Humans Relating to Virtual Agents. Conversational AI and VR in Professional Training

A phenomenological account

Klaus Neundlinger

www.ice-research.org

ice research

Katrin
Sent you a message
"Wir sind in Raum 003. Wir warten..."

Presentation for the Seminar 'Philosophy of AI',
held by Prof. Stefania Centrone, STS Department TU München
February 3rd, 2024

1) My perspective as a practitioner: Training of social situations on the workplace. Social skills training by using Virtual Reality and Conversational AI

The perspective I present in my talk is twofold. On the one hand, I am a practitioner in the field of professional education and training. I have been involved in the design and delivery of workshops, courses and programmes that focus on organisational culture and social skills. Our training programmes address social interaction in collaborative settings, such as teams, departments, and cross-departmental projects. In our training settings, we aim to strengthen people's ability to assert themselves in everyday interaction, i.e. to express their interests and goals and to actively participate in problem-solving and decision-making, as well as their ability to take a different perspective, to listen to others and to commit to the team's goals.

Therefore, we focus on collaboration from the perspective of individuals interacting in specific situations, rather than addressing and analysing organizational processes and structures or dealing with the question how these can be changed to optimize collaborative outcomes. Our approach prioritises the experiences of those involved in the collaboration.

Against the background of this practical approach, I have participated in research and development projects focused on exploring the application of interactive technologies, such as VR or AR, in professional training. Virtual environments offer users the opportunity to interact with other human users via avatars or with virtual agents controlled by computers. In certain instances, these virtual agents utilize conversational AI and may also draw on LLMs. In virtual environments, users engage in conversations with an interlocutor who appears as a human being, while communicating with an AI-based technology. A question that may arise in the context of this seminar is whether these virtual agents can be attributed an 'artificial' intentionality that includes also, at least in the experience of the users, a body, movements, gestures, facial and vocal expressivity, emotionality etc.

2) My perspective as a philosopher: intentionality as act, not as mental state or representation

This leads me to the second perspective I want to take. Being a philosopher with a specialization in classical phenomenology, I am used to conceive of the term of *intentionality* not only as a structure of human mind, but also as a process that involves the lived body. It is precisely this conception I want to outline from my experience as co-designer and user of virtually embodied human-machine interactions for training purposes. I am not concerned with the technology behind these applications, but rather with the experience that this technology facilitates. And I believe phenomenology can make a valuable contribution to the interpretation of this type of human-machine interaction when at stake it is the question how our lifeworld changes with the use of these technologies. To be sure, this is to be understood as an epistemological account, and therefore the terms used in the context of this interpretation must be explained and justified as such.

What characterizes the phenomenological understanding of intentionality? A phenomenological approach to any form of objectual consciousness consists in a thorough description of the respective intentional act (be it perception, memory, expectation, imagination, …) which is rooted in time-consciousness and in what Husserl calls 'passive synthesis', i.e. in bodily movements and receptive (pre-predicative) processes that are the pre-condition of rational and abstract thinking ('active synthesis').[1] This the reason for which phenomenology rejects the Cartesian division between *res cogitans* and *res extensa*, between mind and world. As Sartre and Merleau-Ponty put it, consciousness must be understood as being-to-the-world, as an embodied pre-reflexive being among the things and the others, and not the ideal representation of reality. Gallagher and Zahavi show in their introduction to phenomenology and its relation to cognitive sciences that these positions are

[1] Husserl, E. (2001) *Analyses Concerning Passive and Active Synthesis*. Dordrecht: Springer.

neither obsolete nor have they been refuted by cognitive sciences or other philosophical positions that use the term ‘intentionality’.[2] To repeat and stress it, the crucial difference between the phenomenological conception of intentionality and other accounts is that in phenomenology, the constitution of any objectual appearance is related to an act, be it passively experienced or actively accomplished, by a subject. Intentionality is not a mental state or any form of representation that would contain an object. Rather, the intentional act *constitutes* its respective object. As Sartre[3] puts it, a table is not *in* the consciousness, not even as a *representation*, but *in* the room. – And this has nothing to do with naïve realism, since he shows how the perception of the object is linked to the existence of a pre-reflexive ego for which it is constituted as being there in the room, manifesting itself from a determinate perspective and pointing to other perspectives to be taken in form of a virtual infinity of further perceptive acts.

