AI and Ethics
https://doi.org/10.1007/s43681-021-00092-x

OPINION PAPER

Robotomorphy

Becoming our creations

Henrik Skaug Sætra[1]

Received: 15 June 2021 / Accepted: 30 August 2021
© The Author(s) 2021

Abstract
Humans and gods alike have since the dawn of time created objects in their own image. From clay figures and wooden toys—some granted life in myths and movies but also dead representations of their creators—to modern-day robots that mimic their creators in more than appearance. These objects tell the story of how we perceive ourselves, and in this article, I examine how they also change us. Robotomorphy describes what occurs when we project the characteristics and capabilities of robots onto ourselves, to make sense of the complicated and mysterious beings that we are. Machines are, after all, relatively comprehensible and help dispel the discomfort associated with complex human concepts such as consciousness, free will, the soul, etc. I then argue that using robots as the mirror image by which we understand ourselves entails an unfortunate reductionism. When robots become the blueprint for humanity, they simultaneously become benchmarks and ideals to live up to, and suddenly the things we make are no longer representations of ourselves, but we of them. This gives rise to a recursive process in which the mirror mirrors itself and influences both the trajectory for machine development and human self-perception.

Keywords Metaphor · Computational metaphor · Mechanism · Self-image · Computational theory of mind · Anthropomorphism · Robots · Robotomorphy · Behaviorism

1 Introduction

Humans and gods alike have since the dawn of time created objects in their own image. In the beginning, there were clay figures and wooden toys—some granted life in myths and movies but also dead representations of their creators. While Hobbes [1] provides an illustrative example of the early use of machines and automata as a way to understand ourselves, this article highlights how modern machines are used as mirrors for humanity. Today's sophisticated robots running artificial intelligence (AI) systems mimic both mental and physical aspects of their creators. This makes robots different from computers and non-embodied machines. While the computer is like a small make-up mirror that only lets us see our heads, robots are full frame mirrors that allow us to see intimations of ourselves in full. Their sophistication not only impresses, but inspires a certain confusion regarding who makes who, and how, in our relations with our machines [2]. Robots tell a story of how we perceive ourselves and the human condition, and in this article, I argue that they also change us.

Robotomorphy [3] is a term used to describe what occurs when we project the characteristics and capabilities of robots onto ourselves, to make sense of the complicated and mysterious beings that we are. The term mainly refers to robots, but as robots are based on computers and intimately related to other machinery, it is important to not limit this examination to robots alone. Before advanced robots there were advanced computers, and these also influence how we talk about and conceptualize the mind, for example. These discussions related to computers and other machinery are thus encompassed in the term robotomorphy, as robots are the embodiment and most human-like manifestation of these phenomena, while others are introduced as well. While we humans are frustratingly complex, machines are relatively comprehensible and provide a way to dispel the discomfort associated with human concepts such as consciousness,

✉ Henrik Skaug Sætra
Henrik.satra@hiof.no

[1] Østfold University College, Remmen, 1757 Halden, Norway

Published online: 04 September 2021

free will, the soul, etc.[1] However, using robots as the mirror image by which we understand ourselves entails an unfortunate reductionism that influences what is perceived as important, which theories are accepted, and, ultimately, power relations between individuals and groups.

Furthermore, it possibly introduces deep misconceptions about what we are, but more importantly, what we should be. When robots become the blueprint for humanity, they simultaneously become benchmarks and ideals to live up to, and suddenly the things we make are no longer representations of ourselves, but we of them. What we make and what we are co-evolve [4, 5], and this gives rise to a recursive process in which the mirror mirrors itself and influences both the trajectory for machine development and human self-perception.

This article develops the concept of robotomorphy and highlights the various ways using robots as mirrors influences our very understanding of ourselves, and also our societies through the implications this has for, for example, scientific activity. I begin by examining the use of machines as a mirror and some of the dangers involved in using metaphors. I then detail how robots also become ideals and benchmarks once the metaphor is accepted. Lastly, I explore the recursive nature of robotomorphy.

2 Machines as mirrors

The fascination with machines and automata as models of humans is not at all new, as demonstrated by Hobbes's famous quote:

> *For seeing life is but a motion of limbs, the beginning whereof is in some principal part within, why may we not say that all automata (engines that move themselves by springs and wheels as doth a watch) have an artificial life? For what is the heart, but a spring; and the nerves, but so many strings; and the joints, but so many wheels, giving motion to the whole body, such as was intended by the Artificer? Art goes yet further, imitating that rational and most excellent work of Nature, man* [1].

