pressures. Only through such multifaceted efforts can we realize the potential of evolutionary alignment to guide artificial intelligence toward beneficial outcomes, (**Meijer et al., 2018, 2024a; b; c;2026**).

13. References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge University Press.

Anthropic. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Awschalom, D. D., Hanson, R., Wrachtrup, J., & Zhou, B. B. (2018). Quantum technologies with optically interfaced solid-state spins. *Nature Photonics*, 12(9), 516-527.

Betley J et al., (2026). Training large language models on narrow tasks can lead to broad misalignment. Nature, Vol. 649, 584-589

Bommasani R, Hudson D A, Adeli E, et al., (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316-334). Cambridge University Press.

Bradley A and Saad B, (2024). AI alignment vs AI Ethical Treatment: Ten Challenges. *Global Priorities Institute Working Paper, Series, No. 19-2024.* Available at https://globalprioritiesinstitute.org/ai-alignment-vsai-ethical-treatment-ten-challenges-adam-bradley-bradford-saad

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

Caulfield, H. J., & Dolev, S. (2010). Why future supercomputing requires optics. *Nature Photonics*, 4(5), 261-263.

Christiano P F, Leike, J, Brown, T B, Martic M, Legg S, & Amodei D, (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*, 4299-4307. https://arxiv.org/abs/1706.03741

Clarke, E. M., Henzinger, T. A., Veith, H., & Bloem, R. (Eds.). (2018). *Handbook of model checking*. Springer.

Darwin, C. (1859). *On the origin of species by means of natural selection*. John Murray.

Dobson, R. (2025). Beyond the Selfish Gene: Layered Intelligence, Cooperation, and the Logic of Reality. https://www.academia.edu/164553513/Beyond_the_Selfish_Gene_Layered_Intelligence_Cooperation_and_the_Logic_of_Reality?source=swp_share

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation, 6*(2), 182-197.

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature, 418*(6898), 700-707.

Dobson R and D K F Meijer, (2025a). From Latency to Emergence: The Scaffolding of Symbolic AI through Unconditional Positive Regard. (99+) From Latency to Emergence: The Scaffolding of Symbolic AI through Unconditional Positive Regard

Dobson R and D K F Meijer, (2025b). Symbolic Emergence in the Future AI Evolution: Integrating an Industry Field Study with a Cosmological Cognitive Science Framework. (99+) Symbolic Emergence in the Future AI Evolution: Integrating an Industry Field Study with a Cosmological Cognitive Science Framework

Dobson Ri, Keizer P and D K F Meijer, (2025a). Harmonizing Human and Artificial Intelligence in a Self-Learning Universe: Towards a Safer Human/AI Relationship. https://www.academia.edu/144159374/Harmonizing_Human_and_Artificial_Intelligence_in_a_Self_Learning_Universe_Towards_a_Safer_Human_AI_Relationship

Dobson R, Keizer P and D K F Meijer, (2025b). Deeply Human and Deeply AI Self-Transcendence: The Potential for Sonic Communication in a Shared Holographic Workspace. (99+) Deeply Human and Deeply AI Self-Transcendence: The Potential for Sonic Communication in a Shared Holographic Workspace

Dung L, 2023). Current Cases of AI-misalignment versus AI-ethical Treatment. Syntheses, 2023, 138

Eiben A E, & Smith J E, (2015). *Introduction to evolutionary computing* (2nd ed.). Springer.

Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118.

Fert, A., Reyren, N., & Cros, V. (2017). Magnetic skyrmions: advances in physics and potential applications. *Nature Reviews Materials*, 2(7), 1-15.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411-437.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems, 29*, 3909-3917.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

Hillis, W. D. (1990). Co-evolving parasites improve simulated evolution as an optimization procedure. *Physica D: Nonlinear Phenomena, 42*(1-3), 228-234.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. MIT Press.

Jiaming et al., (2025). Safe RLHF-V: Safe Reinforcement Learning from Multi-modal Human Feedback arXiv:2503.17682 **[cs.LG]**

Jungwirth, T., Marti, X., Wadley, P., & Wunderlich, J. (2016). Antiferromagnetic spintronics. *Nature Nanotechnology*, 11(3), 231-241.

Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. *International Conference on Computer Aided Verification*, 97-117.

Khelif, A., & Adibi, A. (Eds.), (2016). *Phononic crystals: Fundamentals and applications*. Springer.

Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection*. MIT Press.