Therefore, my theoretical interest in the immersive technology described above, in its combination with conversational AI, differs from perspectives in the tradition of logical and analytical philosophy, but also, to a certain amount, from that of Hubert Dreyfus. I am not concerned with the question what computers can’t do or whether machines are able to pass the Turing test. Rather, I am interested in describing how the lifeworld we experience, and take part in, changes as we become more accustomed to ‘natural’ forms of dialogue during certain types of human-machine interaction. In phenomenological terms, my question is what type of intentionality constitutes these virtual worlds in which human-computer interaction appears as more and more similar to human dialogue.

Methodologically, we must accomplish another important step. In the course of a phenomenological analysis, any reference to psychological, physical, chemical or neurobiological knowledge that could explain what happens in our brain and body when we accomplish the aforementioned acts (perceiving, remembering, expecting, communicating, …) is suspended by the switch from the natural attitude to the phenomenological attitude which Husserl calls *Epoché*.[4] What we are expected to do when we engage in the analysis of a specific type of intentionality, like that of perceiving, remembering, communicating and so on, is to establish a thorough description of how these acts constitute their respective object in the experience, an experience that cannot be reduced to general psychological concepts or models like sense data etc. What we gain by these steps, and by accomplishing the description, are the specific qualities of the respective acts, not a general scientific explanation of their physical or biological foundations. This also means that, by carrying out a phenomenological description of the respective act, we do not refer to, neither aim to construct, a generalized model of intentionality that could serve as a basis for comparing human with artificial intelligence. What we aim to accomplish, in a first instance, is to better understand the type of relationship that constitutes human-machine interaction experienced as if it was an interaction between humans.

3) The use case: Training social skills with a virtual agent

In order to concretize these theoretical considerations, I would like to refer to a research project I coordinated.[5] Its aim was the development of a prototype for a Virtual Reality (VR) based social skills training unit. The storyline for the VR scene was developed in collaboration with mid-level managers in an international corporation based in Austria. In the scene, the player takes over the role of a team manager sitting in their office. A virtual agent named Mira Horvath, who is introduced as a collaborator of the player’s fictitious team, steps in, expecting her superior (the player) to start a

[2] Gallagher, S., Zahavi, D. (2012) *The Phenomenological Mind*. London and New York: Routledge.

[3] Sartre, J.-P. (1972) *Being and Nothingness. An Essay in Phenomenological Ontology*. London: Routledge.

[4] Husserl, E. (1960) *Cartesian Meditations. An Introduction to Phenomenology*. Dordrecht: Springer.

[5] Neundlinger, K., Frankus, E., Häufler, I., Kriglstein, S., Schrank, B. (2023) *Virtual Skills Lab – Transdisziplinäres Forschen zur Entwicklung sozialer Kompetenzen im digitalen Wandel*. Bielefeld: transcript.

scheduled meeting with her. She wants to present her team leader a template she has been working on. Suddenly, a message pops up, reminding the team leader to attend another meeting of higher priority. The task for the player is now to “say no in an appreciative way”, i.e., to postpone the meeting with Mira again.

For a number of reasons, the focus of this project was to test how the immersive nature of the virtual environment contributes to an emotionally engaging dialogue for training purposes. We did not aim to facilitate an open-ended conversation that could take unpredictable directions, although, based on conversational AI, this could have been a feasible solution. Actually, there are already solutions available that go in this direction. Instead, our aim was to confront the learner in any case with the unpleasant reaction of the virtual agent, representing the employee, when she realises that her team leader is sending her away without having paid attention to what she considers a result, i.e. an achievement. Actually, the training scene inevitably ends with the virtual agent leaving the office, showing clearly how disappointed she feels by her manager's behaviour.