The machine is here used as a mirror to understand ourselves—the heart a spring, the joints wheels. By this Hobbes demonstrated two things at once: firstly, how machines can be used as mirrors; secondly, how such use of mirrors tends to be associated with a reduction of what is mirrored. In Hobbes's case, it served as a tool for developing his mechanistic view of humans [6]. In contrast to Descartes, who believed that certain psychological phenomena could not be explained by mechanism, Hobbes argued that even the mind was mechanistic, and he was the first true proponent of such a view [7]. Mechanism has ever since been a source of vivid scholarly debates, as it also become the foundation of behaviorist revolution in psychology [8], and more recently the new mechanists who seek to explain all human phenomena through mechanisms. Mechanisms are descriptions of causally determined systems or sequences that help explain the phenomenon we are interested in [9]. Examples of such mechanisms are neurotransmitter release, long-term and spatial memory, hearts, and so on [9], some of which we will explore in more detail below. One purpose of making robots is to have them perform tasks, but they also serve as tools and metaphors for trying to understand how humans function, as Hobbes did in the quote above. In her famous essay "A manifesto for cyborgs", Donna Haraway [2] deals with more sophisticated mirror images than Hobbes's automata, as she suggested that *the cyborg* could be a fictional mapping of our social and bodily reality, and that it could be a resource for our imaginations.

In addition to being a source for imagination, I am interested in the more direct use of machines as human mirrors. As with actual mirrors, one might project oneself onto one's mirror image, and while humans tend not to see their clones as separate beings, a range of animals fail to recognize that the mirror shows themselves. However, when we who understand how mirrors work see our mirror images, this directly and profoundly influences how we perceive ourselves. What we see is what we believe we look like, but just like hearing oneself on a recording tends to be awkward because our voices are different when not heard through ourselves, what we see of ourselves is not necessarily what we *are*, or how others see us.

While anthropomorphism emphasizes how people often see themselves in other things, robotomorphy is about how we see robots in ourselves. Both phenomena are of great interest, but while anthropomorphism continues to attract much academic attention, the other side of the mirroring relationship between humans and robots is often neglected. Like anthropocentrism, robotomorphy can be used both to analyze and understand the interaction between particular individuals and robots and, perhaps more importantly, the long-term shifts in human self-perception.

Perhaps, the most basic example of how we use our machines as mirrors is how the brain, and our memory, is likened to the functioning of a computer. In 1960 Miller et al. [10] coined the term *working memory*, describing something akin to the random access memory (RAM) of a computer. This is the short-term storage which is seen as distinct from the long-term storage (which in a computer

[1] I do not argue that AI systems based on deep learning and artificial neural networks are easy—or even practically possible—to predict, but rather that they are more comprehensible than human beings, and that despite their unpredictability, we know how they function and the processes that lead to the unpredictable results.

is the hard drive, or increasingly often a flash drive). This approach to the mind is referred to as the *computational theory of mind* (*CTM*), which stems from the 1940s and the work of Warren McCulloch and Walter Pitts [11]. According to this view, our minds are like computers implemented through neural activity. While the theory has had its ups and downs, and remains controversial in many circles, it remains influential, as shown by how the computer metaphor finds its way into introductory textbooks on human learning [12]. Modern theories of working memory, and the three-component model of human memory preempted by William James [13] and articulated as the dual-store model by Atkinson and Shiffrin [14], does not state that the human mind is exactly like a computer. However, the CTM assumes a certain likeness, and the computer is often used as a metaphor. Other examples are the sociobiological narratives where individuals are seen as "satisfaction- and utility-maximizing machines" implemented through a form of "genetic calculus" [2]. While metaphors help make communication effective by transposing a known concept to another, this process is also associated with certain challenges, as I will shortly return to.

The CTM is not the only way in which we transpose our creations upon ourselves. Take, for example, the suggestion that we by looking through "the lens of computer science" can "learn about the nature of the human mind, the meaning of rationality, and the oldest question of all: how to live" [15]. Christian and Griffiths [15] proceed to state, with some enthusiasm, that the computational metaphor "can utterly change the way we think about human rationality". Their basic idea is that by taking inspiration from modern algorithms, those used to excel and exceed us at games like chess, go and StarCraft, we might do better than by simply relying on the faulty and irrational instinctive human decision-making processes. Granted, they acknowledge that computers lack common sense and general intelligence. This is an insight that has recently been pointed out by, for example, Marcus and Davis [16], and preceding them is a long lineage of experts trying to convince those that do not fully understand computers that despite their impressive tricks, there are many things computers cannot do [17, 18].

The admiration for computational excellence abounds, and as robots become increasingly capable, the desire to learn from them how to optimize ourselves emerges. While blatant, the statement by Christian and Griffiths [15] does serve as a striking example of how machines that supposedly function like ourselves inspire efforts to learn from them how to overcome the inefficiencies of the human mind. These efforts connect with long and strong traditions in Western philosophy—particularly the ones where rationality is portrayed as the truest and most authentic source of human excellence and uniqueness. After all, if there is one thing robots still are not, it is encumbered with human emotions. This was one of the more prevalent and recurring areas of philosophical examination with regard to the android *Data* in Star Trek, as he both struggled to interact successfully with human beings due to this lack of emotions, while he also somewhat paradoxically seemed to yearn for emotions, which might in itself suggest that he already had certain emotions. The real life "robot mirrors" reflect an image in which the rational computer is what we see, and in which the human realities of emotions and their connection to human decision-making is stowed away, neglected, and at times forgotten, just as human phenomena such as consciousness, mind, imagination, and purpose were purportedly stowed away with the rise of behaviorism [8]. Perhaps this is why we are so inclined to depict fictional robots as encumbered with just these emotions and the problems they at times lead to, for example in the case of Data, but also in that of Marvin, the depressed robot in *Hitchhikers Guide to the Galaxy*. While the fictional representations of robots are of great interest, I mainly focus on actual robots in this article, as real robots potentially influence us and change our perceptions of ourselves in a less explicit manner than do robots in books, TV shows, and movies.

