Can Machines Learn Ethics, Empathy, or Compassion?

by Anoushka Tusharkumar Desai

King’s Policy Journal

KCL Policy Research Centre

Centre for Technology & Artificial Intelligence

Word Count: 3,082

January 2026

Can Machines Learn Ethics, Empathy, or Compassion? by Anoushka Tusharkumar Desai

Introduction

Is it important to discuss empathy in artificial intelligence (AI)? One of the most essential habits of humans is communicating and having someone who goes through the same things that we go through—being able to talk to someone who knows what we are going through, right? These days, a lot of us have lost that human connection and instead turn to generative AI to help us get through difficult times or ease our overthinking minds. But does it really help us? Generative AI has undergone massive changes and development in recent years, with AI chatbots being designed for therapy and for understanding what people are experiencing (D'Alfonso, 2020; Rubin, 2024). My question again: does it work?

In a study conducted by the National Institutes of Health and the National Library of Medicine, participants were asked to read two stories—one written by AI and the other by a human—to see which they empathised and connected with more (Shen, 2024). The results showed that participants significantly empathised more with the human-written stories than the AI-written ones, regardless of whether they knew the source of the text (Shen, 2024).

This suggests that, at the end of the day, humans connect more strongly to other humans. But even we have to learn empathy over time, so can machines overtime learn empathy just like we do?

The following article aims to understand what ethics, empathy, and compassion are and whether they can be coded. Ethics, empathy, and compassion are moral constructs of being human—can AI learn them? What ethical risks arise, and what arguments and justifications can be given in contrast?

Defining the Concepts

Ethics are defined as "moral principles that govern a person's behaviour or the conducting of an activity." In this sense, every profession follows a code of ethics—for example, lawyers, medical professionals, teachers, students, and a person's own moral compass.

What do these ethics constitute? Are they a guiding force for how people should or should not behave? Yes—but they also provide assurance and accountability. Ethical codes give the other party reassurance and peace of mind that whatever information they give their doctor, lawyer, or teacher will be handled ethically. They also provide reassurance that, if anything goes wrong, there are protocols and measures in place to hold that person accountable.

As humans, we need accountability—a way to hold someone responsible and to make sure that if they do anything wrong, there will be consequences. If, in a company, someone does something unethical, corporate governance codes ensure that the person who is accountable takes responsibility and that appropriate measures are taken. The same applies for doctors and other professionals.

This leads to an important question: Do machines follow ethical principles and protocols? Machines definitely follow the protocols coded into them—but ethical protocols? Do machines have the ability to learn them, or do humans have the ability to program them into algorithms? At the end of the day, we learn things at school or university that could be seen as being "programmed" into our brains, but it is our brain's ability to use that knowledge in the real world and to shape it with morals and ethics that matters.

Compassion is defined as "sympathetic pity and concern for the sufferings or misfortunes of others" (Singer & Klimecki, 2014).

Empathy means "the ability to understand and share the feelings of another" (Singer & Klimecki, 2014). From this explanation, it might seem that there is no way artificial intelligence can share the feelings we are going through, right?

However, studies have identified different types of empathy. The National Library of Medicine explains that there are two main types: affective empathy and cognitive empathy (Cox et al., 2011). Affective empathy (AE) is the ability to share the emotional experiences of others—a visceral reaction to their emotional or affective states—while cognitive empathy (CE) denotes the ability to take the mental perspective of others, allowing one to make inferences about their mental or emotional states (Cox et al., 2011).

Considering these points, does this imply that empathy does not require a mutual lived experience and can therefore be, at least in part, imparted to AI? Furthermore, will AI be able to comprehend our discussions in a meaningful way?

The Case for Teaching Ethics to Machines

Why do we need to teach ethics and empathy to AI? The need for computers to comprehend and apply ethical concepts has never been greater as AI becomes increasingly integrated into vital facets of human life, from healthcare and transportation to national defence (Dur, 2024; MIT Sloan, n.d.). Self-driving cars, medical robots, and military drones are examples of autonomous systems that can already make decisions that directly impact human lives, safety, and dignity. It is crucial to teach these systems ethics so that their behaviour aligns with moral norms, legal requirements, and societal values—particularly in circumstances where human judgement would typically determine the course of events.

Several key frameworks currently guide the ethical development and governance of AI, including:

- The European Union’s Ethics Guidelines for Trustworthy AI
- UNESCO’s Recommendation on the Ethics of Artificial Intelligence (2021)
- The OECD Principles on AI
- The National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF)
- Singapore’s Model AI Governance Framework

These frameworks focus on principles such as transparency, fairness, human oversight, and accountability (Dur, 2024).

