Human-Compatible AI and AI-Powered Science: Insights from AAAI Spring Symposium and Beyond

Takashi Kido
Teikyo University, Advanced Comprehensive Research Organization
kido.takashi@gmail.com

Abstract

For over a decade, we have been organizing and leading AAAI Spring Symposium sessions at the intersection of Well-being and AI, fostering discussions on AI's ethical, societal, and scientific implications of Artificial Intelligence. This study explores the evolution of key themes, particularly **Human-Compatible AI** and **AI-Powered Science**, emphasizing their growing importance in shaping the future of AI.

Human-compatible AI ensures that it aligns with human values while enhancing both individual and societal well-being. It focuses on value alignment, explainability, and mitigation of cognitive biases. **Well-being AI**, an extension of this concept, focuses not only on fairness, but also active support for psychological, cognitive, and social well-being.

Meanwhile, **AI-Powered Science** is transforming research paradigms and raising critical questions about reliability, trust, and the role of human intuition in AI-generated knowledge. By integrating AI-Powered Science with Well-being AI, we envision a future in which AI actively contributes to human flourishing.

Drawing on insights from past symposia and personal research, this paper discusses the progression from early concerns about AI fairness to debates on AI's role in knowledge creation and human-AI collaboration. Finally, it argues for **Well-being AI** as the next stage in AI's evolution, emphasizing the necessity of AI systems designed not only for efficiency but also for human happiness and growth.

Introduction

AI is rapidly transforming health care, education, and creativity. The rise of **Generative AI (GenAI)** presents immense opportunities for enhancing individual and societal well-being, but also introduces ethical and technical challenges (Kido &Takadama, 2024). 2024 marks the beginning of the AI for Science era, with AI playing a pivotal role in accelerating discovery. Recent Nobel Prize-winning research has demonstrated AI's ability to drive breakthroughs across scientific disciplines, solidifying its role in modern research (Hinton et al., 2006) (Jumper et al. 2021).

However, as Stuart Russell cautions in *Human Compatible: Artificial Intelligence and the Problem of Control*
AI systems optimized purely for objective functions may misalign with human intentions unless explicitly designed for ethical trustworthiness (Russell, 2019). This issue is particularly pressing as AI becomes integral to scientific discovery, raising concerns about knowledge reliability, bias in AI-driven research, and the role of human intuition in validating AI findings (Kido and Takadama, 2022).

To address these challenges, we propose a dual framework:

1. **Human-compatible AI:** Ensuring that AI systems align with human values, integrating fairness, interpretability, and control (Kido & Takadama, 2024).
2. **AI-Powered Science:** Leveraging AI for research while mitigating risks of bias, misinformation, and security threats (Swan et al., 2023).

Additionally, we introduce **Well-being AI**, which extends Human-Compatible AI by actively promoting **physical, mental, and social well-being**. AI must move beyond ethical concerns to actively support human flourishing (Kido 2024).

The Evolution of AI: From Interpretable AI to Socially Responsible AI

The integration of AI into society has necessitated a shift in design principles. Early AI systems prioritized explainability, ensuring that users could understand AI-driven decisions. This stage of AI development is often referred to as **Interpretable AI**, in which the focus is on making AI decisions transparent and understandable (Kido & Takadama, 2019). However, transparency alone

is insufficient to address concerns about fairness and ethical accountability (Kido & Takadama, 2022).

To overcome these challenges, researchers introduced **Socially Responsible AI**, which aims to mitigate bias, promote fairness, and enhance trust in AI systems (Kido and Takadama 2023). Although **Socially Responsible AI** represents a significant advancement, it has limitations. Fairness is often defined at the societal level, making it difficult to account for individual differences in needs, values, and cultural contexts. An AI system can be fair, yet fails to provide personalized, human-centered outcomes.

