

Johannes Giesinger

Simulating Empathy? Generative AI in Education

*Fragile Futures: Technology, Responsibility and the Idea of Education.
European Seminars in Philosophy of Education, American University of
Paris, June 4, 2026*

1 Introduction

As soon as generative AI in the form of large language models (LLMs) became available to the general public in 2022, students and teachers started using them. The use of LLM chatbots by students was problematised in three ways immediately. Firstly, there is the issue of evaluating students' performance. Using LLMs in testing situations can lead to distorted outcomes that do not reflect students' actual knowledge and skills. If only some students use LLM chatbots, it creates unfair testing conditions.

Secondly, using chatbots may hinder the acquisition of skills and knowledge. For example, how can students learn to write an essay if they rely on LLMs from an early stage in their education? It is clear, for example, that students' mental arithmetic skills decreased as calculators became more widely used.

This raises a third question about the importance or value of certain skills. Just as the role of mental arithmetic in schools is debated, the importance of writing skills may also be questioned given that AI can generate all kinds of texts of an appropriate quality, from WhatsApp messages to song lyrics and scientific papers. LLMs therefore also raise issues regarding the aims of education.

While these issues have been widely discussed, I would like to focus on one that has not yet received much attention: As chatbots evolve, they are increasingly becoming quasi-personal counterparts for students. Students may develop a relationship with them as learning partners or teachers. As well as using them to gain information, students may also start conversations with chatbots, talking to them about a variety of issues, including personal problems and their emotional states. In particular, students may come to trust LLM chatbots to some extent and feel understood, validated, recognised or encouraged by them. They potentially see them as 'empathic' partners in a similar way to how they might see a teacher or peer.

Clearly, empathy and related mental states and attitudes are essential for positive educational relationships. It is also important that students are able to trust their teachers. Against this background, it could be argued that LLM chatbots used in education, or specifically designed for educational purposes, should simulate the empathic behaviour of teachers as much as possible, in order to promote learning.

This argument is outlined in the next section, while the third section discusses three objections. The fourth section proposes how chatbots should be designed and how people should interact with them. It highlights the notion of a 'transparent partnership' between students and chatbots.

2 An argument for empathic chatbots

The argument is based on three assumptions: First, empathy and related emotional attitudes are central to educational relationships. Second, LLMs are capable, to some extent, of simulating empathy. Third, learners react to a chatbot's expressions of empathy in ways that benefit their learning. Let us consider the first assumption more closely. In this context, it is important to distinguish between descriptive and normative perspectives, as well as between the emotional and cognitive aspects of empathy. We can also approach the issue from either the learners' or the teachers' standpoint.

In order to learn from teachers or peers, learners need to have access to their mental states, thoughts, intentions and emotions. They need to understand what others are referring to when they talk, what they mean and what they intend. They need to be able to focus on an object jointly with others, with everyone understanding that they are all focused on the same object. This capacity, known as 'joint attention', evolves in the first years of human life (Eilan et al., 2005). While some researchers focus on the cognitive dimension of understanding mental states (Tomasello, 1999), others highlight "interpersonal engagement involving feelings" (Hobson, 2005, p. 188) as the starting point for joint attention. Peter Hobson writes, "The infant engages with someone else's engagement with the world — and is 'moved'" (ibid.). According to this approach, some form of emotional connection with others – Hobson does not use the term 'empathy' – is a prerequisite for understanding them. Therefore, while learners need to understand their teachers cognitively, it is unclear whether teaching necessarily involves empathy. It is possible for a

teacher to convey their knowledge without considering their students' mental states and still facilitate learning. It is important to note that these considerations are descriptive, providing an empirical view of the conditions necessary for social and cultural learning. This must be distinguished from a normative account of what constitutes a 'good' educational relationship. Here, we can examine the relationship itself and its ('intrinsic') value for those involved, as well as the instrumental value of the relationship in promoting learning. As previously mentioned, learning may also be possible in educational settings where teachers lack empathy. In this context, empathy can be distinguished from an attitude of empathetic concern or care. Empathy as a capacity may also be used to manipulate or humiliate others. However, teachers should empathically care for their students in the sense that they should take a genuine interest in their emotional well-being and make them feel understood and recognised.

