Challenges in Human-Compatible AI for Well-Being: Harnessing Potential of GenAI for AI-Powered Science

Takashi Kido[1], **Keiki Takadama**[2]
[1] Teikyo University, Advanced Comprehensive Research Organization
[2] University of Tokyo
kido.takashi@gmail.com, takadama@g.ecc.u-tokyo.ac.jp

Abstract

At the AAAI Spring Symposium 2025, we explored the challenges of integrating **Human-Compatible AI** and **AI-Powered Science** to enhance social and individual well-being. Our discussion was guided by two perspectives.

Individual Impact of AI on Well-being: This perspective examines how AI influences personal autonomy, mental health, and emotional fulfillment. It seeks to ensure that AI enhances individual agency, rather than undermining critical thinking and independence.

Social Impact of AI on Well-being: This perspective focuses on the broader societal implications of AI, including fairness, misinformation, and economic impact. It emphasizes AI's role in fostering inclusive social structures while mitigating risks, such as bias and automation-driven unemployment.

This paper provides an overview of the motivations driving our exploration, defines key concepts, outlines major research challenges, and proposes strategies for integrating **Human-Compatible AI** and **AI-Powered Science** in a manner that balances innovation with ethical responsibility.

Motivation

The rapid evolution of **Generative AI (GenAI)** has transformed healthcare, education, and creativity, demonstrating the immense potential for improving individual and societal well-being. However, these advancements have introduced significant ethical and technical challenges. Recent Nobel Prize-winning research underscores AI's critical role in scientific discovery while highlighting the growing concern of ensuring that AI systems align with human values (Swan et al., 2023) (Kido, 2024) (Kido and Takadama, 2024).

One fundamental issue is the **problem of control** in AI, as emphasized by Russell. As AI systems become more complex and autonomous, the risk of misalignment with human intention increases. The over-reliance on AI in decision making, echo chambers caused by recommendation algorithms, and automation-driven inequality highlight the need for a structured approach to AI development (Kido and Takadama, 2019, 2022, 2023) (Kido, 2024)

.

To address these challenges, we proposed a dual framework.

1. **Human-compatible AI**: Developing AI systems that align with human values, ensuring fairness, interpretability, and control (Kido and Takadama, 2024).
2. **AI-Powered Science**: Leveraging AI for real-world applications while mitigating risks related to bias, misinformation, and security.

This symposium examines these perspectives, fostering discussions on how AI can enhance both individual and collective well-being, while maintaining ethical safeguards.

Scope of Our Interests

Artificial Intelligence (AI) is increasingly becoming an integral part of human life, influencing various domains such as healthcare, education, and creative industries. However, ensuring that AI operates in a manner that aligns with human values, while fostering both individual and societal well-being, presents significant challenges. This section delves into the key research areas within **Human-Compatible AI** and **AI-Powered Science**, emphasizing their ethical, technical, and social implications.

Human-Compatible AI

Human-Compatible AI aims to develop AI systems that function within human-defined ethical and societal boundaries, such as ensuring that AI remains beneficial to humans, prioritizing human preferences over machine objectives,

and maintaining a level of uncertainty to avoid unintended consequences (Russell, 2019).

The key focus areas are as follows:

- **Responsible AI for Personalized Healthcare, Education, and Mental Health**: Designing AI models, including **Large Language Models (LLMs)** that prioritize privacy, autonomy, and fairness in delivering personalized services.
- **Interpretable AI for Personal Decision-Making**: Creating transparent AI systems that enhance human trust by making decision-making processes understandable, especially in sensitive fields, such as healthcare and finance (Kido and Takadama, 2019)(Yamanaka and Kido, 2024)
- **AI-Augmented Creativity and Personal Growth**: Exploring the role of AI in assisting human creativity and self-improvement while ensuring that AI-driven recommendations do not erode individual autonomy.
- **Ethical Design Principles for GenAI and LLMs**: Establishing frameworks for AI development that emphasize **transparency, accountability, and alignment** with human values to prevent manipulation and reinforcement of biases.

AI-Powered Science

AI-Powered Science leverages AI technologies to drive innovation and address real-world challenges in various disciplines. The major areas of interest are as follows:

- **Advancements in Bias Detection and Fairness in Machine Learning**: Developing sophisticated techniques to detect, mitigate, and prevent bias in AI systems, thereby ensuring equitable outcomes in different applications (Kido and Takadama, 2022).
- **AI-Driven Computational Sociology and Public Discourse**: Investigating the role of AI in shaping social interactions, information diffusion, and public discourse, with a focus on mitigating the effects of **echo chambers and misinformation**.
- **AI-Driven Societal Transformations and Workforce Transitions**: Analyzing how AI-induced transformations affect employment structures and developing strategies for a **fair and inclusive workforce transition** (Kido and Takadama, 2023).
- **Breakthrough Applications in Healthcare and Beyond**: Studying the impact of **Generative AI (GenAI) and LLMs** in fields such as healthcare, education, and creative industries to maximize their positive societal contributions (Swan et al., 2023) (Yamanaka and Kido, 2024).

By addressing these areas, research in **Human-Compatible AI and AI-Powered Science** aims to ensure that AI technologies contribute to both individual empowerment and societal advancement, while adhering to ethical principles and human-centered design.

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. As the planners of the AAAI SSS25 symposium, we aim to foster discussions that shape the future of the responsible AI. Through interdisciplinary research, ethical frameworks, and innovative applications, we strive to develop AI that is not only powerful, but also aligned with human values.

Acknowledgments

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

References

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. 2022. The challenges for fairness and well-being: how fair is fair? Achieving well-being AI. In *Proceedings of the AAAI 2022 Spring Symposium,* Stanford, CA, March 21–23, 2022, 1–3.

Kido, T. and Takadama, K. 2019. The challenges for interpretable AI for well-being – understanding cognitive bias and social embeddedness. In *Proceedings of the AAAI 2019 Spring Symposia,* Stanford, CA, March 25–27, 2019.

Kido, T., Oono, K., and Swan, M. 2017. The challenges for machine learning and subjective computing in well-being AI. In *Proceedings of the AAAI 2017 Spring Symposia,* Stanford, CA, March 27–29, 2017, 751.

Swan, M., Kido, T., Roland, E., and dos Santos, R. P. 2023. Math agents: computational infrastructure, mathematical embedding, and genomics.