3

Robots, Humans, and Their Vulnerabilities

Beate Roessler

3.1 WHERE DO WE HUMANS GO?

In digital societies, systems are becoming ever more powerful, and algorithms ever more complex, efficient, and capable of learning. More and more human activities are being taken over by computers, robots, and AI, and the technologies are becoming ever more deeply and far-reaching integrated into our social practices. It has become impossible to see and understand people, relationships, and social structures independently of these technologies. Especially over the last 2 years, we read almost every day in the newspapers, including and especially the serious ones, that AI leads to the elimination of humans; that the point where AI is more intelligent than we are is approaching. That this has immediate consequences for human life is evident, but it is not just about individual aspects of human life. In a recent article, Acquisti et al. summarize their argument as follows: "Technologies, interfaces, and market forces can all influence human behavior. But probably, and hopefully, they cannot *alter human nature*" (Acquisti et al. 2021, 202, emphasis mine; see, for the following, Roessler 2021a, 2021b).

What I am interested in here is how we should spell out this claim: what does it mean that we hope technologies do not change our human nature and what would this human nature be? Or, put differently, what would it mean to change human nature through technologies and why would it be bad to do so? There has been quite some discussion of this or similar problems in the literature and the most helpful and intriguing is, to my mind, Frischmann and Selinger's (2018) *Re-Inventing Humanity*. Selinger and Frischmann write in an article in *The Guardian* newspaper (Selinger and Frischmann 2015, emphasis mine): "Alan Turing wondered if machines could be human-like, and recently that topic's been getting a lot of attention. But perhaps a more important question is a reverse Turing test: can *humans become machine-like* and pervasively programmable?" This latter question is the topic of their book. Additionally, in the introduction to their book, they write:

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

> As we collectively race down the path toward smart techno-social systems that efficiently govern more and more of our lives, we run the risk of *losing ourselves* along the way. We risk becoming increasingly predictable, and, worse, programmable, like mere cogs in a machine. (Frischmann and Selinger 2018, 1, emphasis mine)

To quote one last passage, this time by Pasquale: "The future that [the robot] Adam imagines . . . reduces the question of human perfectibility to one of transparency and predictability. But disputes and reflection on how to lead life well are part of *the essence of being human*" (Pasquale 2020, 209, emphasis mine).

In this picture, we have Turing on the one hand, trying to build a computer which could be mistaken to be human: we need to work on our technological counterpart to make it as good as human. On the other hand, we have Frischmann, Selinger, and Pasquale, who show us that people – humans – are becoming more and more similar to machines: they contend that we are working on ourselves in order to be ever more perfectly technologically human. In short, we try to improve humans technically so that they become similar to robots; and we try to make robots that become indistinguishably similar to a certain image of the perfect human. Both sides assume – intuitively plausibly – that we know what a 'human being' is and where at least roughly the limits lie between genuinely being human and technologies.

While there is no uncontested concept of *the* human nature, it does seem plausible to argue that human nature is neither something purely accidental, historically completely variable, and relative. It is possible to distinguish characteristics which express what is meant by being human, even though these expressions differ historically, and culturally. Such a concept or idea of human nature could give us critical guidance for analyzing digital societies without risking calling "human" whatever humans (learn to) do under digitally changing conditions. This concept of a human nature can clearly not be reduced to its biological essence: if that was the case, we would not have this discussion in the first place. The question is what it means to have this very special sort of (biological) human nature and how we would best analyze it.

To engage with this rather complex question adequately, I suggest approaching it through a novel whose very topic is the relation between humans and machines: Ian McEwan's (2019) novel *Machines Like Me*. I want to illustrate the problematic by taking up this different perspective on the technological world because, in this novel, McEwan describes the relationship between a human being and an almost-human being; an extraordinarily well-constructed, sensitive, and intelligent robot whose name is Adam. My hope is that, by reading and interpreting *Machines like Me*, we can learn something about how we should think about human beings. Incidentally, it is also possible to interpret other novels, for example *Klara and the Sun* by Kazuo Ishiguro (2021), or, to go a little further back, Mary Shelley's *Frankenstein* (1831), or *War of the Worlds*

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

(1895–1897, by H. G. Wells), but my question would remain the same: what does the image of the robot, of the monster, of the Aliens tell us about the idea and characterization of the human being?

These characteristics by McEwan, I suggest, are generalizable, and I will show, during this chapter, that they can help us to understand the meaning of being human, especially in its relation to the technological world. Furthermore, I will very briefly criticize attempts to transcend this notion of a human being as a finite and vulnerable being and also attempts to imitate and replace "soft" human characteristics, such as emotions and affects in robots, through technology (HRI, Human Robot Interaction technologies). Placing the human being in relation to robots and discussing the extent to which social robots can replace humans helps to understand, or so I will argue, which beings we want to and do refer to as human. I will also argue that it is helpful to refer to the phenomenon of the Uncanny Valley to make sense of a clear line of demarcation between robots and humans – this will in any case be my argument in the end.