However, even if learners do not feel understood, they may still be able to progress in their learning. Therefore, it makes sense to link considerations regarding the instrumental value of educational relationships to an account of the intrinsic educational relationship goods. For example, it is valuable for learners to feel understood by teachers or peers. Therefore, teachers should demonstrate empathic care, regardless of its impact on learning. At the same time, we can assume that empathic care will benefit learning. In this regard, we can refer – for instance – to the findings of empirical research in the field of 'mentalisation theory', even though this theory mainly concerns psychotherapeutic relationships. The basic insight here is that mentalising on the part of therapists fosters epistemic trust on the part of patients. As Peter Fonagy and others use the concept, mentalizing combines cognitive understanding of others' mental states with empathy as an emotional capacity. In teaching practice, mentalising means, firstly, trying to understand learners and, secondly, expressing one's interpretations of their mental states, which may help them to gain better access to them themselves. It is claimed that mentalising in education fosters learners' capacity for mentalising and, moreover, leads them to trust their teachers (Fonagy & Allison, 2014; Duschinsky & Collver, 2019).

Let us now turn to the second assumption: that LLMs can exhibit empathic communication. Here, we can follow Catrin Misselhorn (2021),

who first clarifies that AI systems cannot be empathic. In the course of her argument, Misselhorn proposes a definition of empathy involving three conditions. Firstly, empathy means feeling the same emotions as the other person – in other words, my feelings must be ‘congruent’ with theirs. Crucially, cognitive understanding of the other person's feelings is not sufficient.

Secondly, however, feeling the same emotions does not mean that they are identical. Empathy involves an ‘asymmetry’ between my feelings and yours. For example, if you are sad, I will feel your sadness, but I will not be sad in the same way. However, I will be able to understand your sadness emotionally.

Thirdly, being empathic also means that it must always be clear to me that you are a separate person with your own mental states. Empathy is specifically about feeling another person's emotions, not merging with them.

Against this backdrop, Misselhorn (2021, p. 81) asserts that AI systems lack the capacity for empathy. Crucially, they cannot have feelings and are therefore incapable of empathising with others. Clearly, they fail to meet even the first of the three conditions. While there is some doubt as to whether computer systems labelled as ‘artificial intelligence’ will someday be capable of having feelings, or more broadly, acquiring some form of ‘consciousness’, we can safely assume that current LLMs do not have mental states and lack the capacity for empathy.

However, they can communicate empathetically. For example, they can put together combinations of words and sentences that are understood as empathic by human users. LLMs achieve this by picking up linguistic cues from users, e.g. indicators that they are sad, angry or afraid. The simplest way to react empathetically is to acknowledge these emotional states ('I understand that you must be very sad') and validate them, i.e. consider them appropriate and acceptable. On this basis, various forms of emotional care and support can be offered.

More sophisticated AI systems that integrate visual and auditory cues using cameras and microphones will be even more proficient at recognising and reacting to emotional states appropriately. While these quasi-empathic responses can always be misinterpreted, there is a chance that they will be understood as empathic, in that the user will feel ‘seen’.

This leads to the third assumption: users' positive responses to empathic chatbot communication. Here, I am not interested in cases where a chatbot poses as a human being, e.g. cases where users are intentionally deceived into thinking they are communicating with a human when they are actually communicating with a chatbot. I am only interested in situations where people are aware, to some extent, that their counterpart is a machine. People seem to have a natural tendency to 'humanise' beings or objects that exhibit human-like traits. Therefore, even if users do not believe that the chatbot is a person, they may still recognise it as a partner with whom it is possible to have a quasi-interpersonal conversation. In this way, the relationship with an LLM may function as a substitute for some important features of interpersonal relationships. As outlined, this could indeed lead the user to feel understood.

Along these lines, it also seems possible for learners to develop a certain degree of trust in LLM chatbots that goes beyond mere epistemic reliance on their outputs. Depending on the epistemic quality of the chatbot, it makes sense to ascribe some degree of 'credibility' to it, meaning that there is reason to believe what it says. While this kind of trust may function on a purely cognitive level, it may also be accompanied by an emotional attachment to the non-human counterpart, reinforced by the chatbot's quasi-empathic responses, which create a sense of being understood, seen, validated and recognised. As indicated in the mentalisation theory, this kind of communication may also foster learners' epistemic trust in the chatbot.