This limitation has led to the emergence of **Human-Compatible AI**, which ensures that AI not only adheres to ethical standards but also aligns with individual human experiences and values (Kido and Takadama, 2024). Figure 1 summarizes this evolution, illustrating the transition from **Interpretable AI** to **Socially Responsible AI** and ultimately to **Human-Compatible AI**, which prioritizes both societal fairness and individual well-being.

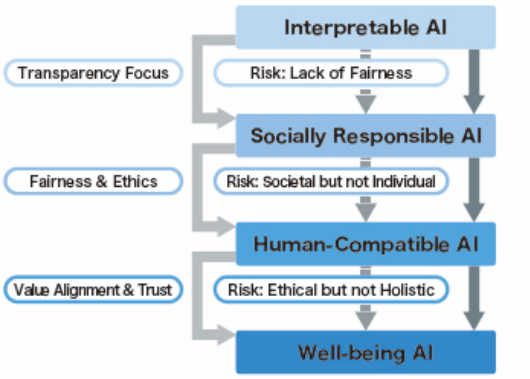

Figure 1: Evolution of AI – From Interpretable AI to Well-being AI

Human-Compatible AI: Towards AI that Aligns with Human Values

As AI becomes an integral part of human decision making, concerns about its alignment with human values and the risks of over-reliance are becoming increasingly pressing. The rapid adoption of **Generative AI (GenAI)** has transformed how people interact with AI, shifting it from a passive tool to an active creator and decision influencer. This shift raises two fundamental questions.

- Can AI be controlled to prevent unintended consequences or goal misalignments?
- Humans become overly dependent on AI, leading to cognitive decline and loss of autonomy.

These concerns highlight the need for a paradigm shift in AI design. **Human-compatible AI** must go beyond fairness and interpretability; it must be designed to align with human values, ensure trust, and support cognitive decision making without overriding human agency.

Figure 2: Discussion on Well-being and GenAI provides a structured visualization of these challenges and opportunities. This figure, based on discussions from the *2024 AAAI Spring Symposium, "The Impact of GenAI for Social and Individual Well-being"*, maps the intersection of AI, human well-being, and Generative AI technologies, highlighting both the potential and the risks associated with AI deployment.

- The left side (Well-being AI perspective) highlights AI's positive contributions, such as improving healthcare, mental well-being, and personalized education.
- The right side (GenAI perspective) identifies risks, such as misinformation, bias, ethical dilemmas, and excessive reliance on AI-generated content.
- The central intersection defines the core challenge of **Human-Compatible AI**, balancing AI's ability to enhance human capabilities while ensuring that people remain in control.

To achieve this balance, **Human-Compatible AI** must integrate the following three key elements:

- **Value Alignment:** AI must dynamically adapt to ethical perspectives and social norms, rather than rigidly optimizing fixed objectives.
- **Trust and Accountability:** AI decisions should be explainable and justifiable to ensure human oversight and prevent blind reliance.
- **Cognitive Support:** AI should enhance human intelligence rather than replace it, encouraging critical thinking instead of passive consumption.

AI should not be a detached, automated authority; it must be a collaborative system that enhances human agency and decision-making. Recent research has explored how Retrieval-Augmented Generation (RAG) models improve AI's role of AI in human decision-making by enhancing contextual understanding and reasoning (Yamanaka and Kido 2024).As illustrated in Figure 2, the key to responsible AI development lies in ensuring that AI empowers individuals and society while mitigating the risks of misalignment and overreliance (Kido, 2024).

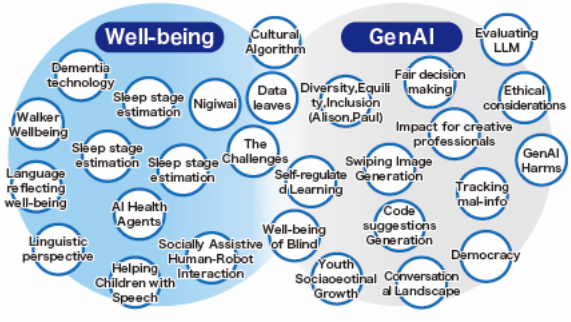

Figure 2: Discussion on Well-being and GenAI

AI-Powered Science: A New Paradigm for Discovery

AI has revolutionized scientific research and provided unprecedented capabilities for discovery and innovation. The 2024 Nobel Prize-winning research (Hinton et al., 2006) (Jumper et al., 2021) demonstrated AI's potential to advance physics, chemistry, and medicine, accelerating breakthroughs in genomics, materials science, and climate modeling (Swan et al., 2023).

