

MDPI

Article

Artificial Forms of Life

Sebastian Sunday Grève 1,20

- ¹ Chinese Institute of Foreign Philosophy, Beijing 100871, China
- ² Department of Philosophy, Peking University, Beijing 100871, China; ssg@pku.edu.cn

Abstract: The logical problem of artificial intelligence—the question of whether the notion sometimes referred to as 'strong' AI is self-contradictory—is, essentially, the question of whether an artificial form of life is possible. This question has an immediately paradoxical character, which can be made explicit if we recast it (in terms that would ordinarily seem to be implied by it) as the question of whether an unnatural form of nature is possible. The present paper seeks to explain this paradoxical kind of possibility by arguing that machines can share the human form of life and thus acquire human mindedness, which is to say they can be intelligent, conscious, sentient, etc. in precisely the way that a human being typically is.

Keywords: artificial intelligence; artificial general intelligence; human–machine interaction; machine consciousness; Turing test; Chinese room; humanoid robot; Ludwig Wittgenstein; Hubert Dreyfus; John Searle

1. The Biological Objection

Perhaps the only important objection that Alan Turing failed to discuss in his famous essay 'Computing Machinery and Intelligence' (1950 [1]) is the idea that machines could never be truly intelligent because they are not alive. This is what I am going to call 'the biological objection'. Turing's omission of this particular problem at that point in time can be forgiven, because it arguably arises only at a later stage in the dialectic. Indeed, the biological objection may be regarded as the ultimate problem of the metaphysics of artificial intelligence.

The closest analogue Turing discussed was what he called the 'argument from consciousness', according to which machines could never be truly intelligent because they are not conscious. The argument from consciousness has attracted much more attention than the biological objection, but it seems reasonable to expect the balance to shift in the opposite direction. One reason why the biological objection has not really entered the public conversation until now is that most people, reasonably, take it for granted that a machine cannot be alive. Moreover, most scientific definitions of life agree with common sense on this point. By contrast, the possibility of machine consciousness has long fascinated the public imagination and been a subject of scientific debate. It is precisely the deep-rootedness and intuitive nature of the common belief that a machine cannot be a living being that reveals the necessity of the biological objection as a last resort for those who cannot accept the possibility that machines could ever be truly intelligent, even in the face of perfectly human-level or superhuman performance by a machine: 'I may not know whether this thing (program, robot, etc.) might not have some kind of consciousness, but at least I know it is not a living being, so whatever "intelligence" or "consciousness" it may possess must be essentially different from ours'.

One formulation of the biological objection, then, is to say that machines could never acquire mental capacity *x* because they are not alive. But the most general and most important instance of it is given in terms of human mindedness, that is, the sort of mindedness that is typical of human beings. It may be formulated as follows:



Citation: Sunday Grève, S. Artificial Forms of Life. *Philosophies* **2023**, *8*, 89. https://doi.org/10.3390/ philosophies8050089

Academic Editors: Juliet Floyd and Sandra Laugier

Received: 18 July 2023 Revised: 5 September 2023 Accepted: 11 September 2023 Published: 22 September 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Philosophies **2023**, 8, 89 2 of 21

The biological objection: Human mindedness requires being alive, but machines are not alive; so machines cannot have human mindedness.

I argue against this objection mainly by presenting an argument to the contrary, according to which machines *can* acquire human mindedness.

Before I proceed, I should briefly say what I will generally mean by 'machine' in this essay. Turing in his 1950 paper [1] restricted 'machines' to mean digital computers, that is, the same type of engineered and programmed artefact as the vast majority of our modern-day computing devices. It seems both natural and convenient to follow Turing on this point here, adding the further stipulation that these artefacts must be primarily made from non-living materials such as silicon or plastic.

1.1. Searle, Dreyfus, and Fuchs

It is not widely appreciated that the biological objection constituted the hidden premise in John Searle's critique of what he called 'strong AI', which he famously expressed in the well-known Chinese room thought experiment. What he actually had to say in this connection is therefore worth quoting at length:

It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality (I am, I suppose, the instantiation of any number of computer programs), but as far as we know it is because I am a certain sort of organism with a certain biological (i.e., chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena. . . . It might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those that human beings use. . . . That is an empirical question, rather like the question whether photosynthesis can be done by something with a chemistry different from that of chlorophyll. (Searle 1980 [2], 422)

Searle believes that human mindedness, or something very similar, could be instantiated by an organism with a chemical constitution different from that of a human being, but not by anything that is not a living organism (and hence not by a machine).

It is at least arguable that implicitly proceeding from the biological objection was also the secret strength of Hubert Dreyfus's influential critique *What Computers Can't Do* (1972/92 [3]). I will quote a few passages from this work in Part 2 below. If Dreyfus did indeed do that, it would certainly be in line with some strands in the phenomenological tradition, upon which he prominently drew, and more recent and more direct expressions of the biological objection have indeed come from the same broad tradition. In particular, Thomas Fuchs has recently presented what seems to be the most general, explicit, and emphatic account of the objection to date. For example, Fuchs writes: 'Only living beings are conscious, sensing, feeling, or thinking' (2022 [8], 254). He argues as follows:

The maintenance of homeostasis and thus the viability of the organism is the primary function of consciousness, as manifested in hunger, thirst, pain, or pleasure. Thus arises the bodily self-experience, the *sense of life*, which underlies all higher mental functions. . . . Without life there is no consciousness and also no thinking. (Fuchs 2022 [8], 254)

Fuchs is thus led to consider whether future humanoid robots might not only simulate life but actually be alive.³ He argues that no machine can ever be alive because machines cannot be autopoietic systems.⁴

For reasons to be explained, I actually do not want to argue directly about whether machines can be alive or not, nor even about whether human mindedness requires the state of being alive. Let me just briefly state the following two points in this connection.