3.2 IAN MCEWAN ON ROBOTS AND HUMANS

Ian McEwan's (2019) novel *Machines Like Me* is set in a rather different, alternative 1982: the war against the Falklands has been lost, the miners' strike is still on, unemployment is rising by the day, John Lennon as well as John F. Kennedy are still alive – and, above all, so is Alan Turing.[1] Turing has been working successfully on AI and the construction of a robot, and the first set of these robots is on sale: 12 Adams and 13 Eves as they have been subtly called. The protagonist, Charles "Charlie" Friend, spends the little inheritance he received after the death of his mother on buying one of them and, since he is too late for an Eve, he gets an Adam. The plot of the novel has different threads: there is the relationship with Miranda, Charlie's upstairs neighbour who he fell in love with long ago and who he starts dating. Miranda, after some time, has an affair with Adam; furthermore, she herself has a difficult personal history which she lies about and which is only revealed little by little, leading to the terrible unfolding. This thread in the complicated plot is important because it forces Miranda and Charles to lie – and after Adam has found out about this piece in Miranda's past, he intends to inform the police, since, as a robot, he can't lie. He must be, he wants to be relentlessly upright. Therefore, Charlie kills Adam. Also, rather uncannily, in the last third of the novel an increasing number of suicides by some of the Adams and Eves is being reported. But the main plot is simple: Charlie buys Adam, programs him together with Miranda, develops a rather friendly relationship with him and in the end kills him.

[1] In *Machines Like Me*, Alan Turing is referred to as: "Sir Alan Turing, war hero and presiding genius of the digital age" (McEwan 2019, 2).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

Let me emphasize just some points here, first, the idea and the process of programming Adam. With the robots, there comes a 470-page online handbook about how to program them, but Charlie writes:

> I couldn't think of myself as Adam's "user": I'd assumed there was nothing to learn about him that he could not teach me himself. But the manual in my hands had fallen open at chapter 14. Here, the English was plain: preferences; personality parameters. Then a set of headings – Agreeableness. Extraversion. Openness to experience. Consciousness. Emotional stability. . . . Glancing at the next page I saw that I was supposed to select various settings on a scale of one to ten. (McEwan 2019, 6)

Charlie feels uncomfortable to choose the settings since he is very aware of their reductive character. And it's not only the reductive character of the program, it's also the predictability which comes with the program, and which goes against our intuitions that human beings – although being perfectly able to follow rules and rationality – can also be unpredictable, in the sense of being unexpectedly creative when dealing with rules and given programs. Interestingly, Charlie has done a degree in anthropology at college. Why anthropology? Because the subtle sub-text (or sometimes not so subtle) is the question of the *Anthropos*, the borderline between what is and what is not human.

A second point concerns the problem of *self-knowledge and decision-making*, with the character of Turing declaring at the end of the novel:

> I think the A-and-E's [the Adams and Eves] were ill equipped to understand human decision-making, the way our principles are warped in the force field of our emotions, our peculiar biases, our self-delusion and all the other well-charted defects of our cognition. Soon these Adams and Eves were in despair. They couldn't understand us because we couldn't understand ourselves. Their learning programs couldn't accommodate us. If we didn't know our own minds, how could we design theirs and expect them to be happy alongside us? (McEwan 2019, 299)

Emotions, however, often guide human's actions for better or worse. And humans – in McEwan (2019) and in general – see themselves as being defined by not only rationality but also sentimentality. Furthermore, the suicides of the Adams and Eves exhibit something like an *uncanny zone*: Isn't it specifically human to kill oneself, to set an end to one's life? What if there is no clear cut borderline between humans and robots?

A third intriguing problem I want to point out is the problem of lying, as Turing explains to Charlie:

> Machine learning can only take you so far. You'll need to give this mind some rules to live by. How about a prohibition against lying? . . . But social life teems with harmless or even helpful untruths. How do we separate them out? Who's going to write the algorithm for the little white lies that spare the blushes of a friend? . . . We don't yet know how to teach machines to lie. (McEwan 2019, 303)

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

Lying, we can say, is also a form of creatively, self-reflectively following rules. And lastly, but centrally, I want to emphasize the robotic corporeality of Adam and the relationship between Adam and Miranda since the fact that Adam has a deceptively *human body* is a problematic subtext throughout the book. After having slept with Adam, Miranda insists that he is not more than a vibrator in human-like form, that he is "a fucking machine" (McEwan 2019, 92). She points out that she has a purely instrumental relationship to Adam, not a relation of mutual respect. Whereas Charlie's take on the situation is rather different: "'Listen', I said, 'if he looks and sounds and behaves like a person, then as far as I'm concerned, that's what he is'" (McEwan 2019, 94). But Charlie, as we will learn later in the novel, does not really mean this. He is still convinced that there's a categorical difference between Adam and himself, although he states the opposite, out of jealousy, out of defiance. He is vulnerable, not only in the bodily sense, but also in an emotional-mental sense. And, again, his contending that Adam is a person leads us directly into the uncanny field between humans and robots. Where do we draw the line?

All these themes not only demonstrate a human characteristic, but at the same time the *sociality* of human existence: the themes are, each in their own way, meaningful too, because humans always live in relationships with other humans. And this seems to be vital for understanding the characteristic differences between Charles and Adam, between human beings and robots, and therefore for understanding the essential characteristics of human beings. Embodiment/corporeality, finiteness, vulnerability, and self-knowledge, together with the (subtle, competent, possibly deviant) use of symbols, are among the classic characteristics of the human being. What is at issue in the novel is the messiness of being human, being thrown into the world without a precise 'program', and the ever-present possibility of being unable to cope with that world. This messiness expresses itself in emotional as well as bodily vulnerability, something which Adam isn't conscious of or worried about as he should be – and would be – if he was human.