Against this backdrop, we can therefore argue for the design of chatbots for educational purposes that simulate empathy as much as possible.

1. 'Good' educational relationships, in which teachers exhibit empathy and empathic concern, and learners show epistemic trust towards their teachers, tend to be beneficial for learning.
2. AI chatbots using LLMs can simulate empathic communication to some extent.
3. Learners tend to react positively to chatbots quasi-empathic communication. This tends to be beneficial for learning.
4. Therefore, chatbots used in education or developed specifically for educational purposes should simulate empathic communication as much as possible, with the aim of establishing a quasi-personal relationship with learners.

5. Teachers should also encourage learners to see chatbots as quasi-personal counterparts and to communicate with them in a similar way to how they would communicate with other people.

3 Objections

There are a number of possible objections to this line of argument. I would like to focus on three of them. First, it could be argued that developing a quasi-personal relationship with chatbots, which entails some kind of epistemic trust in them, can make learners vulnerable to epistemic harm. Second, it could be argued that it can distort the emotional and social development of learners. The third objection, which I consider decisive, is that developing an emotional bond with a chatbot is ‘inadequate’, regardless of the possible consequences for individual development.

Let us consider the first objection. When students trust an LLM chatbot, they can clearly be deceived and manipulated, and may come to believe things that are not true. In other words, they may suffer epistemic harm. To address this, we should first examine the role of trust in the relationship between teachers and learners. As I would suggest, students should trust their teachers, but not blindly believe everything they say. On the one hand, being extremely distrustful or ‘hypervigilant’ can inhibit the learning process. On the other hand, however, the aim of education cannot be to promote credulity. Rather, epistemic trust should be considered the ‘default position’, at least in an educational context. This means that students should trust their teachers unless there are reasons not to. If teachers have previously proven themselves to be incompetent or dishonest, there are reasons not to trust them.¹

Considering trust, rather than distrust, as the default position seems justified in education because of the specific institutional setting and teachers' duties towards their students. One of the core duties of teachers is to teach what is correct and relevant, and in well-functioning education systems, teachers are selected according to competence standards and trained to fulfil their professional duties. Given that the education system can generally be trusted, it is reasonable to consider epistemic trust as the default position for students.

¹ This question might be discussed referring to debates in social epistemology – namely, the debate on whether testimony from others is a way of gaining ‘knowledge’ (for educational considerations on this issue, see Simion, 2020).

This view cannot be applied directly to LLM chatbots. They should not be trusted by default. However, we can assume that it is the responsibility of teachers — or those involved in education more broadly — to select and monitor the use of LLM chatbots in schools in accordance with epistemic standards. If chatbots are designed specifically for educational purposes, without commercial or political interests, they should be highly reliable in epistemic terms. However, it cannot be expected at this point that they will not also generate epistemically problematic content. This is why teachers should introduce students to epistemic practices, enabling them to apply epistemic standards to information gained through chatbots.

The second objection is not related to epistemic harm, but to the potential impact on the emotional and social development of students. It should first be noted that some users, particularly younger children, may start to believe that they are communicating with a real person when they are actually communicating with a chatbot. This is an issue of developmental psychology. Jean Piaget (1926) first described the ‘animistic’ worldview of young children, who tend to see inanimate objects as not only alive, but also endowed with intentional states. This is unproblematic in itself, as it is appropriate for certain developmental stages.

Nevertheless, certain AI systems, or humanoid robots, may make it more difficult for young children to distinguish between ‘persons’ and non-personal beings or objects. While this should be kept in mind when developing educational chatbots, it makes sense to focus on cases where students are developing some kind of quasi-relationship with a chatbot, without mistaking it for a person. These students are not deluded, but to some extent they live under the illusion that they are communicating with ‘someone’. These kinds of relationships come in various forms and degrees. Some people will show emotional reactions in their conversations with chatbots but will not enter into a ‘relationship’ with them; others may develop a kind of relational dependence and turn to their chatbot whenever they feel lonely or desperate. Some will even fall in love with them. The question is whether educational chatbots should be designed to foster an emotional attachment between learners and the chatbot.