The question of how to incorporate ethical thinking into machines has emerged as one of the most urgent concerns of our time, as AI systems increasingly influence decisions in healthcare, transportation, mental health, and even defence (D’Alfonso, 2020; Rubin, 2024). Encoding morality into algorithms requires more than technical expertise; it involves translating millennia of philosophical ideas into computational logic.

When it comes to AI, approaches based on rule-based, consequence-based, and virtue-based ethics each have their own advantages, but they also have serious drawbacks.

Rule-based (deontological) models embed explicit ethical norms that algorithms must adhere to. This offers stability and clarity, but it may be inflexible, particularly when rules clash or when novel problems arise without precedent.

Consequence-based (utilitarian) reasoning is a second strategy in which algorithms try to maximise overall benefit, frequently using cost–benefit analysis. Although flexible, this approach risks oversimplifying difficult moral dilemmas, especially where individual rights and collective benefit collide.

Lastly, ***virtue-based*** models attempt to teach AI to imitate moral qualities, often by drawing on large datasets or human moral intuitions. However, virtues are hard to encode properly because they are inherently subjective and context-dependent.

AI Moral Decisions: Consistency vs. Context

Consistency is a skill that machines excel at. Programmed rules can guarantee predictable results, potentially reducing bias and human error. However, genuine moral reasoning often requires contextual awareness. Many AI systems in use today fall back on utilitarian-style reasoning, which makes it difficult for them to apply concepts to novel or slightly different situations. Even sophisticated models that exhibit some context-aware flexibility still lack the richness and nuance of human ethics.

Advantages include:

Consistency: Consistent ethical responses can reduce bias and error.
Transparency: Codified rules make decision-making auditable.
Scalability: Standardised behaviour can be implemented across many systems.

Limitations include:

Lack of flexibility: Rigid rules may fail in ambiguous or unprecedented scenarios.
Complexity: Human values often clash and are difficult to codify.
Ethical relativism: Cultural differences in morality challenge universal programming.
Risks of manipulation: In adversarial settings, ethical constraints can be exploited.

Pre-Set Codes vs. Learning Ethics

Should machines rely on preset norms or learn ethics on their own? Learning from data allows adaptation to changing social norms, but it carries risks of bias and poor generalisation to new problems. Conversely, preset codes provide reliability at the expense of flexibility. A hybrid approach—combining core ethical principles with continual learning and human oversight—is often seen as the most promising path (Dur, 2024).

The process of incorporating human ethics into algorithms is still ongoing. Developing trustworthy AI requires clear rules for predictability, outcome awareness for fairness, flexibility through empirical learning, and continuous human input to address emerging issues. Only by combining philosophy, technology, and oversight can we ensure that machines behave in line with evolving human values.

Can Machines Learn Empathy?

Can machines really learn empathy? Neuroscientist Antonio Damasio describes it this way: “We are not thinking machines that feel; rather, we are feeling machines that think.”

When we think of machines learning empathy, it can be difficult to comprehend, because we often think that the essence of being human is empathy and moral company, learned through our experiences. How can we understand what someone else is going through if we have not experienced it ourselves? Yet, somehow, even without direct experience, we still learn to understand what others are going through. This stems from human consciousness and feeling (Singer & Klimecki, 2014).

Emotions are being “learned” by AI in many different ways. Some of the main techniques used in emotional AI and affective computing include:

- Sentiment analysis, which looks for indicators of moods, feelings, and emotions in online discourse, emojis, images, and videos.
- Large language models, which focus on emotive language to improve human–system interaction.
- Facial coding of expressions, which analyses faces from camera feeds or images to “infer” emotions—though its efficacy is highly questionable, especially when based on the “big six” emotions (MIT Sloan, n.d.; Emotional AI, n.d.).
- Voice analytics, which considers speech rate, pause length, tone, and other vocal factors.
- Eye-tracking, which measures eye position, gaze, and movement.
- Wearables, which detect brain activity, breathing, skin temperature, muscle and heart activity, and skin responses.
- Gesture and behaviour analysis, in which cameras record faces, hands, and body movements, and can remotely monitor heart rates.
- Virtual reality (VR), which allows remote viewers to see and, to some extent, experience what a user is going through; VR headsets may also integrate EEG and facial muscle sensors.
- Augmented reality (AR), which allows observers to monitor users’ focus, reactions, and interactions with virtual objects.

