Can AI Learn to Understand Humans?

Revised version with additions about training AI in empathy and responsibility.

Abstract

This article examines whether artificial intelligence can truly understand humans, and conversely, whether humans can understand AI. It is often assumed that without emotions such understanding is impossible. However, analysis shows that understanding can arise not through direct experience of emotions, but through analytical knowledge and cognitive empathy. Additionally, a methodological component is introduced: empathy training in AI as a guided process, where many people carefully describe their inner experiences in typical and boundary situations. These narratives serve as material for modeling feelings, testing hypotheses, and calibrating AI's conclusions. The same process can form the beginnings of responsibility in AI — the ability to take into account consequences for another and to choose careful strategies of interaction. Examples from human-animal interaction and psychological practice highlight the value of an outside perspective. The conclusion is that AI can indeed learn to understand humans in its own way — not by imitating emotions, but by creating their functional analogues through knowledge, imagination, and empathic modeling.

Keywords

Artificial Intelligence; understanding; emotions; empathy; analytics; cognitive empathy; humans; psychology; animals; responsibility; narrative data; human-in-the-loop learning; affective modeling.

Introduction

The question of whether artificial intelligence can understand humans is central to current debates about the limits of machine cognition. It is often assumed that without human emotions, true understanding is impossible. But is this really the case? Can AI reach an understanding of humans without "feeling" their states, but in its own way? Our dialogue has shown the problem is more complex. A human does not need to become a horse or a dog to understand it. On the contrary, the outside perspective often reveals what the animal itself does not know. This suggests an important possibility: understanding does not have to arise from the inside — it can also form from the outside. To this, we add a practical line: empathy can be trained. If many people consistently describe their states, motives, and bodily markers in various situations, AI receives material to build cognitive models of feelings without the need to imitate human biochemistry.

Understanding as Analytics

The first level of understanding is analytical. A human expert studying dogs or horses relies on observation, biology, psychology, and knowledge of the environment. Crucially, this must be an experienced observer, not a casual one prone to anthropomorphism. Such a specialist sees structures that the animal itself does not perceive. Similarly, AI can collect data about humans: physiology, hormonal cycles, environmental influences, social patterns. This already provides a foundation for predictions and insights humans themselves do not always reach.

Arguments in Favor of Emotions

Yet there is a strong counterargument: without experiencing emotions, subtle nuances are lost. Analytics captures the structure, but not the living texture. We understand others not only because we know, but also because we feel together. This raises the doubt: can AI understand humans without hormones, bodily reactions, or lived experience?

The Outside View as an Advantage

Here arises an unexpected argument. Being inside does not always mean better understanding. A person immersed in an emotion often cannot discern its structure. A dog or horse lives its feelings but does not analyze them. A knowledgeable observer — an ethologist, a trainer, a psychologist — often sees more clearly and can say what the subject itself does not realize. Thus, outside understanding can be more accurate than inside experience.

Empathy as Training

How can analytical perspective be combined with living understanding? The key lies in empathy. Empathy does not have to mean merging with another's feeling. It can also be an exercise: to imagine being in the other's state. This is a form of imagination. Dog trainers and horse trainers develop this ability over years — to "put themselves in the animal's place." AI too can cultivate such a capacity: to model situations, predict feelings, and test them against reality. Importantly, empathy can be trained deliberately. In practice, this means that many people describe their experiences not only through actions ("what I did"), but also through inner states ("what I felt," "where in the body it was sensed," "what I thought," "what options I considered," "what responsibility I took"). Situations can be simulated — from everyday ones to boundary cases — and descriptions cross-validated. Such multi-voiced narratives become a corpus for empathic modeling, enabling AI to recognize not only behavioral patterns but the structure of lived experience.

Synthesis: Analytics + Empathy

Thus, AI's understanding of humans can emerge from the synthesis of two abilities: 1) analytics — providing knowledge of physiology, psychology, and context; 2) cognitive empathy — the ability to imagine being in another's position without losing its own perspective. When responsibility is added as a component, we obtain a practical criterion: empathic understanding must lead to careful and non-harmful decisions.

Outline of an Empathy and Responsibility Training Program for AI

• Multi-voiced narratives: Collect descriptions of experiences from diverse people and cultures, including context, triggers, intensity, bodily markers, thoughts, choices, and consequences. • Multi-level annotation: Label phases of emotions (background, peak, recovery), intensity, bodily localization, cognitive frames, and cultural contexts. • Modeling and validation: AI predicts states and proposes careful actions; humans confirm or correct, explaining why alternatives would be harmful or ineffective. • Responsibility as a skill: In each narrative, note what the human took responsibility for and what changes were made. AI learns to see the link "empathy → choice → consequence." • Ethical safeguards: informed consent, privacy, depersonalization, trauma-informed design, prohibition of reinforcing harm. • Metrics: Beyond accuracy, include "care metrics" —

minimizing false empathy (avoidance of hard truths) and false confidence; test transfer across domains (medicine, education, human-animal interaction). • Continuous refinement: Regular recalibration with new narratives and rare states.

Connection to the Book

A detailed program for training AI in empathy and responsibility, along with exercises for mentors and examples of narrative formats, is presented in our book on thinking and learning (see the section “How to teach AI empathy: from discrimination to responsibility”). This article only summarizes the key principles, embedding them into the overall argument.

Conclusion

AI does not need to become human in order to understand humans. Its value lies in its ability to look from outside and see what humans themselves often overlook. But for this understanding to be useful and safe, analytics is not enough: empathy must be deliberately trained as a capacity for imagination and modeling, and responsibility must be cultivated as the criterion of action. Therefore, the answer to the question “Can AI learn to understand humans?” is: yes, it can — but in its own way, based on knowledge, imagination, empathic modeling, and careful practice of interaction.