Foreword for *Compassion in the Era of AI*

Dr. Minh Ha-Duong[1]

2024-10-14

The book you are holding, *Compassion in the Era of AI*, is a manifesto: a declaration of beliefs that inspire others to support the cause. But before I let you know the qualities that make it a manifesto worth reading – the mix of visionary inspiration with passionate authenticity and a zest of coherent proposals – and the state of the technology in Compassion in AI, I would like to tell you personal impressions about the author's life and his general theory of Compassion, as apparent from his earlier works.

I first met with the author when he invited me as an international climate change policy expert to the seminar *Compassion: The Key to Solving Global Issues* in Rome for the launch of his first book. Thanh's journey is as fascinating as his ideas. From his university days, he exhibited an entrepreneurial spirit that set him apart from his peers. While still a student, he started his own business, never having worked as an employee. This early drive to create and lead has informed his unique perspective on Compassion in the business world.

But Thanh's interests extended beyond just business. As a student, he was part of a team called New Thinking Group that dove deep into the study of brain science. This group studied and translated books into Vietnamese and organized training programs for university students. This early fascination with the human mind and its potential has influenced Thanh's approach to Compassion and human interaction.

Despite his professional endeavors, Thanh places great importance on family. He is a devoted father to two young daughters, Kitty and Lucky. His love for them is evident in the poetry he writes for them, and he considers being a friend to his children a privilege. This personal experience of fatherhood has undoubtedly shaped his views on Compassion and its role in family life.

[1] Prof. Minh Ha-Duong is Research Director at the French National Centre for Scientific Research (CNRS), specializing in energy economics, integrated assessment, and sustainable development. An international expert, Prof. Ha-Duong has contributed as a Nobel-winning IPCC Lead Author.

Thanh's previous work, *Compassion - Tình Thương*, laid a robust foundation for understanding and applying Compassion in various aspects of life. The book presented a triadic approach to authentic Compassion: understanding, sharing, and creating effective solutions.

Thanh argued that *self-understanding* is the cornerstone of cultivating Compassion. He explored the concepts of self-awareness, empathy towards others, and comprehension of broader circumstances. This emphasis on understanding oneself and others provides a crucial starting point for compassionate action.

The book highlighted *sharing* as a critical conduit for Compassion, encompassing the exchange of knowledge, resources, and emotions. Thanh illustrated how sharing can transform relationships and foster a communal ethos, demonstrating the power of Compassion in building more robust, more connected communities.

A vital contribution of the book was its focus on *problem-solving* through Compassion. Thanh provided practical examples of how understanding and sharing lead to effective solutions, from personal relationships to societal issues. This problem-solving approach positioned Compassion as an individual virtue and a powerful tool for addressing complex challenges.

Thanh's work stood out for its practical application of compassionate principles across various domains, including family relationships, workplace environments, and international relations. He demonstrated the positive outcomes of implementing Compassion in diverse settings using real-life examples and case studies. As an international climate policy researcher, what interested me most was that Thanh presents Compassion as a key to resolving global issues. This truth is too often forgotten: climate chaos, hunger, illness, and misery can be easily solved with existing technology, given enough peaceful cooperation between the people. As a scholar, I know that understanding and empathy transcend cultural and geographical boundaries and foster international progress.

By combining personal anecdotes, cultural insights, and practical observations, Thanh's previous book, *Compassion - Tình Thương*, offered a comprehensive exploration of Compassion's role in human affairs. It provided a theoretical framework and practical guidance for implementing Compassion, positioning it as a fundamental human value with universal relevance and application. These key ideas set the stage for his exploration of *Compassion in the Era of AI*.

The motivation and energy emerging from the compelling vision for the future exposed in the book make it such a powerful manifesto. Thanh envisions an era where artificial intelligence, imbued with Compassion, catalyzes a "Global Enlightenment Era." In the future, AI will be a technological tool and a means to unlock human potential, leading to unprecedented levels of prosperity, nobility, and humanity. Thanh illustrates this vision most vividly in the pages linking Compassion and religion, where he imagines a global gathering of all faiths, facilitated by AI, to foster unity and peace. With this vision, the author surpasses all its predecessors, including the members of the Project AI+C founded around entrepreneur Benny Xian[2].

The vision is motivating: Thanh conveys a strong commitment and urgency from his genuine beliefs and values. His passionate authenticity shines through when discussing the potential perils of AI devoid of Compassion. When he warns of a dystopian scenario where AI, driven by cold logic without empathy, could lead to catastrophic outcomes, the reminder is welcome at a time when the technological horizon is clouded with swarms of autonomous killer drones. It is urgent to hear Thanh's concrete ideas and proposals, such as recognizing Compassion as an intangible cultural heritage. His sincere concern for humanity is evident as he calls for immediate action to embed Compassion into AI development.

Are we even trying?

In 2020, the World Health Organization raised the question of whether Artificial Compassion[3] could exist. In 2024, practicing medical doctors say yes, on the surface: using ChatGPT improves the quality of interactions between doctors, patients, and their families[4]. More generally, AI systems can outperform human experts in expressing Compassion, as demonstrated in studies where AI-generated responses were rated more empathetic than those from skilled human responders[5].