Human relationships, some argue, are shaped by a myriad of factors other than those easily observable in any specific situation, such as history, biology, and previous experiences of both joy and hardship [19]. Not just relationships, but the mere fact of being human, and experiencing, is arguably characterized by how we are not pure sense and calculation machines but "situated", both bodily and socially [4, 20, 21]. While there may be a partially shared epistemological position present in efforts to model human phenomena by way of computers and robots [22], robots have a particular advantage in allowing us to model more than human cognition in isolation. This is important, as the importance of situatedness is increasingly acknowledged [23]. Brooks et al. [23] state that the two key reasons to research humanoid robots are to pursue the engineering goal of making a "general purpose flexible and dexterous autonomous robot", while the other relates to what I am more interested in here, namely the pursuit of the "scientific goal of understanding human cognition" [24]. The pursuit of the latter approach is referred to as the *synthetic method* [25], in which artifacts are developed as models of human phenomena to gain knowledge of them. I return to this use of robots below.

Antonio Damasio [26, 27] has written extensively on the role of emotions in human decision-making, and particularly how it is inseparable from the concept of human rationality. Pure rationality, completely dissected from reason, makes little sense, he argues. This insight is one that is partially obscured by transposing the machine metaphor—in which reason *is* separated from emotions—to humans. As I will return to later, this particular discrepancy between machines and humans is now attracting increasing attention, and this is

an example of how the process of robotomorphy unfolds in practice. While Damasio [28] himself was previously skeptical of the idea that machines could be made with meaningful representations of emotions, he has more recently joined investigations into machines with certain feelings and biological mechanisms [29]. Implementing homeostasis and similar biological phenomena for robots is one thing, and others are actively pursuing machines with intimations of emotions more generally [30].

2.1 The dangers of metaphor

Metaphors help make the unknown intelligible by comparison with something that is known. This is why a cognitive psychologist or a neurologist—deeply enmeshed in the complexities and nuances of the workings of the human mind—might find the metaphor of the computer attractive. The human mind is complex, and human cognition and our memories likewise. However, a lot of people now know some basic computer science, and will, for example, recognize the distinction between short- and long-term memory—RAM and the hard (or increasingly commonly flash) drive. The experts can thus use knowledge of computers to explain in an approximate manner how human memory functions. The metaphor has worked its magic, and previous knowledge of some *other* fact is used to foster understanding of something else.

However, this process is not perfect, and a range of problems are associated with the use of metaphors. First of all, the non-experts will not really have learned how the human mind works, but they will believe that they sort of do, and they may believe that they have something akin to a computer in their heads. While their minds may previously have been mysteries to them, they may now feel a little bit more like the machines that surround them every day, even if the *actual* workings of their minds can be argued to be just as unfathomable to them as they were before they found some comfort through a superficial understanding. The CTM is so influential that most children will at some point be exposed to a wildly simplified version of it in school, and since most do not go on to become experts in neuroscience or cognitive psychology, the simplified beliefs tend to stick.

A different problem associated with metaphors such as the CTM, however, is that such models and theories will also guide and restrict the experts themselves. Even to experts, the workings of the mind are partly mysterious, and the temptation to dispel such mystery by focusing on what can be known, and fathomed, is great. Our ideals of science and beliefs about knowledge radically shape and guide what counts as relevant, and, not least, how all facts and problems are formulated [2]. The computational metaphor, for example, has arguably led to a systematic devaluation of cultural dynamics [4], and this necessitates increased attention to the interactions between culture, metaphor and technology. The metaphor is powerful, as shown by its ability to thrive despite many efforts to combat it [4, 31, 32].

The computational metaphor is intimately linked to a wide range of other theories and ideological developments, and behaviorism plays a key role. The very term robotomorphy [3] is inspired by the term *ratomorphy*, coined by Koestler [8]. Ratomorphy describes what occurs when scientists seeking to understand human (or animal) nature by studying rats proceed to transpose their findings to humans. Humans are seen as rats of sorts, and assumed driven by and susceptible to the same motivations and inclinations as those found in rats. Maslow [33] similarly lamented the fact that humans had for a long time been misunderstood because scientists had studied other animals and used these findings to draw conclusions about human nature. Humans, he argued, are not primarily motivated by physiological needs, and by assuming that they are we will paint a flawed picture of human nature by trying to learn about humans through, for example, rats.