3.3 CHARACTERIZING THE HUMAN

It is true that McEwan also puzzles his readers, as we saw, by the fact that some Adams and Eves commit suicide. But this too is ultimately an indication of the meaning of "being human": the reader's perception of these suicides is confused, unsure because suicide is considered a human act par excellence. It is an expression of self-knowledge (or an attempt thereto), of autonomy, and precisely not an act following a program – whereas here, in *Machines Like Me* (McEwan 2019), it is a consequence of a program error.

Earlier we saw that corporeality, finiteness, vulnerability, and the self-reflective use of language are among the classic characteristics of the human being. Based on the McEwan characteristics and especially with the help of the concept of human vulnerability, I want to analyze in the following how the concept of the human

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

being can best be understood. Vulnerability can serve as a focus of the other elements we found in McEwan. Mackenzie et al., in their volume on "Vulnerability" argue:

> Human life is conditioned by vulnerability. By virtue of our embodiment, human beings have bodily and material needs; are exposed to physical illness, injury, disability, and death; and depend on the care of others for extended periods during our lives. As social and affective beings we are emotionally and psychologically vulnerable to others in myriad ways: to loss and grief; to neglect, abuse, and lack of care; to rejection, ostracism, and humiliation. (Mackenzie et al. 2013, iv)

Human beings are vulnerable as physical beings, as affective beings, as social beings, as self-reflective beings, and this human vulnerability cannot be reduced to anything biological, although it cannot be separated from the biological either. Nor can human nature be reduced to the "brain" or "rationality," so not to their cognitive or mental abilities alone. But we have two different problems here: on the one hand whether the concept "human being" is clearly and distinctly definable in biological or physiological terms, and thus reducible to these descriptions. On the other hand, the concept of "human being" seems to carry a normative load which we would normally understand as an appeal, or maybe even a duty, to manifesting a certain attitude toward them.

To tackle this two-sidedness of the concept, it is helpful to understand "human nature" as one of the "thick concepts" which Clifford Geertz (1973), Bernard Williams (1985), or Martha Nussbaum (2023) analyze, concepts which are not purely normative or purely descriptive, but express elements of both dimensions (see also Setiya 2024). Thick concepts are both action-guiding and world-guided. "If a concept of this kind applies," Williams writes, "this often provides someone with a *reason for action* … At the same time, their application is *guided by the world*" (Williams 1985, 140–141, emphasis mine). So, when we talk about human beings, we are at the same time guided by the world and we have reasons for action – for protecting their vulnerability for instance. We follow empirical evidence *and* we are prepared to follow normative reasons for action and to respect the other as a human being, to recognize their vulnerability, and to acknowledge them as equal. The normative dimension concerns precisely those characteristics I have discussed: vulnerability, finiteness, and the self-reflective dimension of mutually recognizing each other as human.

This normative dimension of the concept of the human becomes especially clear when one looks at contexts in which the very application of the term is denied. Richard Rorty (1998) writes in his essay on human rights how, during the Balkan wars, the Serbs brutally refused to acknowledge the Bosnian Muslims as human beings, not even calling them "human." Precisely because the use of the concept "human" implies respect for others as equals – as equally human – with the application and use of the concept, the attitude which goes with it is denied

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

as well.[2] Sylvia Wynter (1994), in her famous *Open Letter to My Colleagues*, writes that, when black people were involved in accidents and injured, it was standard practice in the LAPD to report back NHI (no humans involved). Not to refer to humans as humans is a violent form of denial of respect for the other, the denial to give them the basic recognition that we owe human beings.

So, when I speak in the following of human beings, I have such a "thick" concept in mind: it is a thick concept that contains both descriptive and normative elements. Several authors in the history of philosophy have already pointed out this double-sidedness of the concept of human nature and it is taken up again in the present, for example by Moira Gatens (2019), who interprets the "human being" as Spinoza's concept of the "exemplary" (see also Neuhouser 2009 and Williams 1985). When we refer to human beings in daily contexts, we generally have in mind beings which we refer to in biological as well as ethical ways (see Barenboim 2023; Heilinger and Nida-Rümelin 2015). I want to suggest that such a concept of human being can play an essential role in the critique of the digital society: Human beings as human beings always already live in their biological nature, but at the same time in a texture, a fabric of norms and concepts that determine, or govern, or shape, the relationship with the human person herself, with others, with the world. The ways we interpret these facts change over time: in fact, the history of making sense of what a human being is forms part of what it means to be a human being.

This approach obviously does not exclude that we might simply want to stop using this concept: when we transgress human beings, as some theories propose, we should not speak of humans any longer, and maybe we will not do so in the future. But this is not yet the point.

3.4 SHOULD HUMANS BE (MORE) LIKE ROBOTS? OR SHOULD ROBOTS BE (MORE) LIKE HUMANS?

We already saw in Chapter 1 that, if the aim is to explore the possible limits of the technicization of the human, then we always need to take up two perspectives: the human becoming a robot and the robot becoming (more) human. I will here briefly remind you of the first perspective and then come back to the latter in a little more detail.

