Will AI Light Up Human Creativity or Replace It? Toward Well-Being AI for Co-evolving Human and Machine Intelligence

Takashi Kido[1], Keiki Takadama[2]

[1] Teikyo University

[2] The University of Tokyo

kido.takashi@gmail.com, takadama@g.ecc.u-tokyo.ac.jp

Abstract

The rapid evolution of generative artificial intelligence raises a fundamental question: will AI illuminate human creativity or gradually replace it? Building upon prior discussions of Human-Compatible AI and AI-Powered Science, this study revisits the concept of Well-Being AI in the era of generative and multi-agent systems. We argue that AI presents a dual potential: it can amplify human creativity and accelerate scientific discovery, but it may also diminish human autonomy through overreliance and automation. To address this tension, we extend the dual framework by integrating creativity as a central dimension of well-being within human–AI collaborative ecosystems. This paper outlines the conceptual foundations, identifies key challenges, and proposes research directions toward a co-evolutionary human–AI ecosystem that maximizes human potential and enhances physical, mental, and social flourishing rather than replacing it.

Introduction

Artificial intelligence has entered a new phase characterized by generative models, large language systems, and multi-agent collaboration. These systems are increasingly involved in writing, design, research, and decision-making. This transformation raises a fundamental question:
Will AI light up human creativity or replace it?

In previous AAAI Spring Symposia, we examined Human-Compatible AI and AI-Powered Science as dual perspectives for promoting well-being and reflected on Socially Responsible AI for Well-Being (Kido and Takadama 2019; 2022; 2023; 2024). Human-compatible AI focuses on alignment, fairness, and the preservation of human autonomy, whereas AI-Powered Science explores how computational systems can expand scientific discovery and collective intelligence.

However, in the generative era, alignment and acceleration alone are insufficient. As AI systems begin to participate directly in creative and epistemic processes, the central issue shifts from control to co-evolution. Rather than viewing AI as an isolated technical system, we approach it as part of an emerging intelligent network of human–AI collaborative ecosystems in which creativity and autonomy evolve together. From this perspective, Well-Being AI aims not merely to reduce harm or increase efficiency but to maximize human potential by enhancing the physical, mental, and social flourishing of individuals and society (Kido 2024).

In this study, we extend the dual framework of Human-Compatible AI and AI-Powered Science by introducing creativity as a core dimension that connects alignment, scientific discovery, and human flourishing within co-evolving human–machine intelligence.

In this context, the AAAI Spring Symposium aims to provide a shared conceptual foundation for examining how generative and multi-agent AI systems can illuminate, rather than replace, human creativity. As the symposium organizers, we position this symposium as a continuation of prior discussions on Human-Compatible AI and AI-Powered Science, while advancing Well-Being AI as an integrative framework for guiding human–AI co-evolution.

This study makes three contributions.

1. It reframes the creativity–replacement debate from a well-being perspective.
2. It extends the dual framework of Human-Compatible AI and AI-Powered Science by integrating creativity as a central dimension.
3. It proposes research directions for the co-evolution of human and machine intelligence within human–AI collaborative ecosystems.

AI for Expanding Human Creativity

On the positive side, AI technologies are opening new pathways for progress by expanding human creativity and understanding. In the generative era, artificial intelligence no longer functions merely as an analytical tool; it increasingly participates in ideation, hypothesis generation, design exploration, and interdisciplinary synthesis. These developments invite us to reconsider creativity not as an exclusively human attribute but as a relational phenomenon emerging within human–AI collaborative ecosystems.

Generative and reasoning systems accelerate the pace of scientific discovery by revealing the latent structures within complex data. Deep representation learning has demonstrated how neural networks can compress and reorganize high-dimensional information into meaningful internal representations (Hinton and Salakhutdinov, 2006). Such computational abstraction extends the reach of human cognition, allowing patterns to emerge that would otherwise remain inaccessible. In this sense, AI augments human insight by reshaping the epistemic landscape in which discoveries occur.

A striking example of machine-augmented scientific creativity is AlphaFold (Jumper et al. 2021), whose predictive capacity has illuminated protein structures that have resisted traditional experimental methods for decades. Importantly, AlphaFold does not eliminate human scientific reasoning; rather, it reconfigures the collaborative process between computational inference and biological interpretation. The result was not the replacement of scientific creativity but its redistribution across a network of human and machine agents.