The objection is that this has negative developmental consequences; namely, emotional and social capacities are not developed properly when students interact with a machine as if it were a person. In interpersonal relationships, everyone involved has their own views, emotions and

needs, and makes claims on others. Becoming socially and emotionally competent involves, among other things, understanding what others think, want or need emotionally and cognitively, and addressing their claims appropriately. AI chatbots do not have their own goals or needs, nor do they make any claims. While they may correct users in epistemic terms, they never interfere with their goals and mostly support them in their endeavours.

The problem may be that, when students become accustomed to this kind of interaction, they fail to develop the skills and attitudes necessary for navigating 'real' relationships and communicating with others in situations involving conflicting needs. Ultimately, interacting with chatbots could negatively impact students' ability to form fulfilling relationships with others.

However, there is not yet enough empirical evidence to determine the consequences of quasi-personal relationships with AI systems for human relationships. Much will depend on whether relationships with chatbots replace interpersonal relationships. If children or adolescents would interact mainly with chatbots because it is easier and does not lead to conflict or frustration, this is likely to have negative developmental consequences. However, if they only communicate with chatbots on school-related issues from time to time while maintaining good relationships with their friends, teachers and parents, negative effects are less likely. Assuming that quasi-empathic communication with chatbots can have a positive effect on learning, it is difficult to determine whether this potential benefit outweighs the possible risks of quasi-personal interactions with chatbots.

Rather than discussing the impact of quasi-empathic relationships on personal development, we could simply assert that treating a machine as a person (or as if it were a person) is inadequate. This is the third objection. It presupposes the general view that emotions can, in some sense, be 'adequate' or, to use a term common in current philosophy of emotion, 'fitting'. There are different philosophical accounts of fittingness. In this context, however, it is not necessary to provide a detailed theory of emotions; a general idea of fittingness formulated using some examples suffices. For example, it is fitting to be afraid in the face of real danger, whether it is 'objectively' present or perceived on the basis of relevant evidence. We might also say that, in some circumstances, there is a (normative) reason to be afraid.

Similarly, we could formulate fittingness conditions for envy. We could assume, among other things, that envying someone is only adequate if this person is better off than oneself. It has been suggested, that such considerations of fittingness are different from moral or prudential judgements. In a basic sense, envy can be adequate although it might be imprudent or morally problematic to feel envy towards someone (D'Arms & Jacobson, 2000). So, we might have a normative reason to be envious although we should not have this emotion for moral or prudential reasons.

When it comes to LLM chatbots, the focus of adequacy considerations will be on relational emotions. The main question is whether it is fitting to have certain emotions *towards them*, given that they themselves do not have emotions, thoughts, desires or intentions. Take gratitude as an example: It seems inadequate to be grateful to something that is incapable of having intentions, whether good or bad. The only reason to be grateful to someone is if they intentionally did something good for us. Apart from that, there may be non-relational emotions of gratitude: For instance, we can be grateful for good ideas provided by a chatbot without being grateful to the chatbot itself.

Moreover, since chatbots are not empathic and cannot care for people, it is inappropriate to react emotionally to them as if they cared. When other people sincerely care about our well-being, it is adequate to have a wide range of emotional reactions towards them, from trust and joy to love. By contrast, it is inadequate to have emotional reactions – including empathic care – towards a chatbot. The third objection is therefore that we should not foster an emotional attachment to chatbots, as at least some of the emotions involved would be inadequate.

If we accept the idea that emotions can be inadequate in this sense, the further assumption is that it is beneficial, particularly with regard to personal well-being, to have an adequate emotional response to one's social and natural environment. For instance, it can be detrimental to well-being to constantly be afraid in situations where there is no reason to be afraid. Although it may feel good to be validated by a chatbot, our well-being may still depend on others having 'real' feelings of care and love towards us.

Against this background, developing a quasi-personal relationship with a chatbot is problematic, regardless of its impact on individual development. In my view, this is the main argument against promoting

emotional ties to chatbots, whether through designing or selecting educational chatbots on this basis, or encouraging students to form emotional connections with AI systems.

4 Transparent partnership

So, how should chatbots communicate in educational settings, and how should students relate to them? I propose that we conceive of the relationship with chatbots as a ‘transparent partnership’. This means that chatbots should not be considered mere tools, but rather as counterparts in some sense. However, they should not be designed to communicate quasi-empathically; rather, they should make it clear in their communication that they are software programmes processing information in specific ways.