First, the fundamental question 'What is life?' has not yet received any very satisfactory answer. On the contrary, there remains widespread disagreement on the matter even

Philosophies **2023**, *8*, 89 3 of 21

amongst biologists.⁵ Moreover, some definitions of life (for example, as autopoiesis) are such that it is not obvious that a humanoid robot made from silicon, for instance, could not in fact satisfy its conditions. It is perhaps not a coincidence that the idea of autopoiesis as a definiens of life was first introduced as the composite term 'autopoietic machine'.⁶

Second, there are in fact reasons to believe that there are possible chemical environments in which silicon (14Si) may perform an equivalent function to the one that carbon (6C) normally performs on Earth, that is, to be the motor of biochemistry, so that being silicon-based need not in itself be an excluding factor. Similarly, there are reasons to believe that carbon-based living organisms can successfully run and internalise computer programs, including learning algorithms, imposed on them from an external non-living silicon-based source, so that being an artificial computer need not be an excluding factor either.

1.2. An Essentially Emotional Position

The first of two main reasons why I wish to remain neutral on both the question of whether human mindedness requires life and that of whether machines can be alive is that I think I can. In the other two parts of this essay (Parts 2 and 3), I will present an argument according to which whatever human mindedness requires, machines can probably fit the bill. This argument entails, trivially, the conditional that if human mindedness requires being alive, then machines probably can be alive, but it entails no more than that regarding either the antecedent or the consequent, despite (or, indeed, in virtue of) my employment of the idea of a form of life.

The second reason is a little bit more complicated. To put it briefly, it is that I do not wish to argue directly with what I take to be an essentially emotional position. For the real cause of the present appeal of the biological objection is, in my view, not merely ignorance but a problematic mixture of ignorance and emotional bias. It is thus in my view a possible version of an important objection that Turing already discussed:

The 'Heads in the Sand' Objection.' The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so.'

This argument is seldom expressed quite so openly as in the form above. But it affects most of us who think about it at all. We like to believe that Man is in some subtle way superior to the rest of creation. It is best if he can be shown to be *necessarily* superior, for then there is no danger of him losing his commanding position. . . . This feeling . . . is likely to be quite strong in intellectual people, since they value the power of thinking more highly than others, and are more inclined to base their belief in the superiority of Man on this power.

I do not think that this argument is sufficiently substantial to require refutation. Consolation would be more appropriate. (Turing 1950 [1], 444)

I think Turing is right to point out this widespread fear, which inclines people towards an irrational humanist bias on the matter in question, and right too in his evaluation of the discursive situation. Other well-known biases, in particular, sexism, racism, and speciesism, have often had the same kind of effect (i.e., bias towards one's own kind with regard to the question of mental superiority). Notably, these kinds of biases may be variously motivated. In particular, they may be motivated by fear as well as contempt (in other words, they may be instances of xenophobia as well as of misoxeny). Which of these two emotions, fear or contempt, will be stronger in a given individual is partly a doxastic function. For example: all things being equal, someone who believes machines to be potentially very powerful will incline towards fear of the possibility that they may acquire human mindedness, while someone who believes machines not to be potentially very powerful or indeed, perhaps, quite limited in their possible development will incline towards contempt for the idea that they may acquire human mindedness.

In recent centuries, prejudice in humans against the possibility of other beings' mindedness has gradually decreased. This reduction has followed a roughly linear trajectory up

Philosophies **2023**, 8, 89 4 of 21

the ranks of standard biological taxonomy. Major stops thus far have been sex, race, and species. Since Darwin, there has probably been a steady increase in the number of people who accept that other species of animals possess significant mental capacities, but it seems a fair estimate to say that psychological speciesism has only significantly receded in the last half-century or so. More recently, a growing number of scholars from different disciplines have been making the case for the existence of mental capacities in non-animal organisms including plants, protista (e.g., slime moulds), and bacteria; so there is a growing body of scientific evidence that humans have also shown a significant amount of unwarranted bias towards their own kind in the form of what might be called 'regnumism', specifically that they have unduly excluded non-animal organisms from the category of minded beings. ¹⁰

Today, in the face of the idea that machines may acquire human mindedness, this historical development seems to be approaching its final stage, which might be called 'bioism': hence, in the present context, the exclusion of non-living beings from the category of minded beings or, more weakly, from a category of beings that are minded in a particular kind of way (e.g., the sort of mindedness that is typical of human beings). Thus, considering how little we currently know concerning the relevant fundamental questions about life, mind, and machines, the present appeal of the biological objection seems to be essentially a function of an irrational bias towards what is perceived to be one's own kind. To be sure, something's being the result of an irrational (or emotional) bias does not mean it is wrong, but it invariably complicates scientific discourse on the matter.

In the remaining two parts, I will first introduce a certain notion of a form of life and then use it to argue that machines can acquire human mindedness.