This networked dimension is even more pronounced in multi-agent infrastructures. Systems designed to coordinate specialized AI agents for reasoning, modeling, and cross-domain synthesis demonstrate how creativity can emerge from structured collaboration rather than isolated intelligence (Swan et al., 2023). Such infrastructures point toward an intelligent network in which human expertise and machine inference are gathered rather than dispersed, thereby amplifying collective intelligence.

In applied domains, the integration of language models with retrieval mechanisms and embodied execution systems further illustrates this shift. For example, research on LLMs with retrieval-augmented generation in robotic planning contexts has shown how generative models can support structured reasoning, task decomposition, and adaptive execution (Yamanaka and Kido 2024). Here, AI does not function as an autonomous decision-maker detached from human oversight; instead, it operates as a collaborative reasoning partner embedded within broader cognitive workflows.

Beyond scientific and technical applications, generative AI has increasingly contributed to design, education, and healthcare. In these domains, AI systems can facilitate reflection, broaden exposure to alternative perspectives, and support inclusive participation in the learning process. When appropriately guided, such systems act as catalysts for imagination and empathy, fostering personal growth and interdisciplinary creativity. Within the Well-Being AI framework, this expansion of creative capacity aligns with the goal of maximizing human potential by enhancing physical, mental, and social flourishing (Kido 2024).

From the perspective of AI-Powered Science, creativity is not confined to artistic production or novel expression; it includes the generation of new hypotheses, the integration of heterogeneous knowledge, and the acceleration of innovation across scales. When embedded within human–AI collaborative ecosystems, generative systems can illuminate new conceptual connections and enable forms of co-creation that were previously impractical to achieve. Thus, creativity is distributed, iterative, and dialogical.

However, this expansion is not automatic. The amplification of creative capacity depends on relational configurations that preserve human autonomy and critical engagement. The same generative model that supports exploratory thinking under human guidance may contribute to cognitive passivity under conditions of uncritical delegation. Therefore, the expansion of creativity must not be understood as an intrinsic property of AI systems but as a dynamic outcome of how human–AI ecosystems are designed and governed.

In summary, AI for Expanding Human Creativity represents one trajectory of co-evolution, in which generative and multi-agent systems illuminate latent possibilities, accelerate discovery, and redistribute creative labor across human–machine networks. When aligned with the principles of Well-Being AI, such expansion contributes not only to efficiency or productivity but also to the enrichment of human imagination and collective flourishing.

AI and the Risks of Replacement

While generative and multi-agent systems expand creative possibilities, the same technological advances introduce profound risks. The expansion of machine capabilities does not automatically guarantee the expansion of human autonomy. Conversely, without careful design and governance, AI systems may gradually displace the cognitive and moral capacities they were intended to support.

A central concern is the problem of control and alignment. As Russell (2019) argues, the challenge of artificial intelligence is not merely technical performance but ensuring that systems remain compatible with human values and preferences. In the context of generative AI, this problem is more subtle. The risk is not only catastrophic misalignment but also gradual cognitive outsourcing, where humans defer judgment, reflection, and responsibility to algorithmic outputs.

This outsourcing can reshape epistemic behavior. When predictive systems provide fluent and authoritative responses, users may reduce their efforts to verify, interpret, or question the results. Over time, this may narrow the horizon of creativity, encouraging conformity to probabilistic averages rather than exploration. In this sense, replacement does not occur through force but through convenience and necessity.

Earlier discussions of interpretability and cognitive bias have highlighted how opaque systems may reinforce hidden assumptions and social embeddedness (Kido and Takadama 2019). Without transparency and explainability, AI systems risk amplifying biases while obscuring their origin. In generative systems trained on large-scale data, such opacity becomes more consequential as outputs increasingly shape discourse, design, and decision-making.

Fairness and well-being pose additional challenges. Achieving fairness in AI systems requires not only technical debiasing but also a deeper reflection on what constitutes fairness within diverse social and cultural contexts (Kido and Takadama 2022). When generative models are widely deployed across domains such as education, hiring, and governance, disparities may be scaled rather than reduced. The appearance of neutrality may conceal the structural inequalities embedded in the training data and optimization objectives.