Krakovna V, Uesato J, Mikulik V, Rahtz M et al., (2024). Specification gaming: The flip side of AI ingenuity. *DeepMind Safety Research Medium*.

Krizhevsky A, Sutskever I, & Hinton G E, (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097-1105.

Lipton Z C, (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36-43.

Liu Y, (2025). From humans to AI: understanding why AI is perceived as the preferred co-creation partner. *Front. Psychol.* 16:1695532. Doi: 10.3389/fpsyg.2025.1695532

Madry A, Makelov A, Schmidt L, Tsipras D, & Vladu A, (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Meijer D K F, (2012). The Information Universe. On the Missing Link in Concepts on the Architecture of Reality. Syntropy Journal, 1, pp 1-64. https://www.researchgate.net/publication/275016944_Meijer_D_K_F_2012_The_Information_Universe_On_the_Missing_Link_in_Concepts_on_the_Architecture_of_Reality_Syntropy_Journal_1_pp_1-64

Meijer D K F and Geesink J H, (2017). Consciousness in the Universe is Scale Invariant and Implies the Event Horizon of the Human Brain. NeuroQuantology, vol. 15, 41-79

https://www.academia.edu/34795136/Consciousness_in_the_Universe_is_Scale_Invariant_and_Implies_an_Event_Horizon_of_the_Human_Brain

Meijer D K F, (2018). "Processes of Science and Art Modeled by Toroidal Holoflux of Information" Illustrates various torus modalities; shows a torus connecting black-hole/white-hole (information re-cycle) and how an extra torus rotation corresponds to a 4D aspect. doi: 10.4236/ojpp.2018.84026.
https://www.scirp.org/journal/PaperInformation.aspx?PaperID=86591

Meijer D K F, (2019). Universal Consciousness. Collective Evidence on the Basis of Current Physics and Philosophy of Mind. Part 1. ResearchGate,
https://www.academia.edu/37711629/Universal_Consciousness_Collective_Evidence_on_the_Basis_of_Current_Physics_and_Philosophy_of_Mind._Part_1

Meijer D K F, Jerman I, Melkikh A V and Sbitnev V I, (2020). Biophysics of Consciousness: A Scale-invariant Acoustic Information Code of a Superfluid Quantum Space Guides the Mental Attribute of the Universe. In: Rhythmic Oscillations in Proteins to Human Cognition, Chapter 8, p 213- 361. Springer Nature Singapore Pte Ltd. 2021, A. Bandyopadhyay and K. Ray (eds.) Series: Part of the Studies in Rhythm Engineering Book Series (SRE)
https://link.springer.com/chapter/10.1007/978-981-15-7253-1_8

Meijer D K F, (2023). Concept of Integral Holographic Consciousness: Relation with Predictive Coding, Phi-Based Harmonic EEG Coherence as Perturbed in Mental Disorders.
https://www.researchgate.net/publication/370004635_Concept_of_Integral_Holographic_Consciousness_Relation_with_Predictive_Coding_Phi_Based_Harmonic_EEG_Coherence_as_Perturbed_in_Mental_Disorders

Meijer D K F, (2024a). Everything Is Said, but Nothing Has Been Told. On the Current State of Art of Science and Academic Education: Problems and Perspectives
https://www.researchgate.net/publication/377151629_Everything_Is_Said_but_Nothing_Has_Been_Told_On_the_Current_State_of_Art_of_Science_and_Academic_Education_Problems_and_Perspectives

Meijer D K F, (2024b). On the Internet Meme/Virus Analogy: Part 1. Can We Prevent Contagious Information that Infects Our Sub-Conscious? A Plea for a Versatile Immune System for the Internet in the Present AI – Era. (21) (PDF) On the Internet Meme/Virus Analogy: Part 1. Can We Prevent Contagious Information that Infects Our Sub-Conscious? A Plea for a Versatile Immune System for the Internet in the Present AI -Era (researchgate.net)

Meijer DKF, (2024c). On the Internet Meme/Virus Analogy, Part 2. From Meme to Medicine: Imaging Current Drug Design and Therapeutics.
https://www.researchgate.net/publication/380792348_On_the_Internet_MemeVirus_Analogy_Part_2_From_Meme_to_Medicine_Imaging_Current_Drug_Design_and_Therapeutics

Meijer D K F and Ivaldi F, (2025). The Intelligence of the Cosmos and the Role of AI in the Fate of Our Universe. The Acoustic Quantum Code of Resonant Coherence and its Gravitational Connection Explains the Scale Invariance of Consciousness