Robotomorphy is similar to ratomorphy in that it might engender a situation in which we seek to explain all human phenomena through mechanisms we can build into our robots—or that are at least hypothetically implementable in robots [34]. Such an approach lays the groundwork for a revival of radical behaviorism, which is a phenomenon Koestler also connected with the term ratomorphy. If we accept the robot as an image of ourselves, we become little more than *conditioned reflex-automata* [8], and this is partially accepted by new mechanists directly or indirectly continuing the lineage of Hobbes and other early mechanists [9].

Also in AI ethics we find behaviorism, as John Danaher [35], for example, champions ethical behaviorism. This entails a rejection of the idea that what occurs inside our heads—in the form of unobservable thoughts and moral considerations—matters from an ethical perspective. Only actions matter. This also relates to efforts to understand human relationships, and whether or not robots can be, for example, friends, lovers, or good colleagues [36–39]. To answer these questions, researchers face the question of whether or not human relationships—or the human condition in general—is anything more than the actions they are usually associated with. If a robot says it loves you, and acts as if it loves you, who are we, Levy [38] asks, to disagree? Some draw on Goffman [40] and the idea that human activity is basically about *performances* [41], and the key question becomes: is a lover, or a friend, anything else than what a lover or a friend is expected or supposed to *do?* The discourse on robot relationships of various kinds demonstrates how robotomorphy can change our perceptions of ourselves, as researchers strive to understand and conceptualize things such as friendship and love in ways that are compatible with the capabilities of robots [34].

As we have no simple means to access the cognitive phenomena inside people's heads, this is used to justify the behaviorist approach [36]. Out of sight, out of mind. Or, more specifically, as the mind is out of sight, we will simply focus on behavior. Phenomenology, for example, offers a radically different approach, and one key message of this article is that robotomorphy drastically affects the balance of power between, for example, behaviorism and phenomenology—the approach that emphasizes the importance of human experience, despite this not being observable. My argument is that neither approach is completely wrong, and more importantly that neither approach is complete. Both behavior *and* experience might matter, and robotomorphy might entail a danger of us losing sight of important aspects of what it means to be human. This relates to the historical debates about behaviorism in psychology [8], and Burt's [42] quote could also be said to partly apply to what occurs with robotomorphy:

> *The result, as a cynical onlooker might be tempted to say, is that psychology, having first bargained away its soul and then gone out of its mind, seems now, as it faces an untimely end, to have lost all consciousness.*

3 Robots as ideals and benchmarks

The history of the computer is rich and nuanced, and an important part of this history revolves around the *competition* between the machine and humanity. The computer was quickly able to out-calculate humans, and the competition has been particularly fierce in the world of games, such as chess and go, where DeepBlue and DeepMind's Alpha-software have conquered their progenitors [43, 44]. The question has not only been whether or not the computer can defeat its creators, but also whether or not it can fool them into thinking it is one of them [45, 46]. Robotomorphy tells the tale of how robots partly shift the rules of the game. Rather than humans being tricked into believing that the robots are human, the tables are turned and humans are somehow tricked into believing that they are robots.

But this is not all. As we are tricked by the robotomorphized mirror images into thinking that we ourselves are robots, all of a sudden it is no longer the robot that must live up to human standards. It is we who must live up to the standards of the machines. Robots manifest new standards of strength, speed, and precision, most obviously perhaps in industrial settings [47]. However, robots are now increasingly also being seen as possible ideals in relation to previously human-exclusive phenomena such as care and love [34, 48]. As soon as human concepts such as "care" and "love" are adjusted to accommodate robots, the human-specific and non-computable aspects of these concepts are lost, and what remain are the parts that robots can—naturally—do very well. They can provide what might be referred to as "perfect love" [49], if love is first changed into something that does not exclude robots. They have unlimited patience, the desired amount of servility and devotion [50], and they can also be esthetically ideal, and forever young [34].

These are all important aspects of machine use, but I choose to focus on a different manner in which machines become ideals. Machines are more rational—in a certain sense an embodiment of the enlightenment logic and Western science—and humans are in this respect deeply flawed. Irrational and helpless, we are urged to aspire to the rules and performance of computational science [15]. And this is not limited to a marginal community of evangelist computer scientists. Behavioral science, as exemplified by Thaler and Sunstein's [51, 52] libertarian paternalism and nudge theory, which has mesmerized much of economics and public policy, also tell the story of how irrational and flawed humans must be helped and guided by scientific and more objectively rational considerations. Like wayward and primitive robots, humans must be shepherded, controlled, nudged, and prodded onto better paths—the paths determined by those with unsullied rationality and computationally superior capabilities.

How we shape the social construct that is humanity's self-image—constantly and changingly—is not only important because of what it means for individuals. It is also important because it entails the exercise and shift of power. For one, robots are instruments of normativity. They are used, for example in therapeutic settings involving children with autism, to demonstrate and teach what is normal and desirable, despite the fact that robots themselves are only able to mimic a relatively small subset of the various social actions one would expect from a human being. They are also relatively uniform, and represent ability, functionality, and error-free cognition *and* physicality. Humanity is in a sense a rather motley crew, whereas robots eliminate much of this disorder and introduce uniformity, both in outward appearance and in behavior.