Recent analyses of generative AI in social and individual well-being further emphasize the need to examine how AI systems influence personal development, social cohesion and moral agency (Kido and Takadama 2024). While generative systems can assist in reflection and creativity, they may also foster dependency, homogenization of expression, and diminished self-efficacy if used uncritically.

Therefore, the problem of replacement is not limited to economic automation or job displacement. It extends to cognitive, cultural, and ethical dimensions. As AI systems increasingly participate in writing, planning, evaluation, and even emotional interactions, humans may gradually relinquish their interpretive authority. The loss is subtle: it is not the disappearance of creativity but its erosion through delegation.

Importantly, even highly accurate systems do not eliminate these concerns. An imperfect AI is risky because errors are inevitable and may go undetected. However, a technically perfect AI can also be dangerous if it discourages questioning. When individuals cease to interrogate outputs, they weaken their capacity to judge and choose. Human autonomy is compromised not only by machine superiority but also by reduced human engagement.

These risks underscore the importance of safeguarding human autonomy in human–AI collaborative ecosystems. Alignment, interpretability, fairness, and human-centered governance are not peripheral technical details; they are structural conditions that determine whether AI contributes to expansion or replacement of human labor. The trajectory of co-evolution depends on relational configurations, such as how authority, agency, and responsibility are distributed between humans and machines.

Thus, AI and the Risks of Replacement represent a second possible trajectory of co-evolution: one in which increasing machine autonomy coincides with diminishing human agency. The challenge is not to halt technological development but to design systems that prevent cognitive erosion and preserve the freedom to question, interpret, and choose.

Integration: Toward Well-Being AI for Co-evolving Human and Machine Intelligence

The preceding sections outline two contrasting trajectories of human–AI development: expansion and replacement. However, these are not mutually exclusive technological destinies. Rather, they represent relational configurations within the evolving human–AI ecosystems. Therefore, the central challenge is not to choose between optimism and pessimism but to design the conditions under which creative expansion can occur without eroding human autonomy.

Well-Being AI provides a conceptual foundation for this integration. Rather than treating alignment, scientific acceleration, and creativity as separate objectives, Well-Being AI frames them as interdependent dimensions of human flourishing. In this view, AI systems are embedded within intelligent networks of human–AI collaborative ecosystems, where creativity, autonomy, and collective well-being co-evolve.

To clarify the possible trajectories of co-evolution, we introduce a conceptual orientation structured along two dimensions:

- **Creativity**: Expansion vs. Replacement

- **Autonomy**: Human-Preserved vs. Machine-Dominant

These dimensions yield four relational configurations:

	Creativity Expansion	**Creativity Replacement**
Human Autonomy Preserved	Augmented Co-Creation (e.g., LLM-assisted writing with critical reflection)	Skill Erosion (e.g., habitual uncritical AI-generated output)
Machine Autonomy Dominant	Accelerated Discovery (e.g., large-scale autonomous scientific inference such as AlphaFold)	Over-Delegation (e.g., fully automated decision systems without human oversight)

Table 1. Conceptual Orientation of Human–AI Co-evolution

This four-quadrant orientation is proposed as a shared analytical lens for symposium discussions. It offers participants a common conceptual vocabulary for examining how specific technologies, governance models, and application domains position themselves in the landscape of creativity and autonomy. Rather than prescribing outcomes, it supports a structured dialogue across disciplinary perspectives.

This framework does not classify the technologies themselves. The same generative model may contribute to augmented co-creation when embedded within reflective human workflows, but may contribute to skill erosion when used passively. The determining factor is not the algorithm alone but the distribution of agency and responsibility within the ecosystem.

The upper-left quadrant, Augmented Co-Creation, represents the aspirational trajectory of Well-Being AI. Generative systems illuminate new possibilities while preserving human judgment. Creativity is expanded through dialogue rather than being substituted by automation. Human agency remains central, even as machine capabilities increase.