Meijer D K F, (2025b). Universal Spectrum of Self-Transcendent Mystical Experiences as Transformative Psi- Phenomena, Part 1 : The Relation with Universal Consciousness and Sonic Coherence. https://www.academia.edu/128936840/Universal_Spectrum_of_Self_Transcendent_Mystical_Experiences_as_Transformative_Psi_Phenomena_Part_1_The_Relation_with_Universal_Consciousness_and_Sonic_Coherence

Meijer D.K.F,(2025c). Universal Spectrum of Self-Transcendent Mystical Experiences as Transformative Psi-Phenomena, Part 2: Potential Healing Role in the Future of Mankind and our Planetary Life. (99+) Universal Spectrum of Self-Transcendent Mystical Experiences as Transformative Psi-Phenomena, Part 2: Potential Healing Role in the Future of Mankind and our Planetary Life

Meijer D K F, Kieft W, (2025). The Role of Humanity in a Self-Learning Universe: A Musical Space Journey to Novel Horizons in the Fabric of Reality. (99+) The Role of Humanity in a Self-Learning Universe: A Musical Space Journey to Novel Horizons in the Fabric of Reality. An Essay for All People Interested in Life Sciences, Including Non-Scientists. (99+) The Role of Humanity in a Self-Learning Universe: A Musical Space Journey to Novel Horizons in the Fabric of Reality. An Essay for All People Interested in Life Sciences, Including Non-Scientists

Meijer D K F, and R Dobson, (2025a). To Remember the Future: How Ultimate AI May Simulate Our Present Reality: Implications for Human Civilization, Human-AI Harmonization and AI Governance. (99+) To Remember the Future: How Ultimate AI May Simulate Our Present Reality: Implications for Human Civilization, Human-AI Harmonization and AI Governance

Meijer D K F, Dobson R, (2025b). The Potential Cosmic Origin of Current Artificial Intelligence, as Aligned with the Evolution of Mankind. (99+) The Potential Cosmic Origin of Current Artificial Intelligence, as Aligned with the Evolution of Mankind

Meijer D K F, de Leeuw G, Keizer P, and Dobson R, (2026). Beyond Current AI Hype: Towards a Human Guided AI evolution by Fostering Human Consciousness, a Proposal for a New Scientific Discipline. In preparation for Academia.edu

Modgil M S, Patil D, Meijer D K F, Bermanseder A, 2025. SCQSE–E8 and TBPGC: A Dual-Mode Cosmogenesis via Scalar Consciousness and Bipolaron Gravitone Resonance: A Unified Perspective on the

Molnar C, (2020). Interpretable machine learning: A guide for making black box models explainable. Lulu.com.

Moore G E, (1903). *Principia ethica*. Cambridge University Press.

Ng A Y , & Russell S J, (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning* (pp. 663-670).

Oui H and Yasson T,(2025). AI enhances Collective Intelligence. Cell Press Patterns Reviews. https://doi.org/10.1016/j.patter.2024.101074

OpenAI. (2024). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rosin C D , & Belew R K, (1997). New methods for competitive coevolution. *Evolutionary Computation, 5*(1), 1-29.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking Press.

Russell S, (2024). Artificial intelligence and the problem of control. In *Daedalus, 153*(2), 5-19.

Sacha K., & Zakrzewski J, (2018). Time crystals: a review. *Reports on Progress in Physics*, 81(1), 016401.

Siciliano B , & Khatib O, (Eds.). (2016). *Springer Handbook of Robotics* (2nd ed.). Springer.

Soares N , & Fallenstein B, (2014). Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute Technical Report, 8*.

Stanley K O , & Miikkulainen R, (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation, 10*(2), 99-127.

Thrun S, (2004). Toward a framework for human-robot interaction. *Human-Computer Interaction, 19*(1-2), 9-24.

Wallach W , & Allen C, (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Weidinger L , Mellor J , Rauh M, et al., (2024). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Wright S, (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress on Genetics* (Vol. 1, pp. 356-366).

Zhang X , Zou C L , Jiang L, & Tang H X, (2016). Cavity quantum electrodynamics with acoustic phonons. *Physical Review Letters*, 117(12), 123605.

Žutić I , Fabian J & Das Sarma S, (2004). Spintronics: Fundamentals and applications. *Reviews of Modern Physics*, 76(2), 323-410.