2. Concepts, Metaphysics, and Natural History

Another philosopher on whose work Dreyfus prominently drew, outside the phenomenological tradition, was Wittgenstein. Fuchs, too, seems to have Wittgenstein in mind when he stresses the importance of an exclusively human form of life. 11 Each of these three authors centrally employs a notion they designate using the words 'form of life' (German *Lebensform*) when discussing the metaphysics of artificial intelligence, and especially when discussing the possibility that machines may acquire human mindedness. Below is a small selection of the kinds of thing they say:

To imagine a language means to imagine a form of life. . . . It is in their *language* that human beings agree. This is agreement not in opinions, but rather in form of life. . . . We say only of a human being and what is like one that it thinks. . . . The word "hope" refers to a phenomenon of human life. (Wittgenstein 1953/2009 [26], 19, 241, 360, 583)¹²

It should be no surprise that nothing short of a formalization of the human form of life could give us artificial intelligence. (Dreyfus 1972/92 [3], 222)

The basic condition for understanding turns out to be the sharing of a common form of life. . . . Artificial systems are unable in principle to fulfill this fundamental condition. (Fuchs 2022 [9], 3)

Dreyfus also made it clear that he tended to agree with this last sort of statement when he wrote, first, that 'the human form of life cannot be programmed' (224, note 37) and, second, that 'the formalization of our form of life may well be impossible' (281–282).

Fuchs gives a more succinct expression to his reason for believing that machines are categorically excluded from any distinctively human form of life: 'Robots do not belong to this shared form of life, since they do not have a *vital* and thus *phenomenal embodiment'* (Fuchs 2022 [9], 9). This is unmistakably an instance of the biological objection. Fuchs furthermore writes: 'Without life there is no consciousness, no feeling and no thinking' (Fuchs 2021 [7], 27). Thus, his position in fact constitutes what I have described as the stronger version of bioism about the mental (i.e., the general exclusion of non-living beings from the category of minded beings). By contrast, what I described as weak bioism is the exclusion of non-living beings from a given category of beings that are minded in a

Philosophies **2023**, *8*, 89 5 of 21

particular kind of way. For example, there is the sort of weak bioism that is probably the most intuitive variant of bioism today, according to which machines may acquire *some*, or some kind of, mindedness (intelligence, emotion, consciousness, etc.)—as many people seem quite ready to admit nowadays—yet according to which machines, as long as they are not alive, may never acquire the special kind of mindedness that is human mindedness.

Unlike Dreyfus and Fuchs, Wittgenstein never concluded that machines could not acquire human mindedness. Therefore, it seems that either he did not believe that human mindedness required being alive or he did not believe that machines could not be alive. However, he agrees with Fuchs and Dreyfus that the human form of life is constitutive of human mindedness. Thus, it seems that on Wittgenstein's account sharing the human form of life does not require being alive. This may sound paradoxical, but it is in fact Wittgenstein's account. Crucially, Wittgenstein's theoretical perspective is apt to show where the biological objection goes wrong, even if it were ultimately to turn out that human mindedness does require life.

2.1. Avoiding the Charge of Human Chauvinism

What is a form of life, anyway? There exists considerable agreement between Dreyfus, Fuchs, and Wittgenstein. For example, Wittgenstein could have substantially agreed with what Fuchs says in the following passage, except for Fuchs's central contention that sharing a form of life requires being alive:

We perceive others from the outset as embodied participants in a common form of life, in which we do not merely infer selfhood from signs but always already presuppose it. This intercorporeal perception is bound up with our common aliveness, embodiment and life history. We share with others the existential facts of being born and growing, the need for air, food and warmth, waking and sleeping, last not least mortality; and this is the common background against which we also interpret all their verbal utterances. . . . Our everyday sharing of emotions and intentions with others presupposes a sharing of life. Whatever can feel hunger, thirst, pleasure or pain, joy or suffering, so that we can empathize with these states, must be of our kind in the broadest sense, that is, a living being belonging to our species or descended from another species whose expressions of emotion and striving are sufficiently similar to ours. (Fuchs 2022 [9], 10)

Similarly, Wittgenstein writes:

Only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious. (Wittgenstein 1953/2009 [26], 281)

Both Fuchs and Wittgenstein use the notion of a form of life to describe human mindedness and how our psychological concepts may or may not be applied to non-human beings. Both do so in biological as well as (in the broadest sense) sociological terms. Fuchs is quick to posit the view that biology offers the more fundamental insight: 'Sociality presupposes *conviviality'* (2022 [9], 9). Wittgenstein does no such thing.¹³

Before explaining why not, and what it is that Wittgenstein is doing instead, we must briefly address a pair of objections that both authors are likely to encounter in connection with their shared view of the primacy of the living human being: namely, the charge of anthropocentrism and the charge of anthropomorphism. The charge of anthropocentrism, in this connection, relates the view that psychological terms should not be defined purely on the basis of the human case. The charge of anthropomorphism, on the other hand, relates the view that psychological terms should not be uncritically extended from the human case to non-humans. It is indeed crucial, and difficult, to avoid making these mistakes, and perhaps impossible to escape being charged with having committed one or the other. For example, Wittgenstein's words in the quotation above do not entail that whatever can be said to have sensations, etc. must be alive. Resembling and behaving like a living human being implies neither being human nor being alive. This feature of Wittgenstein's

Philosophies **2023**, 8, 89 6 of 21

statement may well appease the anti-anthropocentrist but, simultaneously, provoke the anti-anthropomorphist.

I think John Haugeland was essentially right, in his discussion of the possibility of artificial intelligence, to say that the problem is that "'human chauvinism" is built into our very concept of intelligence' (1985 [29], 5). He explains: 'This concept, of course, could still apply to all manner of creatures; the point is merely that it's the only concept we have—if we escaped our "prejudice", we wouldn't know what we were talking about' (5). The point holds in principle for all common psychological concepts, as both Fuchs and Wittgenstein would no doubt agree. Regarding Haugeland's suggestion that such a concept 'could still apply to all manner of creatures', Fuchs (but not Wittgenstein) would of course have to insist on the qualification 'living creatures'. The importance of this view, that 'human chauvinism' is built into our psychological concepts, will become clearer in the course of the subsequent discussion.