What is being negotiated in this scene in the form of a simulation is not limited to those aspects that can be expected in the context of an interaction in everyday office life: Goals are to be achieved, and in order to achieve these goals, tasks are distributed. The organization of such processes, when it concerns the work of people, is called human resources management. Yet, what the learner's attention is drawn to in this scene is not the question of whether the employee has achieved the goals that were set for her or that she herself derived from her tasks. She has obviously been successful in this, because otherwise she would not have come to present results. Performance appraisal is not the issue here, in terms of reflecting and expressing criteria regarding the evaluation of the work accomplished. The topic of the training unit is to be found in a different dimension of the relationship between a team leader and their collaborators. Although it can be said that the employee appears to the manager from the point of view of human resources management as an objective (maybe objectified) force, a potential that has goals to fulfil in the form of accomplished tasks and deadline, at the same time she appears to him from a different point of view, namely that of social recognition. When playing the simulation with the virtual agent, learners are aware that they are pretending to interact with a human person they already know. Comparably to team leaders in real life, they are immersed, via the virtual environment, in a simulated social situation where they are asked to play a role. Thus, not only cognitively, but also emotionally, they engage in a conversation that seems to be embedded in a personal relationship. In this sense, what is at stake in the interaction between the player and the virtual agent are not general or institutional forms of recognition, but a specific one that unfolds in a concrete relationship, between a team leader and his (or her) collaborator.

4) Training social skills with AI: the role of intentionality

Let's first look at the scene from the perspective of the intentional acts we can describe phenomenologically. By putting on the virtual reality headset, I am sitting at a desk in a bright office space. The atmosphere is pleasant, my office is connected to a meeting room by a glass front. An insert introduces me to the game. I am informed that I am a manager and will shortly be having a meeting with my colleague Mira Horvath. By slipping into the role, I can think about how a manager feels under the given circumstances and what he or she is probably thinking. This cognitive and affective change of perspective gives me access to a repertoire of possible thoughts and feelings. Am I a more fact-oriented manager who makes sure that my employees perform well so that I expect my collaborator to come to the point immediately? Or do I cultivate an empathetic management style and look forward to my employee entering my office in a minute so that I can exchange a few personal words with her?

As mentioned, such considerations, which are relevant for training social skills, do not only concern the cognitive side of changing perspectives. The immersive nature of VR technology also encourages me to empathise with the situation. Similar to a role-playing game, the simulation awakens feelings and memories or evaluations of social interactions that connect this learning experience with our wealth of experience. An intentionality is built up that appeals to deeper layers of emotional experience, our desires and expectations. All of this is possible because we can associate this situation with other situations we have experienced or modify it in the sense of varying it in our imagination.

The different management styles we can try out in this kind of variation before the interaction starts have an impact on how the conversation with the employee might go. Even though there was no possibility to try out these different styles when communicating with the virtual agent in the VR training scene we created, this variation in behaviour could be a didactic goal in a further developed simulation. It would be possible to confront the learners with different courses of the conversation, depending on whether they themselves communicate more soberly or empathically. Furthermore, and this is where conversational AI comes into play, the virtual agent could communicate in different ways, sometimes rather shy and understanding, other times disappointed or aggressive.

In light of the significant advancements in AI and machine learning, a question that may arise is whether this type of training will replace other forms, such as roleplay, and if AI will assume the role of a trainer. This question has didactic, economic, and ethical implications. However, it is also linked to the question of whether intentionality is a concept that can contribute to a better understanding of artificial intelligence. If it is possible for humans to conduct increasingly free, complex dialogues with machines, is it justified to speak of intentionality when referring to the output? Don't ChatGPT and other models already have all the capabilities that we attribute to ourselves as intentional beings? Could 'wordware' be the next step in the evolution of interactive technologies, which would allow for most human-computer interactions to be carried out through speech? If we can enter into a dialogue with the machines not only in writing but also via voice and virtual bodies, if we interact with these machines not only rationally but also emotionally, are we then not dealing with an interaction partner to whom we must at least attribute characteristics that we have so far reserved for the human mind?

Let's stay with the type of interaction that we wanted to train with our VR application: how a manager deals with their employee in a stressful situation. The question of what appropriate behaviour on the part of the manager should look like cannot be reduced to the use of appropriate phrases, and a meaningful training measure should consider this. Ultimately, the encounter between manager and employee evolves as a personal relationship that includes all the dimensions and aspects that characterise such a relationship. It unfolds as a shared history and includes mutual familiarisation and understanding, but also conflicts and misunderstandings. It is a relationship that is embedded in a larger network of relationships and is not unaffected by them. In this sense, all intentional acts that are necessary to achieve the team's goals constitute an intersubjective horizon: the planning and distribution of tasks, their completion, feedback, regular coordination, motivation, clarifying, conflict, the evolving and changing team atmosphere and so on.