[2] We have many different examples for this attitude, the last ones from the recent Israel–Palestine war. Both sides deny the other one the property of being human; and, like Rorty, Barenboim argues: "But any moral equation that we might set up must have as its basis this fundamental understanding: There are human beings ('Menschen') on both sides. Humanity is universal, and recognizing this truth on both sides is the only way. ... Of course, especially now, you have to allow for fears, despair and anger – but the moment this leads to us denying each other's humanity, we are lost. ... Both sides must recognize their enemies as human beings and try to empathize with their perspective, their pain and their distress" (Barenboim 2023, translation mine, B.R.). Many examples from contemporary politics and warzones could be cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

The perspective that humans could be (or even should be) more robot-like covers the approaches which we described in Chapter 1 as, on the one hand, post-phenomenological – represented by Don Ihde and his followers who argue that we're always already mediated through technology. That is the reason why the idea that we humans are becoming a little more like robots is not intimidating: Verbeek argues that technology is "more than a functional instrument and far more than a mere product of 'calculative thinking.' It mediates the relation between humans and world, and thus co-shapes their experience and existence" (Verbeek 2011, 198–199). I agree with Verbeek and Ihde to the extent that their analyses of such mediations help us understand crucial aspects of what it means to be human today.[3]

However, Ihde, Verbeek, and other post-phenomenologists are not prepared to take a critical perspective here – only where does this mediation between humans and technology end? When does such an amalgam become hazardous or even dangerous for humans, so much so that they lose their humanity? The post-phenomenologists cannot answer these questions. However, there is a second understanding of the question of why people should become more technologized; this is the transhumanist understanding which we already encountered in Chapter 1 too.

Transhumanists want to increase the phenomenological connection with technology into the perfectibility of humans through technology. They explicitly build on the concept of the human being in its *ideal* version. Transhumanism endeavors to minimize all the characteristics which I described as typically human: the vulnerability, the dependence on being embodied, and eventually also the finiteness (as we know from Ray Kurzweil's vision of the singularity, see Bostrom 2002, 2005; Kurzweil 2006). Most transhumanists are not interested in *criticizing* concepts such as reason or autonomy (Ferrando 2018; Hayles 2008). On the contrary, they rather desire to get a grip on perfecting human rational and intellectual faculties, thereby overcoming vulnerability technologically and eradicating these human weaknesses or at least reducing them as far as possible. Again, I would argue that these theories are not in a position to draw a line between what one would still call human (albeit trans-human) and those beings who have given up on the 'human' in the concept transhuman altogether and are more like robots. Note, again, that I don't think this is inconsistent or impossible: I only believe that there is a borderline beyond which it is no longer appropriate or meaningful to call such beings human.

We are still left with the *opposite perspective*: why should it be bad for humans if *robots* became ever human-like? This perspective needs some more discussion, and

[3] See, on the debate about the "neutrality" of technology, on the question whether technologies are the frame, the *Gestell* which alienates us from the world (for instance Borgmann 1984; Ihde 1990; Verbeek 2011).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

I will therefore look, first, at the research on social robots and, secondly, at the (im) possibility of translating emotions into technology. As we will see, there are still clear limitations in robot–human interaction and in the attempts to making robots look and function like humans. This is particularly obvious when it comes to the expression of emotions: human facial expressions, as well as human emotional life, is so complex that no possibility of translating feelings into data seems to appear on the horizon.

The research on the meaning of embodiment and affect and the possibility of translating them into technology has recently gained a lot of traction. It is a relatively new development that technological research on robots is no longer just about the cognitive area – as has now been shown particularly well with the ChatGPT – but also about emotions and affects. Emotions not only have a conscious or rational component, they also have an experiential or a phenomenal quality which is especially difficult to translate into data (see for the following Loh and Loh 2022; Seifert et al. 2022a; Weber-Guskar 2021). So far, social robots, especially in health-care, have been met with a predominantly critical sentiment: human care should not be replaced by robots. We see this attitude also in research, where several ethical and philosophical approaches argue against this form of anthropomorphizing, from different perspectives.[4] However, in the research on social robots attempts are made to technologically develop certain human qualities in order to apply them to robots, such that they can be used in health care for elderly people or people with dementia. One of these qualities is "hug-quality," for example, another one is to be able to speak and thereby to express emotions like affection, sympathy, and care. The idea, for example, is that robots should have qualities which make it easier to hug them and easier to be addressed by them. This research on the depth of human communication is looking for developments that can improve the use of robots in care. But all this seems very difficult, following Müller's (2023) argument, as:

> AI can be used to manipulate humans into believing and doing things, it can also be used to drive robots that are problematic if their processes or appearance involve deception, threaten human dignity, or violate the Kantian requirement of "respect for humanity."[5]

So, for one thing, if robots are being used in healthcare, all they ever could do is to have an idea of instrumental care, as opposed to an idea of intentional care. What *humans* typically do when they care for others is *intentional care* and characterizes human interaction in a genuinely different way than *instrumental care*. Robots are

[4] See for instance the so-called relational approach by Coeckelbergh (2022): Coeckelbergh, too, starts with the idea of human vulnerability and seeks to interpret it normatively; see also Block et al. (2023) on research on "hug robots."

[5] See also Seifert et al. (2022a, 189), on the problems of deception and of manipulation; the whole article is very informative and convincingly argues to demonstrate the hidden problems in research programs on human–robot interaction.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

thus "care robots" only in a behavioral sense of performing tasks in care environments, not in the sense in which a human "cares" for their patients. It appears that the experience of "being cared for" is based on this intentional sense of "care" only, a form of care which robots cannot provide – or at least not at this moment. This also shows that research on human–robot interaction is still far behind their aims: emotions, responsiveness, and sympathy cannot yet be translated into data and algorithms. However, these are human qualities and characteristics which are definitive of social interaction. Weber-Guskar (2021) discusses the possibility to use data and algorithms to build emotional robots (what she calls Emotional AI systems) and is critical concerning the development as well as the social function these robots would have in communication. Similarly, Darmanin (2019) argues that the attempts so develop robots with facial expressions close to human facial expressions are completely unconvincing. If you look at the examples accessible on the Internet, he seems to be right: emotions cannot be reduced to simple datapoints (you can see examples for different emotions, like happiness, anger, fear). These expressions have at least nothing much to do with human care as we currently still understand it.