2.2. Wittgenstein's Rejection of the Biological Objection

In several of the following quotations from Wittgenstein's *Philosophical Investigations* [26], he is characteristically presenting his thoughts in the form of a dialogue whose different voices can at times be difficult to identify. I do not have the space to discuss these exegetical matters in detail. I hope my interpretation is coherent enough that readers who are unfamiliar with this aspect of Wittgenstein's work can still find their bearings.

The things whose potential mindedness Wittgenstein considers in the passages I will quote include non-human animals, machines, chairs, and stones. Let us begin with the case of machines:

359. Could a machine think? — Could it be in pain? – Well, is the human body to be called such a machine? It surely comes as close as possible to being such a machine.

360. But surely a machine cannot think! – Is that an empirical statement? No. We say only of a human being and what is like one that it thinks. We also say it of dolls; and perhaps even of ghosts. Regard the word "to think" as an instrument! (Wittgenstein 1953/2009 [26], 359–360)

Three things are noteworthy about this pair of remarks for our present discussion. First, Wittgenstein thought that machines can be meaningfully (and, by chance, truly) said to think, and probably to be in pain. Second, he thought that the precise meaning of saying such a thing will depend on just how much a given machine is like a human being (this is also the topic he pursues in the subsequent section 361, which begins 'The chair is thinking to itself . . . '). Third, he thought it useful in this connection to consider the question of why the human body is not normally considered a machine.

A possible answer to this question that Wittgenstein considers in various places, including in the same context, but ultimately rejects, is life. Two sections are especially relevant. Firstly:

357. We do not say that *possibly* a dog talks to itself. Is that because we are so minutely acquainted with its mind? Well, one might say this: if one sees the behaviour of a living being, one sees its mind. – But do I also say in my own case that I am talking to myself, because I am behaving in such-and-such a way? – I do *not* say it from observation of my behaviour. But it makes sense only because I do behave in this way. . . .

As in section 360, Wittgenstein starts by pointing out in the form of a question that he is here concerned to analyse statements—which may be expressed by saying 'possibly a machine thinks', 'possibly a dog talks to itself', etc.—that are trying to make a logical, conceptual, or metaphysical point rather than to report empirical findings; statements, in other words, whose evidential basis is made up of deductive rather than inductive inferences, starting from views concerning the subject matter in general rather than from

Philosophies **2023**, *8*, 89 7 of 21

observations concerning particular instances. ¹⁴ He then introduces a possible response to the question, which asserts that we do indeed know the minds of dogs so very well, because 'if one sees the behaviour of a living being, one sees its mind'. He does not pause to reject this idea. Instead he notes that, although a corresponding claim of human self-knowledge such as 'I am thinking' or 'I am talking to myself' is not generally based on observation of one's own behaviour, 'it makes sense only because I do behave in this way.' Wittgenstein's repeated emphasis on behaviour as a necessary condition for meaning even in the first-person singular is an important feature of his overall position, which will become clearer in the final part of this essay. He continues his line of investigation about the role of being alive in the following passage, where he explicitly rejects the suggestion that life made an essential difference that was somehow required over and above observable behaviour:

284. Look at a stone and imagine it having sensations. – One says to oneself: How could one so much as get the idea of ascribing a *sensation* to a *thing*? One might as well ascribe it to a number! – And now look at a wriggling fly, and at once these difficulties vanish, and pain seems able to get *a foothold* here, where before everything was, so to speak, too *smooth* for it.

And so, too, a corpse seems to us quite inaccessible to pain. – Our attitude to what is alive and to what is dead is not the same. All our reactions are different. – If someone says, "That cannot simply come from the fact that living beings move in such-and-such ways and dead ones don't", then I want to suggest to him that this is a case of the transition 'from quantity to quality'.

Wittgenstein's rejection of the statement in quotation marks is at the same time his rejection of the idea that in order for us to correctly ascribe mental capacities to something, it must not only display a certain quantity of life-like behaviour but also possess a certain independent quality such as being alive. The alternative conception he suggests appears to be that a certain quantity of life-like behaviour constitutes a threshold at which there occurs a qualitative change ('a transition "from quantity to quality"'), so that mental capacities may be correctly ascribed. ¹⁵ Of course, merely behaving like a living human being cannot be enough, or else it would follow that a string puppet could have mental capacities as long as it is played sufficiently well. Wittgenstein indicates what else he thinks is required in terms of what he calls a 'form of life'.

2.3. 'Form of Life' in Wittgenstein

In a first approximation, Wittgenstein's special use of this term can be explained as a way to stress that an account of any part of human nature, including any given piece of human behaviour, must attend to it as a historical phenomenon in the widest possible sense. Thus, he stresses that 'to imagine a language means to imagine a form of life' (1953/2009 [26], 19), that 'the *speaking* of language is part of an activity, or of a form of life' (23), and that agreement in language is 'agreement not in opinions, but rather in form of life' (241). For the same reasons, he stresses that 'giving orders, asking questions, telling stories, having a chat, are as much a part of our natural history as walking, eating, drinking, playing' (25).