In this sense, the understanding of intentionality as it characterises the development of personal relationships, for example at work, differs from what we can define as the achievements of artificial intelligence, which can nevertheless be used to train individual aspects of managing relationships, such as in our case the way managers deal with employees in a stressful situation. Even if there are now applications that are used by humans to simulate affective relationships, we cannot or should not speak, in these cases, of the intersubjective intentionality that is typical of concrete relationships that take place between living people. What characterises this intentionality is that it unfolds as a concrete common history, by joint acts like the physical encounter, but also the mutual concern, the thinking and with and caring for the other, the appreciation and conflict, the expectations and memories, activities, and successes. We can only have relationships with virtual characters because we have been living real relationships and project our real-life experiences and desires onto them. What makes these simulated relationships convincing is the fact that also relationships in real life are based on projections and desires, but nevertheless in these cases they refer to a concrete other with his or her own history, desires, and projections.

Can we clearly distinguish virtual relationships from real ones? In my view, this question contains a flawed conceptual opposition. It is undeniable that we now live in a world in which virtuality, in the form of communication technologies, has become part of any kind of relationship. However, virtuality in a broad and non-technological sense has always been part of living relationships, in the form of letters, memories and expectations. What counts is the distinction between real persons and virtual agents, and therefore it is important to work out a clear understanding of the intentionality that links us to technologies such as LLMs, as opposed to the type of intentionality that unfolds in interpersonal relationships.

5) Common ground? Culture, language, meaning

When we look at large language models from a lifeworld perspective, we are amazed at how these models manage to communicate with us in such a differentiated way. One question that fades into the background amidst all this astonishment is: How do people create the highly complex forms of interaction that have been characterising their daily interactions ever since? From my point of view, this question, if pursued, is a clue as to how we can better appreciate the interactions that have become possible through the advancements in the use of Large Language Models. These interactions already characterise our everyday lives and will do so even more in the future.

Phenomenology denies itself recourse to theoretical models when it asks itself how we understand others, how we can empathise with other people or even cognitively put ourselves in their shoes. It does not fall back on a theory of mind and does not claim that we simulate how others think and feel in order to understand them. Methodologically speaking, this means that it must achieve a point of view of generality, which is necessary in order to move from me to the other, from my point of view to that of the other, via the method of description. The generality of meaning, that makes communication possible, must therefore manifest itself in the concrete dialogue, and yet on the other hand it must go beyond this. From a phenomenological point of view, how can we distinguish between living speech and the system of meaning to which this speech is always related? More generally, how do I describe the general level of sense that enables the production and reception of concrete meaning, for example when writing and reading a text? If we only have access to our consciousness, how do we reach the consciousness of the other?

In the chapter entitled ‘Dialogue and the Perception of the other’ of his book *The Prose of the World*, Merleau-Ponty writes: „Rationality, or the agreement of minds, does not require that we all reach the same idea by the same road, or that significations be enclosed in definitions. It requires only that every experience contain points of catch for all other ideas and that ‘Ideas’ have a configuration. This double requirement is the postulation of a world. However, it is not a question here of the unity attested by the universality of feeling, since the unity of which we are speaking is invoked rather than verified, and since it is almost invisible and constructed on the edifice of our signs. Thus we call this universality the ‘cultural world’, and we call speech our power of making use of certain conveniently organized things—black and white, the sound of the voice, movements of the hand—to put in relief, to differentiate, to master, to treasure the significations which trail on the horizon of the sensible world, …“[6]

[6] Merleau-Ponty, M. (1973) *The Prose oft he World*. Evanston: Northwestern University Press, p.143.