This distinction is echoed by Pasquale when he writes that the practice of caring can't be reduced – and shouldn't be reduced – to instrumental relationships which are expressed by some changes in the expression of the mouth. I agree with Pasquale, that a society organizing institutional care for people along those lines would not be a society we would want to live in (see Chapter 4, by Pasquale). If we wanted robots to replace human care, then robots would have to be either very obviously only replacing human care in the instrumental and basic sense or able to express and behave precisely like humans in providing intentional care for the ill human. It is precisely this impossibility of translating human feelings (or should we say: humanity?) into technologies that limits robotization – at least for the time being.

3.5 THE UNCANNY VALLEY

Apparently, emotions and lived experiences cannot simply be reduced to data and algorithms, even if algorithms are becoming ever smarter. Emotional as well as physical vulnerability, including diseases, that we feel (and fear) cannot be translated into technologies, in the foreseeable future – whereas in fiction, especially in novels or films, this boundary between humans and robots is messed about with. The young man who is actually a robot in the film *Ich bin dein Mensch*, for example, is deceptively similar to other men, and the woman scientist who is supposed to fall in love with him, or at least befriend him, is fundamentally insecure of her attitude toward him (Schrader 2021).

The novel *Klara and The Sun* also plays with this boundary in unsettling ways: the Artificial Friend (AF) Klara is supposed to be a "friend" of Josie, who is a young teenage girl working for her exams (Ishiguro 2021). These exams are stressful and her

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

whole future depends on the results. Furthermore, every now and then we get mysterious hints that Josie is ill and that her sister had the same illness when she died. Since the novel is told from Klara's first person-perspective, the reader is inclined to understand her quite well: also, it doesn't seem to be too difficult. She describes the way she perceives the world in (smaller and larger) squares and, for this perception, for her survival, the sun is necessary. Necessary not in the sense of natural needs that must be satisfied for an organism to live, but necessary in the sense of electricity without which a computer would not function.

Josie, on the other hand, her illness, her relationship with her neighbour Rick, as well as her authoritative mother, seems to be more of a mystery. While Klara is transparent in her perceptions, Josie remains obscure, even in her fear of illness and death. This seems to be a subtle, yet clear indication to express that Josie is the human of the two. Klara desires to be more human and has very transparent, easy-to-understand emotions. While Josie seems opaque to us, just like people who experience depression and melancholy often seem.

In a second step it becomes particularly clear that the difference between robots and humans is essentially based on the latter's vulnerability: Klara, the robot, cannot get ill, she (or it) gets *broken*. It (or she) cannot be healed, only be *repaired*. Although Klara doesn't want to break down, the robot can make that much clear – it needs the sun and is able to express this need, but it needs it like my mobile needs charging. It can't even try to survive without charging, as humans do, when they don't have food.

At least this is what the reader is led to think. Klara is asked by a woman at a party "'One never knows how to greet a guest like you,' and adds: 'After all, are you a guest at all? Or do I treat you like a vacuum cleaner?'" This question pushes us, the readers, headlong into the unsettling problematic of the relation between humans and robots. What rules are we to follow here? Which conventions apply, which conduct should we habitualize? The reader's confusion and insecurity reach even deeper. In Klara's place we – the readers – are ashamed of the woman's outrageous question, we even feel hurt, while on the other hand we know that Klara's "emotions" are alien, not human emotions, that therefore sympathy with Klara simply doesn't make sense. Ishiguro masterfully balances on the boundary between humans and robots, exploring what it means to be not-quite-human. He moves consistently on the edge of the uncanny valley. This valley itself is mysterious, and I want to briefly have a closer look at it.

The uncanny valley is a surprising valley in the previously steadily increasing curve that records the reactions of people when asked about their feelings toward robots.[6] In observing human empathy toward human-like objects, we find that the

[6] With Freud (2003), the uncanny plays more than just the role of intellectual insecurity, as Jentzsch understated, according to Freud. Both go back to the puppet Olimpia in the story "The Sandman" by E. T. A. Hoffmann, the dancing doll which is made of wood but seems to

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

more these objects, robots, resemble humans, the greater the positive response – up to a point where the objects are so human-like, but only human-*like*, that we enter the uncanny valley: we feel distressed, we feel emotionally unsettled, even extreme unease toward the objects. This is shown in the curve as a deep valley: but the valley is closing and the curve rising again when robots are indistinguishable from humans. This gap or valley is surprising since one would expect that robots, if they were almost (although not yet completely) indistinguishable from humans, would give us a reassuring or confidence-inspiring impression. On the other hand, the valley is understandable: intuitively we would always at least *like to know* whether we are dealing with a human or a robot.