The lower-left quadrant, Accelerated Discovery, captures large-scale computational inference, where machine autonomy plays a dominant role. Such configurations can dramatically extend scientific capacity, as seen in AI-driven structural predictions (Jumper et al. 2021). However, sustained human interpretive authority is essential to prevent epistemic dependency.

The upper-right quadrant, Skill Erosion, illustrates a subtle form of replacement. When users habitually rely on AI-generated outputs without critical engagement, their creative capacity may gradually narrow. The risk here is not immediate automation but rather incremental cognitive outsourcing.
The lower-right quadrant— Over-Delegation —represents the most concerning configuration. When machine autonomy dominates both creative production and decision-making authority, human judgment risks becoming peripheral. In such cases, governance, interpretability, and value alignment are indispensable safeguards.

This four-quadrant orientation functions as a conceptual blueprint, rather than a deterministic map. This enables a systematic examination of how creativity and autonomy interact across domains such as education, healthcare, scientific research, and governance. By identifying relational configurations, researchers can analyze not only technical performance but also the structural distribution of agency within human–AI ecosystems.

Importantly, the framework highlights that expansion and replacement are not binary outcomes but dynamic tendencies shaped by design choices, institutional norms and cultural practices. Therefore, co-evolution is neither guaranteed nor inherently beneficial; it must be intentionally cultivated. Within the Well-Being AI paradigm, integration requires continuous balancing.

- expanding creative capacity without diminishing reflective agency,
- leveraging machine autonomy without displacing moral responsibility,
- accelerating discoveries without undermining interpretive diversity.

In this sense, Well-Being AI shifts the focus from isolated system optimization to ecosystem-level design. The goal is not merely to align AI outputs with human values but to sustain environments in which humans remain capable of questioning, imagining, and choosing.

Thus, the four-quadrant framework serves as an architectural orientation for future inquiries. This invites interdisciplinary collaboration among technical researchers, philosophers, educators, and policymakers to examine how human–AI collaborative ecosystems can be structured to maximize human potential while safeguarding autonomy.

Toward a Co-evolutionary Human–AI Ecosystem

The four-quadrant orientation clarifies that the future of human–AI co-evolution is not predetermined technologically. Instead, it depends on how creative expansion and human autonomy are balanced in specific domains. Three interrelated research directions emerge from this framework.

Expanding Human Creativity

First, research must investigate how generative and multi-agent systems can augment reflection, imagination, and interdisciplinary synthesis rather than simply automate output. This involves developing architectures that support collaborative creativity, such as systems that scaffold ideation, encourage questioning, and facilitate iterative co-creation between humans and machines.

In AI-Powered Science, future studies should examine how distributed computational agents can integrate heterogeneous data, bridge disciplinary boundaries, and accelerate hypothesis generation while maintaining interpretive transparency. The challenge is not only scaling computational capability but also designing mechanisms through which human insight remains central within the discovery process.

Evaluation metrics must also be developed. Traditional performance metrics (accuracy, speed, and optimization scores) are insufficient for capturing creative augmentation. New methodologies are required to assess the breadth of imagination, reflective engagement, and interdisciplinary integration. Such metrics would align technical innovation with the broader goals of Well-Being AI.

The symposium invites contributions that explore both technical architectures and evaluative methodologies capable of sustaining such augmentations.

Safeguarding Human from Machine Autonomy

Second, safeguarding human autonomy is essential. Interpretability, fairness, and alignment must be treated as structural design principles, rather than post hoc corrections.

Future research should explore how explainable interfaces, participatory governance models, and human-in-the-loop architectures can prevent cognitive outsourcing and skill erosion. The goal is not to constrain AI capabilities but to preserve the human capacity to question, deliberate, and choose.

Cross-cultural and interdisciplinary studies are particularly important in this context. Concepts such as fairness, responsibility, and well-being are socially embedded and context dependent. Integrating philosophical, legal, and sociological inquiries with technical research can help ensure that alignment mechanisms reflect pluralistic human values rather than narrow optimization criteria.

Discussions at the symposium will examine how alignment and governance mechanisms can be designed to prevent cognitive erosion while preserving innovation.