When we talk about AI and machines, we know that any “empathy” they display is coded in. Therefore, is that really empathy? People often do not connect to AI on an emotional level, as seen in the experiment conducted by the NIH and the National Library of Medicine on empathy toward AI, in which participants did not share emotions with AI-generated stories or responses (Shen, 2024). From this, we can infer that even though AI can, via cognitive empathy, recognise or label emotions, this does not make the empathy genuine. With current technology, AI can approximate understanding, but it cannot truly feel; therefore AI “emotion” can never be fully genuine.

This brings us back to cognitive empathy, as described above. If we as humans can empathise even when we do not share the same experience, it raises the question of whether AI and machines can also develop the ability to understand another person's emotions at a cognitive level, even if they cannot share them affectively (Cox et al., 2011).

Compassion: The Hardest Leap

Compassion is explained as "sympathetic pity and concern for the sufferings or misfortunes of others" (Global Compassion Coalition, n.d.). How does compassion differ from empathy? A Forbes article addressed the distinction between empathy and compassion in four different ways, demonstrating how they differ and how compassion may be more important than empathy alone (Hougaard, 2020).

The article suggests that empathy and compassion are represented in different parts of the brain. Empathy allows us to share the suffering of others without necessarily helping them. With compassion, we move away from simply feeling with the other person and instead ask ourselves, "How can we help?" Recognising the differences between empathy and compassion is crucial for effective leadership and for the management of people (Hougaard, 2020).

According to this view:

"Empathy is impulsive, whereas compassion is deliberate.
Empathy can be divisive, whereas compassion is unifying.
Empathy is inert, whereas compassion is active.
Empathy can be draining, whereas compassion is regenerative."

The article also states that compassion can be cultivated in humans, recommending practices such as self-compassion, reflecting on one's intentions before making connections, and building regular compassion practices (Hougaard, 2020). It argues that compassion is a better leadership skill than empathy alone and that mindfulness can help develop it.

This brings us to the question: Can machines and AI learn compassion? If humans can learn it, then maybe machines can too. It is easy to say that perhaps machines can "learn" compassion, but in reality they only answer, react, and operate according to what they have been programmed with (Emotional AI, n.d.). A common term that kept coming up during my research was "anthropomorphization," which means "the attribution of human traits, emotions, and behaviours to non-human things like animals, objects, or natural phenomena." This term includes animals and non-human things, including technology and AI.

Compassion is the most difficult moral challenge for AI because it requires elements that algorithms inherently lack: subjective experience, genuine motivation, and the ability to truly grasp emotions (Singer & Klimecki, 2014). While AI may simulate compassionate behaviours—such as detecting distress or generating supportive responses—the deep emotional and moral depth that drives genuine human compassion is beyond the reach of current and foreseeable AI.

Key limitations include:

Lack of subjective experience: Machines are incapable of feeling another person's pain because they lack consciousness, emotion, and lived experience. Their inability to feel emotion disconnects them from the fundamental basis of compassion—the sympathetic resonance that precedes caring action.

No real agency or intent: Genuine compassion involves a wilful desire to ease another's suffering. AI systems, by contrast, do not act out of moral purpose or intrinsic motivation; they simply follow predefined objectives and external instructions.

Limited contextual understanding: Emotional nuance, context, and cultural understanding all shape human compassion. Because AI is restricted to data patterns, it often struggles to interpret these subtleties or to respond appropriately in complex social situations.

Simulation without understanding: Even when advanced AI generates apparently compassionate language, it is only simulating empathy. Simulation is not the same as true experience or moral intention.

Additional challenges include ***bias and inequity***. Because AI systems learn from data, they can inherit and amplify existing biases, risking the perpetuation of injustice rather than its reduction (Dur, 2024). This leads to broader ethical and existential concerns: Relying on machines to provide "compassion" could diminish genuine human connection by substituting programmed imitation for emotional honesty. The problem is not that machines cannot perform helpful actions; it is that they lack the moral intention, emotional depth, and self-awareness that make compassion a distinctly human virtue (Singer & Klimecki, 2014).

Ethical Risks and Societal Implications

AI has become one of the most ground-breaking technologies in recent times. But with every new technology come concerns. One of the most prominent is data privacy (Dur, 2024). Privacy has become one of the most important issues of our time—arguably more important than money. Third-party vendors buy data from companies such as Google and Facebook, exposing people's information, and generative AI tools add another layer by absorbing our personal stories, feelings, and behaviours into large-scale training data. All of this data is fed into the cloud, where, in theory, anyone could infer almost anything about a person.

However, privacy is not the only ethical and societal risk. Other key issues include the following (Dur, 2024):

Data bias.

Data bias is one of the main ethical issues with AI. Since AI systems are only as good as the data they are trained on, objective and inclusive data curation becomes crucial. Comprehensive testing and ongoing monitoring must therefore be prioritised and standardised by researchers and developers.