Nowadays, in our daily digital lives, we seem to be confronted with a number of these uncanny areas: one example is the case of phoning a company where we no longer know whether we are being served by humans or by algorithms since the voices are indistinguishable. This is also the case with automated decision-making and the question whether there are – or ought to be – "humans in the loop": a question which is also one of dealing with the uncanny valley or field.[7]

I'm sure that in the – maybe far away – future it will be part of the rather normal world to move in this border area between beings clearly identifiable as robots, those that come across as uncanny and those which are in fact robots but no longer identifiable as such. Novels such as *Machines Like Me* or *Klara and the Sun*, or films like *Ich bin dein Mensch* or *Her* describe such a world impressively. The most recent example I came across is a short film by Diego Marcon (2021), *The Parent's Room*, just a brief clip which is haunting and truly uncanny not only because of the music and the lyrics (a father has just killed his son, daughter and wife, and is about to kill himself) but mostly because it is not entirely clear whether the figures are human, papier mâché, or a mixture of both. Isabella Achenbach writes about this film by Marcon: "[T]hat extreme representative realism evokes a response of repulsion and restlessness. . . . Marcon creates characters that give the viewer goosebumps with simultaneous feelings of aversion and unsettling familiarity" (Achenbach 2022, 293). Precisely this mixture of rejection, reluctance (aversion) and sympathy, and compassion (familiarity) characterizes the territory of the uncanny valley.

3.6 THE SELF-REFLECTIVE FINITENESS OF HUMANS

The critique or investigation of what it means to be human belongs broadly to the area of anthropological criticism. This criticism enriches the practical-normative

have human eyes and with whom the protagonist Nathanael falls in love. See Hoffmann (2020); see also Misselhorn (2009).

7 The GDPR Art 22 states that "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" (European Union, 2016). See the interesting article by Brennan-Marquez et al. (2019).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

discourse with thick descriptions of human life and helps us criticize certain digital practices, with a whole web of related thick and normative concepts, such as care, love, emotion, autonomy and freedom, respect, and equality. Taken together, they enable us to form further criteria or standards for the good, the right human life in the digital society; and this way, we build a net of anthropological criticism and ethical-political criticism. In trying to explore and search irreducible characteristics of human life, one can be guided and inspired by imaginations, by novels, by films, as we have already seen. They can help us to develop plausible narratives and thus to ask where the limits should be beyond which technologies should not further interfere with human life. Or if they do, we wouldn't be speaking of humans any longer – this is one of the critical questions, which for instance Setiya asks when he criticizes David Chalmers analogizing virtual and humans-as-we-are reality (Chalmers 2022; Setiya 2022).

These questions – and also the question of who the "we" is, which I use throughout here – are controversial and must be ever again openly discussed, contested, balanced, and determined in liberal-democratic discourses. But the criteria, characteristics, and the basic normative framework discussed here must be the background for these disputes. In the following, by way of concluding, I want to point out that there are certain normative narratives that we can use to explore the boundaries between robots and humans – and that there are others which we wouldn't so use. The aim is to be able to refer critically to those contexts in which the use of robots would not concern specific human vulnerabilities. For instance, robots in care contexts, should we use them, or shouldn't we? Should robots be used as teachers? As traffic officers? As police officers? At the checkout at the supermarket?

These are questions which are being researched already by many different universities and other public institutions, as well as by private companies, and they will occupy us even more in the future. I have argued that these questions can best be discussed if we do not simply present a short and precise definition of the human being but seek the help of normative narratives which take up the thick concepts I discussed above. We can then identify contexts within which we do or do not want to use robots and give reasons by describing the characteristics of human beings and of human relationships with these thick concepts such that the gains and losses of using robots would be visible and could be discussed. Let me raise two critical points.

Firstly, what could be the source of the feeling of uncanniness in the uncanny valley? The reason many people feel insecure vis-à-vis an almost-human-like robot is, I suggest, grounded in their vulnerabilities: the assumption, the suspicion of the impossibility of equal, respectful, emotional relationships appears as a possible dehumanization of relationships. Such dehumanization is frightening and perceived as threatening, since we mostly are frightened of the non-natural nonhuman (especially when they pretend to be human). We feel fear of those creatures, fear of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

being hurt in unknown ways. Humans have central characteristics which by definition robots do not have: we are finite, vulnerable, self-reflective beings, always already living with other humans, having relationships with them. If we want to or must expose our vulnerability, then we want to be intuitively sure that we are dealing with another person.[8] Even stronger: we always already presuppose that the other is human when we expose ourselves as deeply vulnerable beings.

The uncanny consequences of not being able to make this presupposition become clear, secondly, when we are uncertain about yet another aspect of this boundary. Remember Adam and Charles in McEwan's novel: Adam's appearance is not uncanny because he is indistinguishable from a human. Rather, it is his behavior which is uncanny: he cannot lie, and he seems mentally and, at first, physically invulnerable. Therefore, when Charles kills him, it seems, at first, rather human that he kills him without the sort of considerations one would expect him to have if he saw Adam as human. But paradoxically, Charles kills and has regrets, he feels pangs of conscience. Does having feelings of remorse and responsibility tell us more about what it means to be human than any clear definition of 'human' or precise instruction for a robot ever could?

REFERENCES

Achenbach, Isabella. "On Diego Marcon's 'The Parents' Room." In *The Milk of Dreams: Catalogue of the 59th International Art Exhibition*, 2:293. Venice: La Biennale di Venezia, 2022. https://store.labiennale.org/en/prodotto/biennale-arte-2022/

Acquisti, Alessandro, Laura Brandimarte, and George Loewenstein. "The Drive for Privacy and the Difficulty of Achieving It in the Digital Age." *Agendadigitale.Eu*, August 2, 2021. www.agendadigitale.eu/sicurezza/the-drive-for-privacy-and-the-difficulty-of-achieving-it-in-the-digital-age/.