The fact that Wittgenstein did not think sharing a form of life required being alive will perhaps seem less surprising once it is understood that the words 'form of life' designate a notion of the highest order of abstraction in Wittgenstein's mature work. Following the revolution in his thinking in the 1930s, form of life replaces the earlier logical form as a key concept in his philosophical method. Whereas previously formalisation aimed at discovering logical forms (which were to be specified using symbolic logic), it now aims at describing forms of life, which are to be specified using real or imagined natural history (including real or imagined language-games, etc.). Commentators have largely been divided over whether Wittgenstein's notion of a form of life is primarily a sociological or a biological one. Much less attention has been given to the fact that Wittgenstein does not say 'feature', 'trait', 'quality', 'mark', 'property', 'way', 'kind', 'characteristic', or what have you, but instead deliberately chooses to say 'form'. As we shall see, the

Philosophies **2023**, *8*, 89 8 of 21

forms of life that Wittgenstein is primarily interested in are typically individuated on both biological and sociological grounds, that is, in terms of natural history in the broad sense in which Wittgenstein employs this notion when he speaks of 'our natural history' (25) or 'the natural history of human beings' (415). So, whilst there is thus a reason why these forms are typically taken (lifted, abstracted) from life, there is no reason to think that these forms may not be shared by something that is not a living being.

Wittgenstein speaks of forms of life on a smaller and bigger scale, just as he uses the term 'language-game' not only to speak of an instrument of logical analysis but also to refer to various subsets of our language or others. Indeed, his term 'language-game' can generally be interpreted to mean *linguistic form of life*.²⁰ Thus, the well-known list of examples he gives of language-games in section 23 of the *Investigations* is equally a list of examples of (linguistic) forms of life:²¹

The word "language-game" is used here to emphasize the fact that the *speaking* of language is part of an activity, or of a form of life.

Consider the variety of language-games in the following examples, and in others:

Giving orders, and acting on them –

Describing an object by its appearance, or by its measurements –

Constructing an object from a description (a drawing) -

Reporting an event -

Speculating about the event -

Forming and testing a hypothesis –

Presenting the results of an experiment in tables and diagrams –

Making up a story; and reading one -

Acting in a play -

Singing rounds -

Guessing riddles -

Cracking a joke; telling one -

Solving a problem in applied arithmetic –

Translating from one language into another –

Requesting, thanking, cursing, greeting, praying. (1953/2009 [26], 23)

The primary reason behind Wittgenstein's stressing here that speaking is part of an activity or, as he puts it, of a form of life is also the primary reason behind his general insistence that any account of human nature must attend to it as a historical phenomenon, namely, that this is the only way to avoid superficial, oversimplifying psychological explanations.

2.4. Forms of Life as Patterns of Activity

A useful example is Wittgenstein's discussion of what it is to point at the shape as opposed to the colour of a given object (or vice versa), say, in the case of a beautiful blue vase:

You'll say that you 'meant' something different each time you pointed. And if I ask how that is done, you'll say you concentrated your attention on the colour, the shape, and so on. But now I ask again: how is *that* done? (Wittgenstein 1953/2009 [26], 33)

Wittgenstein goes on to illustrate with a number of examples, and quite correctly, that we do not always do the same thing when, for instance, we attend to the shape of something; furthermore, and to my mind equally correctly, that there may be characteristic experiences when pointing at the shape but there does not seem to be one characteristic process that occurs in all cases, yet we have a tendency to attempt a general explanation by positing an alleged mental, perhaps unconscious process (such as 'concentrating one's attention'), which in reality explains nothing:

And we do here what we do in a host of similar cases: because we cannot specify any *one* bodily action which we call pointing at the shape (as opposed

Philosophies **2023**, 8, 89 9 of 21

to the colour, for example), we say that a *mental* activity corresponds to these words. $(36)^{22}$

To be sure, Wittgenstein is not saying that there could be no reductive psychological (e.g., neurological) account of such a thing as pointing at the shape. Rather, he is concerned with showing that the way we are presently applying the relevant concepts is practically independent of any such inner process. Thus, he writes:

Even if something of the sort did recur in all cases, it would still depend on the circumstances – that is, on what happened before and after the pointing – whether we would say "He pointed at the shape and not at the colour". (35)²³

Again, I believe Wittgenstein is quite right in what he is saying.²⁴ Now his stressing of the importance of the circumstances, specifically of what happens before and after the pointing, might equally well be expressed by saying that pointing at the shape (as opposed to the colour, for example) is a form of life, where the term 'form of life' is used to emphasise the fact that the pointing is part of an activity. Moreover, Wittgenstein typically uses this term to emphasise that any more general account of the phenomenon in question (say, a given act of pointing), that is, of the kind of phenomenon it is, must attend to it by widening the historical perspective ('what happened before and after') so as to bring into view the fact that this activity is part, in turn, of a *pattern of* activity—or, as he sometimes says, a custom or practice—for example, by reminding ourselves of the variety of activities that have fallen under the same concept.²⁵

Thus, a form of life in Wittgenstein's sense is objectively nothing but a pattern of activity. The reason why he prefers 'form of life' or, as he also sometimes writes, 'pattern of life' (*Lebensmuster*) is that this kind of expression stresses the organic origin of the patterns of activity he is primarily concerned with and hence, importantly, their open structure in terms of an amenability towards future extension (which goes hand in hand with the kind of conceptual growth Wittgenstein famously likened to the growth of a family over many generations).²⁶