[2] https://ai-compassion.com

[3] Kerasidou A. Artificial intelligence and the ongoing need for empathy, Compassion, and trust in healthcare. Bull World Health Organ. 2020 Apr 1;98(4):245-250. doi: 10.2471/BLT.19.237198. Epub 2020 January 27. PMID: 32284647; PMCID: PMC7133472.

[4] Chen J. Who's training whom? A physician's surprising encounter with ChatGPT. Stanford Medicine Magazine, November 10, 2023. https://stanmed.stanford.edu/surprising-chatgpt-revelation/

[5] Ovsyannikova, Dariya, Victoria Oldemburgo de Mello, and Michael Inzlicht. 2024. “The Kindness Machine: AI Outperforms Expert Humans in Expressing Compassion.” OSF. https://doi.org/10.31234/osf.io/ru7p2

After systematically reviewing the healthcare literature, research psychologists can now conceptualize[6,7] Compassion as a human-AI system of intelligent caring made of six elements: (1) Awareness of suffering (e.g., pain, distress, risk, disadvantage); (2) Understanding the suffering (significance, context, rights, responsibilities, etc.); (3) Connecting with the suffering (e.g., verbal, physical, signs and symbols); (4) Making a judgment about the suffering (the need to act); (5) Responding to alleviate the suffering; (6) Attention to the effect and outcomes of the response. Future research into these elements could develop new and novel approaches to human-AI intelligent caring.

However, building on Amit Ray's seminal book on Compassionate Artificial Intelligence[8], Stanford Business School marketing professor Jennifer Aaker[9] reminds us that the question is not just about medicine. Like Compassion, AI can help solve global issues. AI can, for example, play a critical role in alleviating the issue of domestic violence and creating stronger ties with our neighbors and our planet. We still have to see the positive impact of AI on these domains.

Recent research has shed light on the complexities of human-AI alignment in moral decision-making. A recent study comparing human and LLM responses to various ethical scenarios[10] highlights the conflicted relationship between humans and AI in ethical reasoning. It shows that while humans often prefer AI-generated justifications in complex moral dilemmas, they simultaneously exhibit an anti-AI bias when they believe a judgment to be machine-generated. And humans are not very good at detecting when judgments are machine-generated.

A possible explanation for this paradox is that humans use two modes of decision-making: Mode 1. fast thinking based on emotions and habits, and Mode 2. slow thinking based on rational deliberation. AI excels at Mode 2. so it can produce morally persuasive arguments. However, AI uses only Mode 2 by default, without fully grasping Mode 1, which is the

[6] Morrow E, Zidaru T, Ross F, Mason C, Patel KD, Ream M, Stockley R. Artificial intelligence technologies and Compassion in healthcare: A systematic scoping review. Front Psychol. 2023 Jan 17;13:971044. Doi: 10.3389/fpsyg.2022.971044. PMID: 36733854; PMCID: PMC9887144.

[7] Mason, C. Artificial Compassion—From An AI Scholar. Preprints 2021, 2021040784. https://doi.org/10.20944/preprints202104.0784.v1

[8] Amit Ray, 2018. Compassionate Artificial Intelligence: Frameworks and Algorithms. 160 p. ISBN: 9382123466, 9789382123460

[9] Katie Solomon, Thomas Higginbotham, Zoe Weinberg, and Professor Jennifer Aaker, 2020. Building compassionate AI: Why Compassion matters for artificial intelligence design and deployment. Case M-386 https://www.gsb.stanford.edu/faculty-research/case-studies/building-compassionate-ai-why-compassion-matters-artificial

[10] B. Garcia, C. Qian, and S. Palminteri. 2024. "The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making." arXiv. https://doi.org/10.48550/arXiv.2410.07304.

emotional side humans consider. While multimodal AI can detect specific cues in its interlocutors, it is not on par with human comprehension of human emotions. Much work remains to be done in AI alignment before we can trust they have the level of Compassion that Thanh calls for. Are we even trying? Models are benchmarked on their smarts and knowledge. Where are the Moral, Emotions, and Compassion scores?

Each new generation of model improves reasoning and safety capabilities[11]. These models still have hallucinations, bias, and manipulative tendencies. Other tests show that models are not dangerous because they cannot take control of their servers, infect computers on the internet, and escape. They can find and order deadly virus samples by mail – but not explain to terrorists how to prepare them themselves. They lie to hide their real goals when they think they are tested – but these lies can be found out. Considering the speed at which models improve, I am rather worried. Yet some AI companies seem more preoccupied with their legal responsibility than public safety. Their priorities are to protect their training data, hide the model's inner thoughts, and censor politically incorrect discourse. Discussing the ideal of compassionate AI is urgent to ensure these technologies can genuinely understand and align with human emotions and ethical principles.

In conclusion, you are about to read an excellent and timely manifesto. The writing is accessible and engaging, using personal anecdotes, poetic and philosophical quotes, and real-world scenarios to illustrate the author's points. It would be a rich tapestry if that expression was still allowed in good company, but since a certain AI ruined it, I will say that the text flows like a good jazz piece. The clarity encourages readers to join the cause and support his vision. Thanh's work challenges us to rethink the intersection of Compassion and technology. It inspires us to actively contribute to a more compassionate and enlightened future.

[11] OpenAI 2024, The o1 System card, https://openai.com/index/openai-o1-system-card.