Barenboim, Daniel. "Unsere Friedensbotschaft muss lauter sein denn je." *Süddeutsche Zeitung*, October 13, 2023. www.sueddeutsche.de/kultur/daniel-barenboim-israel-aufruf-hamas-1.6287339.

Block, Alexis E., Hasti Seifi, Otmar Hilliges, Roger Gassert, and Katherine J. Kuchenbecker. "In the Arms of a Robot: Designing Autonomous Hugging Robots with Intra-Hug

[8] This is contested in the case of friendships, and there is indeed research on friendships between humans and AI. Humans can and do have good and satisfying relationships with robots – robots which are clearly recognizable as such. This is palpable in the recent development of AI and the "friendships" that are possible between humans and such intelligent (ro)bots (see Calvo et al. 2014, and the whole volume they edited; also Block et al. 2023). Much research has been done on the ethical-philosophical side as well as on the technical side of friendships with robots, especially the relation between robots and children: children see them as friends and companions. Many people report that they do have good, even trusting, and close relations with their bots, describing them as friends without deceiving themselves about the nature of the relation (see Danaher 2019; Prescott 2021; Ryland 2021). This connects to the ethical idea of different forms of friendship which goes back to Aristotle for whom not every friendship relies on or expresses a mutuality of feelings, only if they were to be called true friends (Friedman 1993; Roessler 2015).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

Gestures." *ACM Transactions on Human–Robot Interaction* 12, no. 2 (2023): 18:1–18:49. https://doi.org/10.1145/3526110.

Borgmann, Albert. *Technology and the Character of Contemporary Life: A Philosophical Inquiry*. Chicago, IL: University of Chicago Press, 1984.

Bostrom, Nick. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9, no. 1 (2002): 1–36. https://nickbostrom.com/existential/risks.pdf.

"A History of Transhumanist Thought." *Journal of Evolution and Technology* 14, no. 1 (2005): 1–25.

Brennan-Marquez, Kiel, Karen Levy, and Daniel Susser. "Strange Loops: Apparent versus Actual Human Involvement in Automated Decision-Making." *Berkeley Technology Law Journal* 34, no. 3 (2019): 745–772. https://philarchive.org/rec/BRESLA-2.

Calvo, Rafael A., Sidney D'Mello, Jonathan Gratch, and Arvid Kappas. "Introduction to Affective Computing". In *The Oxford Handbook of Affective Computing*, edited by Rafael A. Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas, 334–348. Oxford: Oxford University Press, 2014. https://doi.org/10.1093/oxfordhb/9780199942237.013.006.

Chalmers, David. *Reality+: Virtual Worlds and the Problems of Philosophy*. New York: Allen Lane, 2022.

Coeckelbergh, Mark. "Three Responses to Anthropomorphism in Social Robotics: Towards a Critical, Relational, and Hermeneutic Approach." *International Journal of Social Robotics* 14, no. 10 (2022): 2049–2061. https://doi.org/10.1007/s12369-021-00770-0.

Danaher, John. "The Philosophical Case for Robot Friendship." *Journal of Posthuman Studies* 3, no. 1 (2019): 5–24. https://doi.org/10.5325/jpoststud.3.1.0005.

Darmanin, Godwin. "On the Possibility of Emotional Robots." *Revista de Filosofia Aurora* 31 no. 54 (2019): 804–817. https://doi.org/10.7213/1980-5934.31.054.DS08.

European Union. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)." *Official Journal of the European Union* 59, no. L 119 (2016): 1–88.

Ferrando, Francesca. "Transhumanism/Posthumanism." In *Posthuman Glossary*, edited by Rosi Braidotti and Maria Hlavajova, 438–439. New York: Bloomsbury Academic, 2018. https://doi.org/10.5040/9781350030275.

Freud, Sigmund. *The Uncanny*, translated by David McLintock. Illustrated edition. 1919. Reprint, New York: Penguin Classics, 2003.

Friedman, Marilyn. *What Are Friends For? Feminist Perspectives on Personal Relationships and Moral Theory*. Ithaca, NY: Cornell University Press, 1993.

Frischmann, Brett, and Evan Selinger. *Re-Engineering Humanity*. Cambridge: Cambridge University Press, 2018.

Gatens, Moira. "Frankenstein, Spinoza, and Exemplarity." *Textual Practice* 33, no. 5 (2019): 739–752. https://doi.org/10.1080/0950236X.2019.1581681

Geertz, Clifford. "Thick Description: Towards an Interpretive Theory of Culture." In *The Interpretation of Cultures*, 311–323. New York: Basic Books, 1973. https://philarchive.org/rec/GEETTD.

Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Kindle ed. Chicago, IL: University of Chicago Press, 2008.

Heilinger, Jan-Christoph, and Julian Nida-Rümelin, eds. *Anthropologie und Ethik*. Berlin: De Gruyter, 2015. https://doi.org/10.1515/9783110412918.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

Hoffmann, E. T. A. *Der Sandmann/The Sandman by E. T. A. Hoffmann: The Original German and a New English Translation with Critical Introductions*, edited and translated by Jolyon Timothy Hughes. Bilingual ed. 1816. Lanham, MD: Hamilton Books, 2020.