Not all forms of life in Wittgenstein's sense of the term are linguistic in nature, although arguably any human form of life in his sense, whether big or small, is essentially a or the form of life of a speaking being. In the case of hope, which Wittgenstein discusses in a number of places, this is relatively obvious. Not all instances of hope need involve linguistic representations. Nonetheless, what we call 'hope' in humans is essentially linguistic in nature, or else there could be no such characteristic instances as hoping that she will be back in two or three days, hoping that their marriage will be happy, or hoping that the war will be over soon.²⁷ It is much less obvious that this should be equally true of things like seeing or hearing and even of being blind or deaf, as Wittgenstein seems to be suggesting, for example, when he counts these things among those that can only be ascribed to a living human being or what resembles one (as quoted above). Yet, as we shall see more clearly in the third and final part of this essay, Wittgenstein's conception of a form of life entails that a lot more than is ordinarily thought to be distinctively human will in fact be so, especially if there is to be such a thing as the human form of life, in the sense of a single form that encompasses all of human life either in general or at a given time and place. Thus, Wittgenstein's conception of a form of life leads to a fuller appreciation in both its wealth and specificity of what it is to be minded in the way typical of a human being (indeed, so full an appreciation of the unique nature of human mindedness that, I think, it may move some proponents of the biological objection to give it up).

3. Human versus Non-Human Mindedness

It is a truism that whatever we can empathise with must be, in some sense, of our kind. Fuchs believes that the relevant kind is 'a living being belonging to our species or descended from another species whose expressions of emotion and striving are sufficiently similar to ours' (Fuchs 2022 [9], 10). If we asked him whether that was an empirical statement, that is, whether perhaps his justification for saying this was that he is so minutely acquainted with

Philosophies **2023**, *8*, 89 10 of 21

all non-living beings, the answer would likely be 'No'.²⁸ The present, final part of this essay will conclude my argument that no matter what human mindedness requires, machines can probably fit the bill. The same argument will show that we can probably empathise with machines as well as with humans regardless of whether machines can be alive or not.

Let us begin by asking the following very general question. How do we even ascribe any sort of mindedness to anything at all, including ourselves? Beginning in this way will also let us see that empathy does not reach beyond the human case, or indeed up to it, as easily as one might think. Wittgenstein famously had something to say about this. The perhaps most original move in his influential remarks about private language is to turn around the stakes of the problem of other minds, and ask whether I could possibly have a language that describes my inner experiences but which only I myself can understand.²⁹ Another original idea that Wittgenstein argues, to my mind, convincingly is that our natural languages are forms of life that entail not only the existence of other minds but indeed what I shall call their 'likemindedness', that is, the qualitative identity between the mental states of another and my own, in a large number of cases. The reason Wittgenstein gives can briefly be stated as being the way in which the reference of some relevant linguistic expressions is fixed. For example, he writes:

How does a human being learn the meaning of names of sensations? For example, of the word "pain". Here is one possibility: words are connected with the primitive, natural, expressions of sensation and used in their place. A child has hurt himself and he cries; then adults talk to him and teach him exclamations and, later, sentences. They teach the child new pain-behaviour. (Wittgenstein 1953/2009 [26], 244)

I take it that some of our basic psychological terms ('pain', 'hungry', 'angry', 'scared', etc.) are indeed naturally taught and learnt, and hence have their reference fixed, in this kind of way. But then, as Wittgenstein was keen to point out, the reference of these terms will be fixed publicly, and the states they consequently designate will be things that are in principle publicly accessible. Learning and speaking the same natural language is thus built on an implicit assumption of likemindedness: specifically that, all else being equal, psychological terms when successfully applied will refer to qualitatively identical mental states of different people, and will do so when used in the first, second, or third person alike (this is what Wittgenstein meant by saying that claims of self-knowledge are not generally based on observation of one's own behaviour but make sense 'only because I do behave in this way').

There is a temptation to think that under normal circumstances the child will, more or less automatically, generalise its early knowledge of ascribing mental states to others that it obtained through interacting with a small circle of people, and extend that knowledge to the class of all human beings. But this is not obviously so. For example, it is not obvious that what is usually described from a moral point of view as children's 'cruelty', especially when the relevant behaviour is directed against other children or animals, is really the result of an immoral attitude towards others (e.g., an intention to cause the other pain) rather than an attitude towards the non-likeminded (e.g., ignorance of the possibility of pain in the other). In terms of empathy we might say that not only must whatever we can empathise with be of the right kind, but it must also *appear* to us as being sufficiently likeminded.

The case of adult human beings lends additional confirmation to the claim that the perception of other human beings as likeminded is less automatic than one may think and in fact requires considerable cultivation and socialisation. Sexism, racism, and nationalism are only the most obvious examples of the widespread lack of empathy in modern human society. In general, strangeness (unfamiliarity, otherness) constitutes the fundamental dimension along which humans tend to fail to understand each other in the broadest possible sense, including failing to perceive each other as being likeminded. It is the same basic principle that makes it even more difficult to perceive non-human animals as being likeminded.

Philosophies **2023**, 8, 89 11 of 21

3.1. Non-Human Animals

Of course, virtually no one would ever have denied that a dog or cat can see, hear, taste, feel, or smell. Many people also readily ascribe beliefs and knowledge to non-human animals as well as emotions such as anger, joy, or fear. But it is generally not well understood what it actually means to ascribe such psychological predicates to non-human animals. We noted earlier that what we call 'hope' in humans may seem to be essentially linguistic, given such characteristic instances as hoping she will be back in two days or hoping their marriage will be happy. Similarly, Wittgenstein writes: 'We say a dog is afraid his master will beat him; but not: he is afraid his master will beat him tomorrow. Why not?' (1953/2009 [26], 650). In response to this kind of problem, people will normally opt for one of two positions. The first is to take the seeming imperfection of some of the mental capacities that non-human animals appear to possess as evidence that they do not really possess any such capacity at all. The second is to take the same facts as evidence that non-human animals do possess the relevant mental capacities, albeit imperfectly.