Where the yardstick is perfect computation and rationality, human quirks and idiosyncrasies are flaws to be purged. The exemplars of human ingenuity, such as Einstein, Picasso and all the others hailed in Apple Computer's iconic "Think different" ad from 1997, were both deeply flawed and highly remarkable. The ad suggests that their peculiarities were part of the reason they were special. Today, computers are perceived by many as new kinds of "geniuses" that have solved the folding of proteins [53] and former pinnacles of human intelligence, such as the most challenging games [54]. These computers—or, more correctly, their developers—excel because the systems adhere to a pure rationality and a narrowly purposeful approach to their tasks. They are not designed to have any major quirks in the way humans do,

as such quirks would be seen as simple errors. The human champions of chess still compete for second place—the position of the best human – and while they play their games everyone watching is most concerned about how Stockfish and AlphaZero[2] evaluate their moves.

Furthermore, the computers have no sickness or disorders, and rarely do we see machines that represent and normalize individuals with specific challenges such as syndromes or disorders. A more pedestrian example is the color of machines, and their genders, and even here robots fail to encompass diversity and inclusion [55]. Normativity—the making of normal, so to speak—is related to power and superiority, and our machines are increasingly often challenged on the basis of their association with problematic social hierarchies [56]. Race is one thing, but feminists have also provided incisive critiques of how machines relate to structures of domination [2]. We might, for example, argue that robots exemplify a masculine ideal of rationality, disinterestedness, and strength—"Man, the embodiment of Western logos" [2]. However, we must also note that the structures and social constructs that are changed through robotomorphy may have been just as oppressive as those we arrive at. Haraway [2] makes this argument as she contrasts older hierarchies of domination with what she refers to as *the informatics of domination.* While robotomorphized humanity may be problematic, so may the preceding, and equally arbitrary and historically determined, images of ourselves have been.

Finally, as robotomorphy entails a celebration of logos and western science, it also poses a challenge to alternative ontologies and perspectives: the indigenous perspective [57], for example, and all other world-views in which some mystery is retained and cherished. As noted, robotomorphy coupled with behaviorism leaves little room for romanticism. If all that matters is what can be implemented in a robot, observed and scientifically proved, the soul, Mother Earth, God, etc. are no longer accepted and become quite simply what Hobbes referred to as superstition:

> Fear of power invisible, feigned by the mind, or imagined from tales publicly allowed, religion; not allowed, superstition [1].

Robotomorphy, when emerging from attempts to understand the situated nature of human cognition, the role of emotions, or even consciousness and human experience in a manner not conducive to reducing these phenomena into something easily handled by existing technology, is not necessarily as problematic as it would be when based on simple and highly reductive approaches. The different approaches to robots together shape the totality of the impact of robotomorphy, and it must be stressed that it is a dynamic concept with ever-changing implications, partly because of the recursive mirroring involved in robotomorphy.

4 Recursive mirroring

Robotomorphy is similar to anthropomorphism and ratomorphy when it comes to mirroring effects and the constitution of new benchmarks. However, it is radically different from both when it comes to robotomorphy's recursive nature. Anthropomorphism entails projecting human characteristics to things, but this is not assumed to affect the things themselves, as it relates mainly to how *we* perceive and interact with them. Similarly, ratomorphy entails a change of how we see ourselves on the basis of knowledge about rats, but this is not assumed to influence rats in any meaningful way.

Robotomorphy could be said to be special because the robots we make are seen as mirrors of ourselves, and if taken seriously, this means that these mirrors themselves might be susceptible to the very same process. If they are like us, they will, like us, be influenced by their mirror images. If we make a robot like ourselves, and if we are influenced by how our reflections appear to us, this could arguably lead to a situation in which the robot also changes when it faces us. Both humans and robots, then, are susceptible to reproducing selves "from the reflections of the other" [2]. Gunkel [58] emphasizes the importance of the concepts of the *face* and the *other*, and argues that robots can indeed be meaningful others. While robots may indeed become more than mirror images, few would argue that today's robots are in fact capable of the kind of cognitive activity that would make robotomorphy irrelevant because they are exactly like us. As argued in Sætra [59], I perceive robots as instruments of human activity, and little more than advanced tools covered by a "veil of complexity" but without the capabilities required to be considered truly autonomous in terms of having their own goals, purposes, and bearing responsibility for their own actions.

Robots are here seen as instruments capable of acting on behalf of humans, as they act on the instructions provided through programming. While these instructions are now so complicated that human designers might not foresee what the robot will do, this is argued not to matter in terms of responsibility and autonomy [59]. Robots, then, are media of sorts.[3] But media are not just neutral channels through which we communicate, and according to McLuhan [61], the media is in fact the message. I propose that robots are

[2] The most commonly used chess engines today.