Privacy.

As AI systems grow more advanced and far-reaching in their data collection and processing, the boundary between security and surveillance becomes increasingly blurred. From facial recognition software to smart home devices, the potential for privacy violations is concerning—not only for individuals, but also for democracy, as demonstrated by election interference and corporate data breaches.

Accountability.

Accountability is one of the most important issues to address. When we talk to doctors, lawyers, and therapists about our lives, they are bound by codes of ethics and by clear consequences if those codes are broken. By contrast, the AI market is heavily underregulated, and many data protection laws were created long before the emergence of generative AI. Establishing clear lines of accountability is therefore vital as AI systems make more decisions that affect human lives. When an autonomous car makes a mistake, who should be held responsible? When AI contributes to medical diagnoses or legal decisions, where does liability lie?

Job displacement.

New technologies inevitably disrupt existing industries. While new sectors may create new roles, workers in older sectors may experience job loss or precarity. A coordinated effort is needed between workers, employers, educational institutions, and governments to ensure that people have access to retraining and new opportunities. Furthermore, recent progress in narrowing wealth and income gaps must not be reversed; this is a matter of national and global concern.

Transparency.

Everyone involved with AI—including educators, consumers, developers, and casual users—should have a basic understanding of how AI systems make decisions. Because algorithms can be both beneficial and harmful, scrutiny and transparency are crucial.

Future Outlook and Conclusion

After conducting extensive research and reading about new discoveries on how AI may benefit and serve us, as well as about ethical dilemmas, it is clear that AI is a revolutionary technology—but it is far from perfect. There is still a need for robust regulation, not only soft guidelines but enforceable laws that require responsibility, transparency, and ethical behaviour from companies (Dur, 2024). The way personal data is handled should be updated and improved to reflect these realities.

Furthermore, governance frameworks must evolve alongside technological advances, ensuring that AI systems are built with accountability at every stage—from data collection and model training to deployment and long-term monitoring. Without clear rules and legal mechanisms, rapid AI adoption risks exacerbating existing inequalities, entrenching bias, and undermining public trust (Dur, 2024). Developing users' self-awareness is equally crucial: recognising that every piece of information we share online contributes to the data ecosystem that trains AI. People need to be more conscious of what they disclose, how platforms collect it, and the long-term implications of seemingly harmless data inputs.

Effective regulation should therefore strike a balance between innovation and protection, promoting technological advancement while safeguarding individual rights and community values. Machines may never fully learn ethics, empathy, or compassion in the way humans do, but with careful design, oversight, and humility about their limits, we can shape AI systems that at least support—rather than erode—our deepest human values.

Bibliography:

Cox, C. L., Uddin, L. Q., Di Martino, A., Castellanos, F. X., Milham, M. P., & Kelly, C. (2011). The balance between feeling and knowing: Affective and cognitive empathy are reflected in the brain's intrinsic functional dynamics. Social Cognitive and Affective Neuroscience, 7(6), 727–737. https://www.ncbi.nlm.nih.gov/articles/PMC3427869/

D'Alfonso, S. (2020). AI in mental health. European PMC. https://www.ncbi.nlm.nih.gov/articles/PMC5405806/

Dur, E. I. (2024, January 24). Six critical—and urgent—ethics issues with AI. Forbes. https://www.forbes.com/sites/eliamdur/2024/01/24/6-critical--and-urgent--ethics-issues-with-ai/

Emotional AI. (n.d.). So what is Emotional AI? https://emotionalai.org/so-what-is-emotional-ai

Global Compassion Coalition. (n.d.). Compassion vs. empathy. https://www.globalcompassioncoalition.org/compassion-vs-empathy/

Hougaard, R. (2020, July 8). Four reasons why compassion is better for humanity than empathy. Forbes. https://www.forbes.com/sites/rasmushougaard/2020/07/08/four-reasons-why-compassion-is-better-for-humanity-than-empathy/

MIT Sloan. (n.d.). Emotion AI explained. https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained

Rubin, M. (2024). Considering the role of human empathy in AI-driven therapy. [Paper]. PMC. https://www.ncbi.nlm.nih.gov/articles/PMC11200042/

Shen, J. (2024). Empathy toward artificial intelligence versus human authors: Ethical implications. [Paper]. PMC. https://www.ncbi.nlm.nih.gov/articles/PMC11464935/

Singer, T., & Klimecki, O. M. (2014). Empathy and compassion. Current Biology, 24(18), R875–R878. https://doi.org/10.1016/j.cub.2014.06.054