3

Digital Simulations of Emotions

A young blond man with freckles steps out of a helicopter onto a beautiful island. Lush vegetation, gentle streams, cascading water. After walking through a dense forest for a while, he finally arrives at a hyper-modern house equipped with maximum-security systems. The house (and the whole island in fact) belong to Nathan. He is the inventor and founder of the largest Internet search engine in the world called "Blue Book" (an allusion to the famous book by philosopher Ludwig Wittgenstein). Nathan is an ingenious and arrogant programmer who has set himself the goal of creating a new species: a robot capable of consciousness. Caleb, the young man with freckles, works in Nathan's company and was chosen by Nathan to test whether one of his first robots has reached this goal.

"Do you know what the Turing Test is?" Nathan asks him shortly after his arrival.

"Yeah," Caleb replies. "I know what the Turing Test is. It's when a human interacts with a computer. And if the human doesn't know they're interacting with a computer, the test is passed."

"And what does a pass tell us?"

"That the computer has artificial intelligence."

The robot Caleb is supposed to test is Ava, an attractive robot woman. Her face resembles that of a young woman. Only her legs and arms are made of shiny metal, and blue wires glow in her belly. When she moves, there is a soft hissing sound, as if a neon tube is humming. In various sessions, Caleb watches Ava through a pane of bulletproof glass. Through the PA system, Caleb talks to her, asks her questions, tests her. Like an enigmatic sphinx, she

J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism*,
https://doi.org/10.1007/978-3-031-12482-2_3

sits across from him and answers all his questions—like a real, self-aware human being. After a while, however, Ava begins to turn things around. Now it is she who starts asking Caleb questions. Looking at her face, Caleb can detect many emotions in her. She is surprised, sometimes flattered, sometimes puzzled, sometimes hurt, and finally in love. And yet, Ava is a machine. So how can she possibly have emotions?

Nathan will soon give Caleb the explanation:

"If you knew the trouble getting AI to read and duplicate facial expressions. You know how I cracked it?"

"I don't know how you did any of this," Caleb replies perplexed.

"Every cell phone just about has a microphone, camera, and a means to transmit data. So I turned on all every microphone and camera across the entire f***ing planet, and I redirected the data through Blue Book. Boom. A limitless resource of vocal and facial interaction."

Ava is an expert in facial and vocal expressions. By observing all the people in the world and their reactions, she has acquired a perfect reservoir of knowledge about facial expressions over time. She knows how to interpret facial expressions and she knows what facial expressions are considered appropriate at what time. Big Data makes her a perfect imitator of emotional expressions. But does that mean she really *has* feelings?

"I want to be with you. [...] Do you want to be with me?" Ava asks Caleb in the fifth session.

Caleb, too, would like to know if Ava really has feelings for him or has just been programmed to pretend to do so. Eventually, Caleb decides to believe her. He regards her as an autonomous and unique being. A being he falls in love with and assumes has fallen in love with him as well.

In another session, Caleb tells Ava about the thought experiment "Mary's Room". This thought experiment really exists. It was put forward by the Australian philosopher Frank Cameron Jackson in his essay "What Mary didn't know" (1986).

"Mary is a scientist, and her specialist subject is color. She knows everything there is to know about it, the wavelengths, the neurological effects, every possible property color can have. But she lives in a black and white room. She was born there and raised there and she can observe the outside world on a black and white monitor. One day, someone opens the door, and Mary walks out. And she sees a blue sky. And at that moment, she learns something that all her studies couldn't tell her. She learns what it feels like to see color."

Ava looks at Caleb motionless. Judging by the expression on her face, this story is taking a toll on Ava. This isn't surprising. After all, isn't she just like

Mary? A person who knows everything but only from second-hand information from the Internet? In Ava's face, Caleb reads disappointment, but also a fierce determination. She makes it clear to Caleb that she too wants to leave her room one day. Preferably—so she tells him—with him. On their first date, she tells him about her biggest dream: Standing at a busy intersection, watching the people go by.

When she finds out that Nathan plans to switch her off soon to recycle parts of her for a new robot, she is determined to do everything she can to escape. Caleb wants to help her and comes up with a plan.

By the end of the film, Caleb has managed to break the code of the maximum-security wing. Ava escapes. Shortly after, Ava kills Nathan, her creator. Nothing stands between her and her freedom anymore. But then something happens that neither Caleb nor the viewer expected at this point: Ava cold-heartedly leaves Caleb behind, locked up in a room. The viewer is also shocked at this moment, because like Caleb he has gotten the feeling in the course of the film that Ava is a sentient being who not only suffers from her situation but has also fallen in love with Caleb.

As Caleb desperately pounds on the door which will keep him inside the house until he'll starve to death, she walks through the house in a white dress and white shoes like an elf. With organic material taken from other deactivated robots, she now walks out into the world. Her brown shoulder length hair caresses her delicate face. As she breathes in the air of the forest for the first time, she smiles. She touches branches and curiously looks at her new life. She feels no remorse and does not even look back.

Like Mary, she now steps out of her room into the big wide world, ready to have her own experiences. Will she learn to not only imitate emotions but also to have them? Or will she remain a machine forever? This is the essence of all philosophical questions around which AI enthusiasts keep circling.