Ihde, Don. *Technology and the Lifeworld: From Garden to Earth*. Bloomington, IN: Indiana University Press, 1990. https://philarchive.org/rec/IHDTAT-3.

Ishiguro, Kazuo. *Klara and the Sun*. New York: Knopf, 2021.

Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. London: Penguin, 2006.

Loh, Janina, and Wulf Loh, eds. *Social Robotics and the Good Life: The Normative Side of Forming Emotional Bonds with Robots*. Bielefeld: transcript Verlag, 2022. https://doi.org/10.1515/9783839462652.

Mackenzie, Catriona, Wendy Rogers, and Susan Dodds. 'Introduction: What Is Vulnerability, and Why Does It Matter for Moral Theory?' In *Vulnerability: New Essays in Ethics and Feminist Philosophy*, edited by Catriona Mackenzie, Wendy Rogers, and Susan Dodds, 1–30. Oxford: Oxford University Press, 2013. https://doi.org/10.1093/acprof:oso/9780199316649.003.0001.

Marcon, Diego, dir. *The Parents' Room*. 2021. www.youtube.com/watch?v=B94pgamC3sk.

McEwan, Ian. *Machines Like Me*, 1st ed. New York: Nan A. Talese, 2019.

Misselhorn, Catrin. "Empathy with Inanimate Objects and the Uncanny Valley." *Minds and Machines* 19, no. 3 (2009): 345–359. https://doi.org/10.1007/s11023-009-9158-2.

Müller, Vincent C. "Ethics of Artificial Intelligence and Robotics". In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. Stanford, CA: Metaphysics Research Lab, Stanford University, 2023. https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/.

Neuhouser, Frederick. "Die normative Bedeutung von 'Natur' im moralischen und politischen Denken Rousseaus." In *Sozialphilosophie und Kritik*, edited by Rainer Forst, Martin Hartmann, Rahel Jaeggi, and Martin Saar, 109–133. Frankfurt am Main: Suhrkamp Verlag, 2009.

Nussbaum, Martha C. *Justice for Animals: Our Collective Responsibility*. New York: Simon & Schuster, 2023.

Pasquale, Frank. *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Cambridge, MA: Belknap Press, 2020.

Prescott, Tony. "Will Robots Make Good Friends? Scientists are Already Starting to Find Out." The Conversation. *Academic Journalism Society*. February 15, 2021. http://theconversation.com/will-robots-make-good-friends-scientists-are-already-starting-to-find-out-154034

Roessler, Beate. "Mark of the Human: On the Concept of the Digital Human Being." *European Data Protection Law Review* 7, no. 2 (2021a): 157–160. https://doi.org/10.21552/edpl/2021/2/5.

"Was bedeutet es, in der digitalen Gesellschaft zu leben? Zur digitalen Transformation des Menschen." *Abschlussmagazin des DFG-Graduiertenkollegs "Privatheit & Digitalisierung"* 1681, no. 2 (November 2021b): 20–25.

"What Is There to Lose?" *Eurozine*. February 26, 2015. www.eurozine.com/what-is-there-to-lose/

Rorty, Richard. "Human Rights, Rationality, and Sentimentality." In *Truth and Progress: Philosophical Papers*, 167–185. Cambridge: Cambridge University Press, 1998.

Ryland, Helen. "It's Friendship, Jim, but Not as We Know It: A Degrees-of-Friendship View of Human–Robot Friendships." *Minds and Machines* 31, no. 3 (2021): 377–393. https://doi.org/10.1007/s11023-021-09560-z.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004

Schrader, Maria, filmdirector. *Ich bin dein Mensch*, 2021.
Seifert, Johanna, Orsolya Friedrich, and Sebastian Schleidgen. "Imitating the Human. New Human–Machine Interactions in Social Robots." *NanoEthics* 16(2) (2022a): 181–192. https://doi.org/10.1007/s11569-022-00418-x.
Selinger, Evan, and Brett Frischmann. "Will the Internet of Things Result in Predictable People?" *The Guardian*, August 10, 2015. www.theguardian.com/technology/2015/aug/10/internet-of-things-predictable-people.
Setiya, Kieran. "Human Nature, History, and the Limits of Critique." *European Journal of Philosophy* 32, no. 1 (2024): 3–16.
"Intellectually Simulating. The World as an Illusion of Technology." *TLS*, January 21, 2022. www.the-tls.co.uk/philosophy/contemporary-philosophy/reality-plus-david-chalmers-book-review-kieran-setiya/.
Shelley, Mary. *Frankenstein* (1831 Edition), Independently Published, 2021.
Verbeek, Peter-Paul. *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago, IL: University of Chicago Press, 2011.
Weber-Guskar, Eva. "How to Feel about Emotionalized Artificial Intelligence? When Robot Pets, Holograms, and Chatbots Become Affective Partners." *Ethics and Information Technology* 23, no. 4 (2021): 601–610. https://doi.org/10.1007/s10676-021-09598-8.
Wells, H. G. *The War of the Worlds* (original 1895–1897), Grapevine (2019).
Williams, Bernard. *Ethics and the Limits of Philosophy*. Roermond, Netherlands: Fontana Press, 1985.
Wynter, Sylvia. "'No Humans Involved': An Open Letter to My Colleagues." *Forum N.H.I.: Knowledge for the 21st Century* 1, no. 1 (1994): 1–17.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 23:49:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009383851.004