Proponents of the first position would likely be accused by proponents of the second of being anthropocentrist (i.e., defining psychological terms purely on the basis of the human case), while proponents of the second would in turn be accused of being anthropomorphist (i.e., extending psychological terms uncritically from the human case to non-humans). Moreover, it can easily look as if we had no choice but to opt for one of the two positions. But that is a false dichotomy. Here we are actually encountering the kind of 'human chauvinism' that, as Haugeland said, is built into our very concepts. Even once we acknowledge that evolutionary history has not been a linear trajectory through stages of human development, our concepts will still be such that we will necessarily have to measure non-human mindedness against our human paradigm—or, as Wittgenstein put it, 'Only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious'.

What this kind of reflection on evolutionary history and our psychological concepts shows is that the seeming imperfection of some of the mental capacities that non-human animals appear to possess appears as an imperfection of these mental capacities only on the absurd assumption that non-human animals must match up to human forms of life (hope, fear, love, etc.)—as if non-human animals were somehow imperfect humans—when in reality the most that can be said is that, for better or worse, they more or less perfectly resemble those human forms of life. Thus, dog fear is not imperfect human fear, cat love is not imperfect human love, and so on. Rather, we describe dogs, cats, and other animals in psychological terms, because their behaviour resembles that of humans in relevant ways, albeit imperfectly, and our describing them in these terms has proved to some extent successful. So when we ascribe fear to a dog, for example, we are not saying that the dog is in qualitatively the same mental state as a human to whom, all else being equal, we might ascribe fear of the same thing. Nor are we saying that the dog is in a mental state that is qualitatively the same up to a certain level, as if human fear were the same as dog fear plus x. Rather, we are saying that the dog has a mental state that resembles human fear (which might therefore also correctly be called 'dog fear').

3.2. The Specificity of the Human Form of Life

In fact, this account of the relation between human and non-human psychology implies the specificity of all human forms of life, that is, the view that not only is human fear not dog fear plus x but that the same is true for all (known) non-human animal species and all human forms of life, including any ancestor species from which the human species has evolved and the most basic-seeming capacities (e.g., walking, eating, drinking, or playing, as well as attention, sensation, or memory). From the perspective of evolutionary biology, any such ancestor species may of course be regarded as an imperfect version of the human species. However, it does not follow from this that the human form of life or any part of it was the same as any prehuman form of life plus x. On Wittgenstein's conception of a form of life, a given ability, disposition, or capacity is to be understood in

Philosophies **2023**, *8*, 89 12 of 21

terms of a pattern of activity. Moreover, the answer to a question of the type 'What is this or that capacity (e.g., human vision)?' is to be given, according to the later Wittgenstein, in terms of such a pattern of activity. Not only are abilities, dispositions, capacities, etc. thus constituted by these patterns of activity, but so are any particular states, processes, acts, actions, activities, etc. that are their manifestations (exercises, actualisations), as well as anything that might go in between, such as children in early developmental stages or people with certain disabilities. Unlike developing or disabled humans, non-human animals (including ancestors of the human species) will not exhibit imperfect versions of human patterns of activity even in cases of identity in isolation, that is, even if on a given occasion the behaviour of a non-human animal perfectly resembles that of a human in all relevant respects. In other words, there could at most appear to be partial overlap between the two (human and non-human animal) patterns of activity, or forms of life, that in fact constitute the nature of any behaviour on a given occasion. In general, each form of life and all its parts are essentially constituted by their relation to the whole.³⁰

Wittgenstein's discussion of whether a dog can learn to simulate pain is a good illustration of this contextualist point:

250. Why can't a dog simulate pain? Is it too honest? Could one teach a dog to simulate pain? Perhaps it is possible to teach it to howl on particular occasions as if it were in pain, even when it isn't. But the right surroundings for this behaviour to be real simulation would still be missing.

Isolated from its surroundings, this dog's behaviour may perfectly resemble in all relevant respects that of a human simulating pain, so that there is an abstract level of description at which dog and human could be said to share a pattern of activity here. Yet, as Wittgenstein indicates, the larger patterns of activity that in fact surround and constitute the patterns of activity manifested on this kind of occasion are different in the case of dog and human respectively, so there could at most appear to be some overlap.³¹ The surroundings that Wittgenstein says would be required for the dog's behaviour to be 'real' simulation are of course ones that would sufficiently resemble those represented by the larger pattern of activity that is the human form of life of simulating pain. He reiterates the point in a more general register in the following passage of remarks:

583. ... Could someone have a feeling of ardent love or hope for one second – *no matter what* preceded or followed this second? — What is happening now has significance – in these surroundings. ... And the word "hope" refers to a phenomenon of human life. (A smiling mouth *smiles* only in a human face.)

584. Now suppose I sit in my room and hope that N.N. will come and bring me some money, and suppose one minute of this state could be isolated, cut out of its context; would what happened in it then not be hoping? – Think, for example, of the words which you may utter in this time. They are no longer part of this language. And in different surroundings the institution of money doesn't exist either. . . .