Integrating Human and Machine Intelligence toward Well-Being AI

Third, integration requires a shift from system-level evaluation to ecosystem-level designs. Human–AI collaborative ecosystems must be studied as socio-technical networks in which agency, authority, and creativity are distributed dynamically.

Therefore, future research should address the following issues:

how educational systems can cultivate AI-augmented creativity without diminishing foundational skills,

how governance frameworks can balance innovation with accountability,

how embodied and digital systems can support sustainable co-evolutionary forms.

Interdisciplinary methodologies are essential for modeling these ecosystems. Combining technical experimentation with philosophical reflection and policy analysis can help articulate design principles for maximizing human potential while safeguarding the autonomy of users.

Ultimately, the research agenda of Well-Being AI calls for a redefinition of progress itself. Rather than measuring advancement solely in terms of computational efficiency or economic productivity, progress should be evaluated in terms of expanded human capability, preserved agency, and enriched collective growth.

By bringing together diverse disciplinary perspectives, the symposium seeks to articulate ecosystem-level principles for responsible and creative co-evolution.

Conclusion

The rapid evolution of generative and multi-agent artificial intelligence confronts us with a foundational question: Will AI illuminate human creativity or gradually replace it?

This study argues that the answer depends not on technological inevitability but on relational design. By extending the dual framework of Human-Compatible AI and AI-Powered Science, we introduce creativity as a central dimension within Well-Being AI. Through this lens, expansion and replacement emerge as alternative trajectories shaped by the interaction between human autonomy and machine capability within collaborative ecosystems.

The four-quadrant framework presented in this study offers a conceptual blueprint for examining these trajectories. It emphasizes that AI systems do not inherently determine outcomes; rather, outcomes arise from the configurations of agency, governance, and engagement.

Thus, Well-Being AI shifts the focus from mere alignment and optimization to the maximization of human potential. The objective is not simply to prevent harm but to cultivate environments in which creativity is illuminated, autonomy is preserved, and human and machine intelligence co-evolve in ways that deepen physical, mental, and social flourishing.

The future of AI will not be defined solely by algorithmic sophistication but by how researchers, designers, and policymakers collectively shape the ecosystems within which human and machine intelligence interact.

This symposium invites participants to engage in this shared endeavor: to design AI systems that illuminate creativity, preserve autonomy, and maximize human potential within co-evolving human–machine societies.

References

Hinton, G. E., & Salakhutdinov, R. R. 2006. Reducing the dimensionality of data using neural networks. Science 313(5786): 504–507.

Jumper, J.; Evans, R.; Pritzel, A.; et al. 2021. Highly accurate protein structure prediction using AlphaFold. Nature 596: 583–589.

Kido, T. 2024. AI and Well-Being: Enhancing Health, Happiness, and Cultural Understanding. In Proceedings of the International Conference on Human-Computer Interaction (HCI 2024), 93–102. Springer.

Kido, T., and Takadama, K. 2019. Challenges for Interpretable AI for Well-Being: Understanding Cognitive Bias and Social Embeddedness. Proceedings of the AAAI 2019 Spring Symposium. Palo Alto, CA: AAAI Press.

Kido, T., and Takadama, K. 2022. The Challenges for Fairness and Well-Being: How Fair Is Fair? Achieving Well-being AI. Proceedings of the AAAI 2022 Spring Symposium. Palo Alto, CA: AAAI Press.

Kido, T., and Takadama, K. 2023. AAAI 23 Spring Symposium Report on Socially Responsible AI for Well-Being. AI Magazine 44(2): 211–212.

Kido, T., and Takadama, K. 2024. The Challenges for GenAI in Social and Individual Well-Being. Proceedings of the AAAI Spring Symposia, 365–367. Palo Alto, CA: AAAI Press.

Russell, S. 2019. Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking.

Swan, M.; Kido, T.; Roland, E.; and dos Santos, R. P. 2023. Math Agents: Computational Infrastructure, Mathematical Embedding and Genomics. arXiv preprint. arXiv:2305.09123 [cs.AI].

Yamanaka, J., and Kido, T. 2024. Evaluating Large Language Models with RAG Capability: A Perspective from Robot Behavior Planning and Execution. In Proceedings of the AAAI Spring Symposia; 452–456. Palo Alto, CA: AAAI Press.