[3] I wish to thank David Gunkel for pointing out to me the relevance of McLuhan and how the theory of robots I advance in Sætra [60] can be translated to the idea of machines as media, or messages.

not autonomous beings that influence us, but that robots are the creations of humans that communicate something of fundamental importance: how we grapple with understanding ourselves. Since the earliest depictions of robots in R.U.R., the play in which the word was first used and in which where they depicted as artificial worker servants [58, 62], robots have been used in fiction to highlight and analyze the human condition. This might entail examinations of the nature and morality of work and servitude, or how Data in Star Trek is used to highlight and explore the importance of emotions in human relations, or the nature of relationships, as the robot Adam does in McEwan's [63] *Machines like me.*

However, as previously mentioned, some of the work done in actual robotics is also aimed at trying to understand how human beings function by using the synthetic method. While robots are imperfect and partial representations of what we believe humans are, some of the work in computer science, AI, and robotics can in fact be seen as an effort to understand and reproduce how humans are assembled and function. Through this function of robots, we have arguably made progress on understanding a range of phenomena related to human physiology and psychology, and this points to the importance of understanding the constructive and beneficial role of trying to make robots that resemble ourselves. The synthetic method involves using robots to understand what human beings really are, and is, when successful, highly important. Biomimetic robots are not only inspired by nature, but are built as direct representations of various structures and systems found in nature [64]. Since the 1980s, Winfield [64] argues, there has been a significant shift toward biologically inspired robotics, even if not all biological inspiration leads to biomimetic robots. Some are inspired by biology, but mainly to serve other purposes and perform certain functions, or even jobs, and the biological inspiration will not necessarily result in particularly deep imitation of real biological systems. Biomimetic robots are part of the discipline known as *artificial life*—a discipline in which both computer simulations and robots are used to examine living systems. However, as already noted, humans are situated knowers, and humanoid robots may thus have a particular edge when true-to-life models of humans are the goal [23]. It must also be noted that turning to the biological realm for inspiration does not necessarily remove us from the fundamental mechanism discussed elsewhere in this article, as mechanists from Hobbes and onwards consider human beings to be biological machines.

The novelty of the concept of robotomorphy is that it provides a warning against believing that the insight gained through the use of the synthetic methods in fact does represent the *true* nature of humanity. One problem highlighted by robotomorphy is that the synthetic method might in certain cases lead to the creation of self-fulfilling prophecies whenever no objective evidence can be found to ascertain whether or not the models created through robots are in fact real representations of aspects of humanity. As robotomorphy shows, when we make such models, we change how people perceive themselves, and thus how they will act in accordance with this new self-perception, and thus potentially validate models and hypotheses that are no more true than other alternatives. Despite this potential pitfall, the synthetic method has allowed us to make increasingly complex, and intuitively increasingly correct, models of humans.

We made simple machines, which in turn inspired and facilitated models of humanity's computational nature. We arguably perfected logic, computation, and narrow reason, and thus find new ideals in our constructions. Yet something is still amiss. The machines lack a certain sophistication for us to accept them as perfect mirror images, as we still struggle to use them to explain the role of feelings, of biology, of morality. This gives rise to new efforts to construct humanoid robots that model broader aspect of human existence, including our embodied nature and the role of emotions [29, 30]. And as we do so, we wrestle with concepts such as "feelings" and "morality" and do our very best to reduce them into something computable—something that can be implemented in our robots.

This, then, means that even if the robots are limited, recursion occurs as we see ourselves a certain way and make machines to represent this, and this in turn changes and shapes how we perceive ourselves, leading us to make new and slightly different robots. The notion that we become what we behold, that we shape our tools, which in turn shape us [5], nicely captures the recursive nature of robotomorphy. Anthropomorphism explains how we see ourselves in other things, ratomorphy how we see rats in ourselves, but robotomorphy explains how we through robots construct messages conveying images of ourselves, which in turn change our images of ourselves. It is a recursive process, and it is the basis of a symbiotic and dynamic relationship between humans and our creations. As noted by Haraway [2], machines do not dominate or threaten us, as we are, in fact, they.

5 Conclusion

Robotomorphy as a concept is useful for describing how the robots we create end up shaping our perceptions of ourselves. It is linked to anthropomorphism, but it details a clearly distinct phenomenon worthy of the attention of anyone concerned with how technology influences individuals and society.

This article has told the story of how robots function as mirrors, and the argument is that the effects of these mirrors are inevitable and important. However, I have not made the argument that robotomorphy is wrong, or that

the machine metaphor is entirely misguided, erroneous, or dangerous. Through the synthetic method, new knowledge about humans' physical and cognitive capabilities is created, and these advances are potentially of great importance both in relation to physical and economic health, in addition to the economic and social benefits that can be reaped from more flexible and capable autonomous robots. However, I have argued that the behaviorism often associated with robotomorphy does not provide a complete image of human existence, and that there might be clear benefits associated with factoring in human experience in addition to human behavior.