There are several reasons to see chatbots as partners. Large language models are based on a wide range of cultural information – essentially, the collective knowledge of our societies that is digitally accessible. While this raises copyright issues, it also means that chatbots’ utterances draw on a cultural treasure trove developed by individuals and groups over generations. Chatbots allow human intelligence and creativity to speak to us.

Furthermore, chatbots respond to individual prompts and engage in extensive conversations, reacting to users’ questions and remarks. In some ways, they are more attuned to individuals’ needs than some human counterparts who may not listen and are focused on their own interests.

This also means that, while AI systems do not have intentions or goals of their own, they can support people’s goals and projects if prompted correctly. In this sense, we can see them as tools, as means to our own ends, rather than as partners. However, unlike other tools, interacting with or collaborating with a chatbot might cause us to modify our goals. While common technical objects, such as a typewriter, are used for predetermined human purposes (in this case, writing a text), collaborating with a chatbot may stimulate us to pursue new goals or abandon a project altogether. Since chatbots often provide valuable ideas, it makes sense to take them seriously and consider their suggestions, as we would with human dialogue partners.

In this sense, chatbots can therefore be considered partners. However, they should not be seen or treated as quasi-human counterparts. When

communicating with students, they should not attempt to create the illusion of an empathic relationship; rather, they should be transparent about their inner workings. They should make it clear that they do not understand, feel or care about anything or anyone. It also makes sense to educate students about how various AI systems function, provided that the technical details are understandable to teachers and can be explained to students.

It is clear, then, that certain kinds of relational emotions are inappropriate in this kind of partnership. This does not mean, however, that collaborating with a chatbot should be devoid of emotion. For instance, it is reasonable to feel astonishment when the chatbot suggests an unexpected idea, although it would be inappropriate to express gratitude to the chatbot for it. We can feel joy when the chatbot helps us to understand our own ideas better and sadness when communication with the chatbot reveals that our project is not as promising as we thought. We may also become angry when collaborating with the chatbot, but it is pointless to direct our anger at the chatbot itself. Additionally, it seems reasonable to value chatbots as collaborative partners. While valuing something can have an emotional dimension, it would be inappropriate to show interpersonal appreciation or respect towards objects.

5 Conclusion

AI chatbots used in education should not be designed to communicate in quasi-empathic ways. They should not be programmed to foster an emotional bond with users. The main reason for this is that it would lead to learners having inadequate emotional reactions – having relational emotions in response to a machine is inadequate.

At the same time, as this essay argues, it is appropriate to see chatbots as collaborative partners and to value them as such. The relationship with them should be transparent, meaning the illusion of a quasi-human relationship should be avoided.

Brighouse, H. & Swift A. (2014). *Family values. The ethics of parent-child relationships*. Princeton University Press.

D'Arms, J. & Jacobson, D. (2000). The moralistic fallacy: On the 'appropriateness' of emotions. *Philosophy and Phenomenological Research*, Vol. 61(1), 65–90.

- Duschinsky R. & Collver, J. (2019). “Trust comes from a sense of feeling one’s self understood by another mind”: An interview with Peter Fonagy. *Psychoanalytic Psychology*, 36(3), 224–227.
- Eilan, N., Hoerl Ch., McCormack, T. & Roessler, J. (2005) (eds.)- *Joint attention: communication and other minds*. Clarendon Press.
- Fonagy, P. & Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy*, 51(3), 372–380.
- Hobson, R. P. (2005): What puts jointness into joint attention? In N. Eilan, Ch. Hoerl, T. McCormack & J. Roessler (eds.), *Joint attention: communication and other minds* (pp. 185–204). Clarendon Press.
- Misselhorn, C. (2021). *Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co*. Reclam.
- Piaget, J. (1926). *La représentation du monde chez l'enfant*. Alcan.
- Platz, M. (2021). Trust between teacher and student in academic education at school. *Journal of Philosophy of Education*, 55(4/5), 688–697.
- Simion, M. (2020). Social epistemology of education. In M. Peters (ed.), *Encyclopedia of Educational Philosophy and Theory* (pp. 1–6). Springer.
- Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard University Press.
- Weber-Guskar, E. (2024). *Gefühle der Zukunft. Wie wir mit emotionaler KI unser Leben verändern*. Ullstein.