However, although **AI-Powered Science** presents immense potential, it also raises challenges related to trustworthiness, interpretability, and fairness. *Figure 3* shows key discussions from the 2024 AAAI Spring Symposium, illustrating the balance between AI's risks and potential. The horizontal axis represents **risk versus potential**, while the vertical axis distinguishes **between social and individual impacts**. **AI-Powered Science** is positioned in the potential region, emphasizing its contributions to research acceleration; however, ethical concerns highlight the need for oversight.

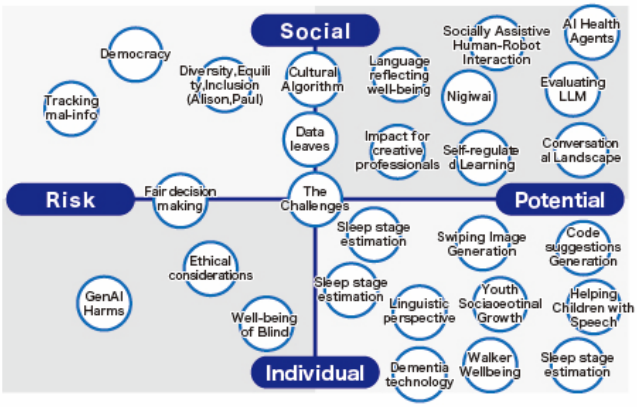

Figure 3: Discussion on Risk and Potential

Key Challenges to Address

1. **Reliability of AI-Generated Knowledge**
   - How can the AI-generated findings be validated and reproduced?
   - AI's black-box nature of AI raises concerns regarding **transparency**.
2. **The Role of AI in Scientific Discovery**
   - Can AI move beyond pattern recognition to contribute to the formation of theory?
   - Although AI can assist in hypothesis generation, human intuition and critical thinking remain essential.
3. **Bias and Fairness in AI-Driven Research**
   - AI models trained on biased datasets risk reinforcing the existing inequalities.
   - AI-driven research must ensure equity in scientific advancements.

The insights from *Figure 3* emphasize that, while **AI-Powered Science** drives groundbreaking discoveries, its development must adhere to trust, fairness, and ethical integrity. As science AI enters a pivotal era, it is crucial to establish clear ethical guidelines and regulatory frameworks to ensure that AI-driven discoveries remain credible, transparent, and aligned with human values. The future of **AI-Powered Science** depends on designing AI systems that not only accelerate research, but also uphold the integrity of scientific inquiry.

The Future of Well-being AI: Bridging Human-Compatible AI and AI-Powered Science

The evolution of AI does not stop with ensuring ethical compliance and alignment with human values. To truly support human flourishing, AI must actively contribute to **psychological, cognitive, and social well-being**. This vision is encapsulated in **Well-being AI**, a concept first introduced in our 2017 AAAI Spring Symposium on Well-being AI: From Machine Learning to Subjectivity-Oriented Computing. We defined Well-being AI as

"An Artificial Intelligence that aims to promote psychological well-being (that is, happiness) and maximize human potential. Our environment escalates stress, provides unlimited caffeine, distributes nutrition-free 'fast' food, and encourages unhealthy sleep behavior. For this issue, Well-being AI provides a way to understand how our digital experience affects our emotions and our quality of life and how to design a better well-being system that puts humans at the center."