People understand each other or make themselves understood by drawing on a repertoire of behaviours and actions into which they are introduced from birth. When speaking, and later when writing, but also when listening and reading, they learn to move in this cultural world, which for them is never separated from sensory experience, but also contains those ‘almost invisible’ meanings that allow them to reach the level of generality necessary for them to understand each other. Generality is reached by this continuous process, it is constituted within the continuous exchange of (more or less) meaningful gestures, phrases, behaviours, and actions. Meanings are interconnected and can be reconfigured in the search for meaning, in the process of understanding. The fascinating thing about the technologies of language (and image generation) is the fact that these technologies have access to our cultural world in a way that was not previously the case. By producing text, finding answers and performing and reacting to speech acts as if they were human, they take actively part in the game of reconfiguring meanings. From a phenomenological point of view (to emphasise once again, this perspective must put into brackets the mathematical or technological mode of explanation and also any form of modelling consciousness and understanding, be it based on neurology, systems theory etc.), LLMs and image generators such as Midjourney and Dall-E achieve this because a level of generality without which meaningful expression and understanding of meaning are not possible is already inherent in living communication between people. On the other hand, for humans meaning that is completely decoupled from the sensory world and bodily processes would also be inconceivable. When we speak and understand, we not only use language as an instrument for conveying our message (that would be a model of communication), but we are also part of the dynamics of meaning production and reception, we are ourselves the language that we speak in that we are embodied beings whenever we express ourselves, and as embodied beings in time we participate in communication even when we are silent.

"Between myself as speech and the other as speech, or more generally myself as expression and the other as expression, there is no longer that alternation which makes a rivalry of the relation between minds. I am not active only when speaking; rather, I precede my thought in the listener. I am not passive while I am listening; rather, I speak according to . . . what the other is saying. Speaking is not just my own initiative, listening is not submitting to the initiative of the other, because as speaking subjects we are continuing, we are resuming a common effort more ancient than we, upon which we are grafted to one another and which is the manifestation, the growth, of truth."[7]

From a phenomenological point of view, it is therefore the "almost invisible" dimension of meaning that has always made interpersonal communication possible and that now allows interactive language technologies to enter into a dialogue with us and to develop, i.e. learn, through this communication. As far as the concept of intentionality is concerned, however, we must differentiate between the actions and acts that are built up in a concrete relationship or in a network of interhuman relationships (team, community, family, ...) and the intentionality that is formed through interaction with technologies based on LLMs. How can we conceptualise such an intentionality that emerges in human-machine interaction and allows machines to participate in the cultural world as never before?

6) Intersubjective intentionality with ChatGPT? Lifeworld considerations

This question is too broad to be answered satisfactorily in this short presentation. I would therefore like to return to the empirical use case of social skills training to briefly outline how I imagine an intersubjective intentionality that relates to LLMs in the realm of everyday office communication. Our use case is a good illustration of the extent to which human and machine learning processes are now intertwined. To the extent that we use machine learning to train interpersonal skills, the algorithm receives training data on how humans interact with each other on this level. One problem of this

[7] Merleau-Ponty, *The Prose of the World*, p. 143/144.

interwovenness of human-human and human-machine interaction is the risk of a truncated or stereotypical view of communication resulting from this process. AI-supported training could be used to standardise behaviour without there being a transparent reflection on the normative presuppositions and effects of such a training programme. As language technologies, to give a concrete example, are already being used to design communication in terms of tone and wording, we must assume that collective communicative behaviour will change significantly by way of this continuous human-machine interaction, especially in the professional sphere. Chat programmes will more and more take over a filter function that comes between me and others.

To be sure, people have been thinking about how to formulate a phrase in a meeting, a letter or an e-mail also in the past and developed different ways of communicating in order to meet the specific habitus of their communication partners. Adapting to communicative situations is therefore nothing new. The question is how this behaviour changes on a collective level when we massively outsource it to technology. In any case, LLMs make it possible to quickly access the entire repertoire of behaviours that the 'cultural world' makes available in the form of texts. Therefore, we should reflect on what a 'conscious' or 'critical' use of these technologies could look like, e.g. in opposition to behaviour-related streamlined communication techniques imposed by the management of organizations.

Let's turn back to the theoretical implications of this argument: A phenomenological perspective on language as living speech makes it possible to capture those aspects of communication—and, to a certain extent, to protect it from levelling—which not only mediate between concrete communicative situations and general levels of meaning, but also harbour something new and innovative.