While I consider robotomorphy to be inevitable, awareness of the mechanisms here described will potentially reduce some of its impact. However, I want to close by emphasizing one key take-away message: even if the machine metaphor leads us to conceptualize ourselves as robots, this in and of itself does not need to be accompanied with the idea that humans should be treated as such, or that robots should be our ideals. Neoliberalism, scientific management, nudging, and various other attempts to rationalize and optimize human affairs on the grounds that rationality and robotlike performance and precision are ideals, might easily gain ground with robotomorphy. It is thus important to keep in mind that what we are is not a determinant of what we *should* be, and that our values and goals as societies are not, and should not necessarily be, determined by how science, and in turn we, perceive and portray ourselves. Artifacts, and cyborgs, have politics [2, 65], and so do robots. This is not to say that they should determine politics, but that it is important that we recognize their political implications and foster popular debate about how we should deal with our machines—both robots and the non-embodied varieties.

Declarations

Conflict of interest The author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Hobbes, T.: Leviathan, p. 1651. Basil Blackwell, London (1946)
2. Haraway, D.Steven Seidman(eds): A manifesto for cyborgs: science, technology, and socialist feminism in the 1980s. In: The Postmodern Turn: New Perspectives on Social Theory, Cambridge, Cambridge University Press. pp. 82–115 (1994)
3. Sætra, H.S.: The ghost in the machine. Human Arenas **2**(1), 60–78 (2019)
4. Hayles, N.K.: Unfinished work: From cyborg to cognisphere. Theory Cult. Soc. **23**(7–8), 159–166 (2006)
5. Culkin, J.M.: A schoolman's guide to Marshall McLuhan. Sat. Rev. Ed. (1967)
6. Sætra, H.S.: Man and his fellow machines: An exploration of the elusive boundary between man and other beings. In: Orban, F., Strand Larsen, E. (eds.) Discussing Borders, Escaping Traps: Transdisciplinary and Transspatial Approaches. Münster, Waxman (2019)
7. Husbands, P., Holland, O., Wheeler, M.: Introduction: the mechanical mind. In: Husbands, P., Holland, O., Wheeler, M. (eds.) The Mechanical Mind in History. MIT Press, Cambridge (2008) . (**ch. 1**)
8. Koestler, A.: The ghost in the machine. The Macmillan Company, New York (1967)
9. Krickel, B.: *The mechanical world* (Studies in brain and mind). Springer Nature, Cham (2018)
10. Miller, G. A., Galanter, E., Pribram, K. H.: Plans and the structure of behavior. Austin: Holt, Rinehart and Winston, inc. (1960)
11. Piccinini, G., Bahar, S.: Neural computation and the computational theory of cognition. Cogn. Sci. **37**(3), 453–488 (2013)
12. Ormrod, J.E.: Human Learning. Pearson Higher Ed (2016)
13. James, W.: The principles of psychology. Holt, New York (1890)
14. Atkinson, R.C., Shiffrin, Kenneth Spence Janet Taylor Spence (eds) R.M.:+ Human memory: a proposed system and its control processes. In: Psychology of Learning and Motivation, vol. 2, pp. 89–195. Elsevier (1968)
15. Christian, B., Griffiths, T.: Algorithms to Live By: The Computer Science of Human Decisions. William Collins, London (2016)
16. Marcus, G., Davis, E.: Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon (2019)
17. Dreyfus, H.: What Computers Still Can't Do: A Critique of Artificial Reason. MIT Press (1992)
18. Dreyfus, H.: What computers can't do: the limits of artificial intelligence. New York: Harper & Row. 1972
19. Turkle, S.: Alone Together: Why We Expect More From Technology and Less From Each Other, p. Hachette UK. (2017)
20. Haraway, D.: Situated knowledges: The science question in feminism and the privilege of partial perspective. Fem. Stud. **14**(3), 575–599 (1988)
21. Lindblom, J., Ziemke, T.: Social situatedness of natural and artificial intelligence: Vygotsky and beyond. Adapt. Behav. **11**(2), 79–96 (2003)
22. Jones, R.A.: Projective anthropomorphism as a dialogue with ourselves. Int. J. Soc. Robot., pp. 1–7 (2021). https://doi.org/10.1007/s12369-021-00793-7
23. Brooks, R.A., Breazeal, C., Marjanović, M., Scassellati, B., Williamson, M.M.: The Cog project: Building a humanoid robot. In: International Workshop on Computation for Metaphors, Analogy, and Agents, pp. 52-87. Springer (1998)
24. Brooks, R.A., Stein, L.A.: Building brains for bodies. Auton. Robot. **1**(1), 7–25 (1994)
25. Cordeschi, R.: Steps toward the synthetic method: symbolic information processing and self-organizing systems in early Artificial Intelligence. In: Husbands, P., Holland, O., Wheeler, M. (eds.) The Mechanical Mind in History. MIT Press, Cambridge (2008)