Since its initial proposal, Well-being AI has evolved to address new challenges posed by AI's increasing role in human decision-making, creativity, and social interactions. The rise of Generative AI (GenAI) has brought unprecedented opportunities but also critical risks that must be carefully considered. Two key concerns have emerged:

1. **The risk of AI overpowering human autonomy** leads to overreliance and potential cognitive stagnation. As AI becomes more integrated into everyday life, there is growing concern that humans may become overly dependent on AI-driven recommendations, weakening their ability to make independent decisions. The challenge lies in designing AI systems that enhance human agency rather than replacing it.
2. **The risk of AI-driven manipulation**, in which opaque systems exploit human biases instead of supporting genuine well-being. AI models trained on vast datasets can learn and amplify existing biases, creating unethical reinforcement loops in decision making, media consumption, and even social interactions. Without proper safeguards, AI

can be used to influence human behavior in ways that are misaligned with well-being.

To mitigate these risks, Well-being AI must be designed not merely as a tool for optimization, but also as a framework for human empowerment. This should ensure that AI systems act as partners rather than replacements, enhancing human decision-making while preserving autonomy. This requires AI systems to integrate the cognitive, emotional, and social support mechanisms that align with human values.

In this light, we propose three primary dimensions of Well-being AI:

- **Psychological Well-being**: AI supports mental health, reduces stress, and fosters emotional resilience through applications such as AI-assisted mindfulness coaching and mental health interventions.
- **Cognitive Growth**: AI enhances learning, creativity, and critical thinking without replacing human intuition. Examples include AI-powered educational tools and knowledge discovery systems that augment human intelligence.
- **Social Engagement**: AI fosters meaningful human connections and mitigates social isolation. Conversational AI, digital companionship, and AI-powered community building tools can contribute to this goal.

These dimensions align closely with the discussions from the 2024 AAAI Spring Symposium on the Impact of GenAI on Social and Individual Well-being, where the balance between AI's potential and its ethical risks was critically examined. By integrating **Human-Compatible AI and AI-Powered Science**, Well-being AI offers a roadmap for AI systems that are not only ethical and fair but also actively beneficial to human life.

Conclusion

This study explored the intersection of Human-Compatible AI and AI-Powered Science in promoting well-being. By balancing these two perspectives, we can mitigate the risks of AI while maximizing its benefits, ensuring that AI is not only ethical and efficient, but also human-centered.

As AI evolves, the transition from Explainability to Justifiability has become a key challenge. AI must not only be transparent, but also capable of providing reasoning that aligns with human intuition, values, and well-being.

Well-being AI represents the next step in AI evolution, an AI that goes beyond responsibility and fairness to actively enhance human happiness, cognitive growth, and social engagement. Through interdisciplinary research, ethical frameworks, and innovative applications, we strive to develop AI that is not only powerful but also aligned with human values and flourishing.

As planners of the AAAI SSS25 symposium, we aim to foster discussions that shape the future of AI in a way that empowers individuals and society. By integrating Human-Compatible AI, AI-Powered Science, and Well-being AI, we envision a future in which AI is not just a tool but a true partner in human progress.

Acknowledgments

We would like to thank the program committees of this symposium for their assistance.

References

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data using neural networks. *Science* 313(5786): 504–507. https://doi.org/10.1126/science.1127647

Jumper, J.; Evans, R.; Pritzel, A.; et al. 2021. Highly accurate protein structure prediction using AlphaFold. *Nature* 596: 583–589. https://doi.org/10.1038/s41586-021-03819-2

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)*, Lecture Notes in Computer Science, 93–102. Springer.

Kido, T., and Takadama, K. 2019. Challenges for Interpretable AI for Well-Being: Understanding Cognitive Bias and Social Embeddedness. In *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. In *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. https://doi.org/10.1609/aimag.v44i2.23015

Kido, T., and Takadama, K. 2024. The Challenges for GenAI in Social and Individual Well-Being. In *Proceedings of the AAAI Spring Symposia*, 365–367, March 2024. 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]. Ithaca, NY: Cornell University Library.

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, March 2024. Palo Alto, CA: AAAI Press.