"The living relation between speaking subjects is masked because one always adopts, as the model of speech, the *statement* or the *indicative*. One does so because one believes that, apart from statements, there remain only stammering and foolishness. Thus one overlooks how the tacit, un-formulated, and nonthematized enters into science, contributing to the determination of science's meaning, and as such provide tomorrow's science with its field of investigation."[8]

In other words, in the context of AI-supported experiential learning on the topic of social skills, for example, it is not just about the best way for a manager to express themselves in a given situation, or about the tone and gesture with which they speak. All of this can now be transformed into data using speech or facial recognition and used for automatised training. What is overlooked, however, is that an encounter with another person can only be described as such if it includes precisely those tacit, un-formulated and non-thematised aspects of speaking to one another that make up a living dialog. In my view, social skills training for the professional field should teach two things:

- on the one hand, awareness of the fact that communication is increasingly informed by AI-based applications, i.e. the constant presence of technologically processed knowledge about communication that draws on the 'cultural world', i.e. on collective knowledge.
- on the other hand, the ability to engage with other people in an open encounter, i.e. to build up and maintain human relationships, even if they are more and more mediated by communication technologies that include artificial intelligence, such as conversational AI.

These two aspects can come into conflict with each other, although this is not due to the technology as such, but to the way in which the intersubjective intentionality in which AI participates is conceived and designed. In conclusion, let me refer to an episode that illustrates how the use of reductive models can lead designers to misinterpret the needs of learners and the embeddedness of human interaction in a determinate social context. In an early stage of our research project, we conducted a series of interviews with experts engaged in the development and application of the VR technology in professional training. One of our interviewees recalled an episode that occurred in the

[8] Merleau-Ponty, *The Prose of the World*, p. 144.

context of one of their research projects. In order to help school dropouts prepare for job interviews, the researchers had developed an application within which the youths interacted with a virtual agent appearing on a monitor. At one point, the agent asked them to talk first about their personal strengths and then about their weaknesses. One young man reacted to the agent inviting him to assess his weaknesses in the following way: "He froze, he didn't do anything—[this is a] mega disaster for a computer scientist, [when] you can't measure anything—, [then he] grabbed the monitor and threw it out of the window. And we thought: […], we cannot do this in this way, we need to give some serious thought to what happens inside people's minds."

Elaborating further on the issue, this researcher explicitly referred to 'Theory of Mind'. As he told us, in later projects, cognitive-science-based models would be used to design dialogues with machines that exhibit empathy towards learners. However, by taking this theoretical view for the design of training simulation, the researchers seemed to disregard the obvious reason for the boy's rage during the interview simulation. As an unemployed school dropout, this young man was probably often confronted with the fact that someone was accusing him of personal failure. He could therefore only perceive this question as a criticism of himself and not as an invitation to show how aware he was about his personal skills or character and that he was ready to learn from mistakes and able to handle also negative feedback. On the contrary, the question triggered a feeling of powerlessness and frustration in him. As sensible as it may seem to make chatbot technologies accessible to young people without perspective for training purposes, it is also important to take the context into account and adapt the didactic design accordingly.

To do this, however, it is necessary to look at the way in which young people *experience* their world from a different perspective than that of the Theory of Mind. It seems much more important to include the lifeworld situation into the design and to ask how technologies can be used in an obviously difficult social environment in such a way that they are not perceived as a threat, but rather as an opportunity to learn and develop relevant capacities for mastering social interaction. For this purpose, one can draw on the tradition of symbolic interactionism (Mead, Schütz, Goffman, Berger/Luckmann) with its differentiated view on frames and constructions that create a situation, but also to more recent developments that try to link phenomenology and cognitive sciences, such as the 4E cognition studies (embodied, enacted, embedded, extended).[9] In this context, qualitative methods like the 'phenomenological interview' have been developed, in order to grasp the dimension of meaning form the perspective of the lived experience of the persons involved.

[9] Newen, A., De Bruin, L., Gallagher, S. (2018) *The Handbook of 4E Cognition*. Oxford: Oxford University Press.