Caleb, too, keeps asking himself the question: Has Ava only learned to imitate certain behaviors in order to give the false impression that she has feelings much like the "cold" actor described by Diderot, whose art focuses primarily on the perfect mastery of physical behavior? The truly troubling question, however, is the following: What if not only Ava's but also our feelings were really nothing more than just pure behavior? That, at least, is what radical positivists claim, advocating the metaphysical thesis that mental states are nothing but patterns of behavior. A positivist's understanding of consciousness identifies mental properties and states, such as being afraid or having desires or beliefs, with particular behaviors. "Jacob is in pain" means—in the positivist's understanding—nothing other than "Jacob behaves in a

certain way, for example, he cries 'ouch' or jerkily withdraws his hand from the stovetop."

It is not a coincidence, by the way, that the film refers to the philosopher Ludwig Wittgenstein several times (once with the name "Blue Book," which is both the name of Nathan's company and the title of Ludwig Wittgenstein's famous book, and another time with the portrait of Gustav Klimt by Margarethe Stonborough-Wittgenstein, Ludwig Wittgenstein's sister, which hangs in Nathan's house) since Ludwig Wittgenstein is considered by most scholars to be a "Behaviorist."

If behaviorism were true, however, we would have to assume that SIRI, the communication software established on many smartphones, has very similar feelings to ours. After all, it reacts as if it were really disappointed or worried. But the software only simulates feelings, it does not have them.

Far more plausible than the behaviorist view on mental states is the realist view: pain characterizes a certain type of feelings that are unpleasant and that we usually try to avoid. At the dentist, we strive to suppress every reaction so as not to disrupt the treatment, but this does not mean that we do not feel pain. Even the imaginary super spartan who shows no emotion even in extreme pain can still actually be in pain. It is simply absurd to identify "having pain" with certain patterns of behavior.[1]

Perhaps the most fundamental argument against the identity of mental states or properties and neurophysiological or digital states or properties is called the "qualia argument." In his famous essay "What is it like to be a bat?" (1974), Thomas Nagel argues that it is not possible to know what it feels like to be a bat (i.e., what the bat feels), even if one examines its brain in great detail. These so-called *qualitative* mental states of the bat are not ascertainable based solely on knowledge of neurophysiological states. So, the qualia argument speaks against the identity of neurophysiological and mental states.[2]

Caleb believes that Ava is in the same situation as Mary from Jackson's thought experiment. She knows—as Nathan told him—everything about the world as well as about people and their feelings, but that doesn't mean she understands what it means to experience the world and to have feelings.

Of course, one can also reject the identity of the mental and the neurophysiological, but still argue that the mental can only occur in connection with the material. Indeed, there is much to suggest that human consciousness is only possible due to the corresponding brain functions. But even those who

[1] Of course, our human ability to mutually ascribe correct mental states to each other depends on there being common patterns of behavior and people expressing their emotional states in similar ways. We can only learn what other people's feelings are because we share certain response patterns.

[2] Chalmers (2010).

hold that human consciousness is based essentially on neurophysiological processes need not subscribe to the identity theory of the mental and the physical. That mental states of humans are *realized* by brain states (i.e., neurophysiological processes and states) does not mean that they are caused by them.

It is undeniable for us humans that we have mental properties, that we have certain mental states, that we have beliefs, desires, intentions, fears, expectations, etc. We are convinced (at least most of us are) that these mental phenomena are realized by processes in our brain, or at least correlate with them. The first-person perspective plays a crucial role in this. However, this must not be radicalized into a solipsistic view according to which I am alone in the world and my mind is the only one that exists. The comprehension of the lifeworld happens essentially through our interaction and cooperation with others to whom we ascribe comparable mental properties. For young, pre-linguistic children, it is not only the haptic experiences of the world, the sensory perceptions, that are important but also the exchange, interaction, and communication with other, older, linguistically capable members of the human species. This role of the Other is not possible without a (presumably genetically anchored) perception of other minds, even in pre-linguistic children. This is how the human conception of the world begins; to doubt that basis would cause our world to collapse.[3] Just as there can be no reasonable doubt for us about other minds, so, as things stand, there can be no doubt about the non-psychic character of the digital. To deny the correlation of the mental and the physical in humans and highly evolved mammals, which bear a sufficient resemblance to us and permit at least a rudimentary recognition of their mental states, is not justified as it mentalizes digital states and processes. Digital states and processes *simulate* mental ones but are not identical to them, even if that simulation were perfect. There is nothing to suggest that mental states and processes can be realized by digital ones. Simulation must not be confused with realization.

In the final scene of the film *Ex Machina*, we see Ava walking through the forest, visibly unmoved. By acquiring her freedom, she has achieved her goal. That however does not prove that Ava has consciousness. After all, as Nathan himself says at some point in the film, she was programmed to want freedom. Seen from that point of view, she was merely acting out her program. Even if the film itself at times suggests that Ava does have feelings, we opt for another interpretation and take the fact that killing two people (Nathan and Caleb) apparently poses no moral problem whatsoever for her as a proof that Ava has no consciousness and therefore no emotions. It was Caleb's fatal mistake to

[3] Nida-Rümelin (2010).

believe her facial expressions and gestures to be expressions of genuine emotions. In this sense, we want to read the film as a warning not to fall into the same trap Caleb fell into when he projected so much more onto Ava than she actually had. We therefore interpret the following utterance by Nathan "One day AIs will look back on us the same way we look at fossil skeletons on the plains of Africa. An upright ape living in dust with crude language and tools, all set for extinction" not as a realistic prophecy but as an expression of masochistic fantasies about the extinction of Western civilization.