26. Damasio, A.: Descartes' Error: Emotion, Reason, and the Human Brain. Quill, New York (2006)
27. Damasio, A.: Looking for Spinoza: Joy, Sorrow, and the Feeling Brain. Harcourt Inc, Orlando (2003)
28. Damasio, A.: The Strange Order of Things: Life, Feeling, and the Making of Cultures. Pantheon Books, New York (2018)
29. Man, K., Damasio, A.: Homeostasis and soft robotics in the design of feeling machines. Nat. Mach. Intell. **1**(10), 446–452 (2019)
30. Cominelli, L., Mazzei, D., De Rossi, D.E.: SEAI: Social emotional artificial intelligence based on Damasio's theory of mind. Front. Robot. AI **5**, 6 (2018)
31. Ibáñez, A., Cosmelli, D.: Moving Beyond Computational Cognitivism: Understanding Intentionality, Intersubjectivity and Ecology of Mind. Springer (2008)
32. Pettman, D.: Love in the time of Tamagotchi. Theory Cult. Soc. **26**(2–3), 189–208 (2009)
33. Maslow, A.H.: Motivation and Personality. Pearson Education (1987)
34. Sætra, H.S.: Loving robots changing love: Towards a practical deficiency-love. J. Future Robot Life (2021). https://doi.org/10.3233/FRL-200023
35. Danaher, J.: Welcoming robots into the moral circle: a defence of ethical behaviourism. Sci. Eng. Ethics **26**(4), 2023–2049 (2020)
36. Danaher, J.: The philosophical case for robot friendship. J. Posthuman Stud. **3**(1), 5–24 (2019)
37. Marti, P.: Robot companions: towards a new concept of friendship? Interact. Stud. **11**(2), 220–226 (2010)
38. Levy, D,: Love and SEx with Robots: The Evolution of Human-Robot Relationships. New York, Harper Perennial (2007)
39. Nyholm, S., Smids, J.: Can a robot be a good colleague? Sci. Eng. Ethics **26**(4), 2169–2188 (2020)
40. Goffman, E.: The Presentation of Self in Everyday Life. Harmondsworth London (1978)
41. de Graaf, M.M.: An ethical evaluation of human–robot relationships. Int. J. Soc. Robot. **8**(4), 589–598 (2016)
42. Burt, C.: The concept of consciousness. Br. J. Psychol. **53**(3), 229–242 (1962)
43. Campbell, M., Hoane, A.J., Jr., Hsu, F.-H.: Deep blue. Artif. Intell. **134**(1–2), 57–83 (2002)
44. Chouard, T.: The Go Files: AI Computer Wraps Up 4–1 Victory Against Human Champion. Nature News (2016)
45. Turing AM.: Computing machinery and intelligence. In: Robert E, Gary R, Grace B (eds) Parsing the turing test, pp 23–65. Springer (2009)
46. Searle, J.: Minds, brains, and programs. Behavioral and brain sciences **3**(3), 417–457 (1980)
47. Danaher, J.: Automation and Utopia: Human Flourishing in a World without Work. Harvard University Press (2019)
48. Sætra, H.S.: First, they came for the old and demented. Human Arenas (2020). https://doi.org/10.1007/s42087-020-00125-7
49. Sullins, J.P.: Robots, love, and sex: the ethics of building a love machine. IEEE Trans. Affect. Comput. **3**(4), 398–409 (2012)
50. Viik, T.: Falling in love with robots: a phenomenological study of experiencing technological alterities. Paladyn J. Behav. Robot. **11**(1), 52–65 (2020)
51. Thaler, R.H., Sunstein, C.R.: Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale University Press, NewYork (2008)
52. Thaler, R.H., Sunstein, C.R.: Libertarian paternalism. Am Econ Rev **93**(2), 175–179 (2003)
53. Senior, A.W., et al.: Improved protein structure prediction using potentials from deep learning. Nature **577**(7792), 706–710 (2020)
54. Google. "AlphaZero: Shedding new light on the grand games of chess, shogi and Go." https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go. Accessed 30 March 2020
55. Cave, S., Dihal, K.: The whiteness of AI. Philosophy & Technology **33**(4), 685–703 (2020)
56. Katz, Y.: Artificial Whiteness: Politics and Ideology in Artificial Intelligence. Columbia University Press (2020)
57. Gellers, J.: Rights for Robots: Artificial Intelligence, Animal and Environmental Law. Routledge, Abingdon (2020)
58. Gunkel, D.J.: Robot Rights. MIT Press, London (2018)
59. Sætra, H.S.: Confounding complexity of machine action: a Hobbesian account of machine responsibility. Int. J. Technoethics (2021). https://doi.org/10.4018/IJT.20210101.oa1
60. Sætra, H.S.: Social robot deception and the culture of trust. Paladyn J. Behav. Robot. (2021). https://doi.org/10.1515/pjbr-2021-0021
61. McLuhan, M.: Understanding media: The extensions of man. McGraw-Hill, New York (1965)
62. Horáková, J., Kelemen, J.: The robot story: why robots were born and how they grew up. In: Husbands, P., Holland, O., Wheeler, M. (eds.) The Mechanical Mind in History. MIT Press, Cambridge (2008) . (**ch. 1**)
63. McEwan, I.: Machines Like Me. Knopf Canada (2019)
64. Winfield, A.: Robotics: A Very Short Introduction. OUP Oxford (2012)
65. Winner, L.: Do Artifacts Have Politics?, pp. 121–136. Daedalus (1980)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.