Here the isolated activity is stipulated as in fact being a part (specifically, a one-minute segment) of an activity that is a manifestation of the human form of life hope. Yet, as Wittgenstein indicates, the isolated activity would not have amounted to hope if the surroundings—that is, what happened before and after—had been significantly different from what they actually were; there would be no overlap between hope and any other form of life that may, counterfactually, surround the isolated pattern of activity. This is what Wittgenstein means by saying that a thing such as hope is 'a part of our natural history' (25) or that it is a human form of life: namely, he means that it is *essentially* a part of our natural history.

At least since Aristotle, it has been argued that there exists an essential difference between what it is to be minded in the way typical of a human and what it is to be minded in the way typical of a non-human animal.³² This difference has usually been

Philosophies **2023**, *8*, 89 13 of 21

expressed by saying that human beings are essentially rational animals, with Wittgenstein characteristically stressing the linguistic nature of human rationality.³³ There are indeed good reasons to believe that it is the fact that rationality (or language) permeates the human form of life which gives it the kind of distinctive unity within which even the most fundamental capacities (walking, eating, seeing, etc.) take on a specific nature such that no non-rational (non-speaking) animal can share any of them.³⁴ But I have not attempted to give a detailed defence of this kind of view. I have mainly tried to make intuitively plausible what I take to be the later Wittgenstein's conception of the specificity of the human form of life both as a whole and in all its parts, and hence of what it is to be minded in the way typical of a human being. I have thus tried to make plausible the view that although human mindedness and non-human animal mindedness are in many ways similar, human and non-human animals are nevertheless entirely non-likeminded, that is, there occurs no qualitative identity between any of their respective mental states. Moreover, it seems at least plausible that it is this same specificity of the human form of life that makes it so difficult to conceive that machines could acquire human mindedness. However, I shall now present an argument in favour of the view that machines nonetheless can acquire human mindedness.

3.3. An Alternative History

Imagine that in human history there had always been roughly equal numbers of humanoid robots and humans in society, but it somehow took thousands of years before any difference between these carbon-based and silicon-based beings was detected. Imagine, for example, that these robots had near-perfect simulations of human body functions including circulatory, digestive, and reproductive systems and that humans and humanoids had thus always mixed at all levels of society; perhaps humanoid bodies are even the same as humans' except for the central nervous system. Suppose, for the moment, that this sort of thing is at least conceivable—I will return to the issue of conceivability—and let us enquire into the likely consequences, in such an imagined scenario, following the discovery of the difference between humans and humanoids.

Should one group have inferred that the other was not really intelligent, sentient, or alive? Or, more generally, that the other was not really likeminded (i.e., that humans and humanoids did not, at least for the most part, have qualitatively identical mental states)? No. All else being equal, neither group would seem to have had a good reason to make such an inference. To see this, it is useful to first note an apparent difference between their concepts and ours. Suppose that at the time of the discovery they spoke modern English or some maximally close variant. Now the word 'human' in their language, at least until the time of the discovery, would seem not to have designated our concept *human* but an analogue concept *HUMAN* that applies to both carbon-based and silicon-based things (i.e., both humans and their robotic analogues). Similarly, their language would seem to have employed not our concepts *intelligence*, *sentience*, *life*, and so on, but instead analogue concepts *INTELLIGENCE*, *SENTIENCE*, *LIFE*, etc., each applying in equal measure to both kinds of HUMAN.

Of course, following the discovery that there are both carbon-based and silicon-based members of society, HUMANS would probably have begun to investigate differences between the two groups further. So they would soon have begun to discover significant differences, perhaps at first ones of a sociocultural sort such as (for example) that silicon HUMANS tend to live near urban centres, physiological ones such as that silicon HUMANS appear to come only in three skin tones or one blood type, or something like that. There are indefinitely many ways in which we can imagine this alternative history of the world to have led up to the time of the discovery, but for present purposes we may as well suppose that some almighty being had seen to it that the world took the requisite course up until the discovery. Similarly, there are indefinitely many ways in which we can imagine things playing out from the time of the discovery, notably including many ways in which HUMAN society continues as one in which humans and humanoids mingle as equals and

Philosophies **2023**, 8, 89 14 of 21

in harmony. However, even if some sort of struggle took place between the two kinds of HUMANS, neither group could possibly have a good reason to think that the other was not really HUMAN, INTELLIGENT, SENTIENT, ALIVE, etc., because these notions have evolved in HUMAN society precisely to refer to whatever it is that HUMANS share. For the same reason, there can be no question that HUMANS (i.e., carbon and silicon HUMANS alike) share HUMAN mindedness.

The question we should be asking instead is: are what it is to be minded in the way specific to carbon HUMANS and what it is to be minded in the way specific to silicon HUMANS the same or different? Insofar as, per our imagined scenario, carbon HUMANS and silicon HUMANS share a single form of life that encompasses all of carbon HUMAN life at least until the discovery, carbon HUMAN mindedness and silicon HUMAN mindedness will indeed appear to be the same. This is therefore the point at which technical questions concerning the supposed scenario must be addressed, specifically how scientifically plausible it is that carbon HUMANS and silicon HUMANS should share a single form of life. There are two main factors that decide what sort of scenario we are trying to imagine: the level of humanoid simulation and the level of human ignorance. It will be useful to suppose the level of human ignorance to be roughly equivalent to some time and place in actual human history (probably, any known major civilisation in human history could serve as a foil). A necessary condition on our scenario is that humans and humanoids are ignorant of their being differently constituted from each other. Therefore, we must at least suppose some such thing as that humans and humanoids have never seen any brains.

For example, this will be the sort of supposition required if we assume that the main difference between human bodies and humanoid ones is that humanoids do not have human brains. It is a curious thought that the rest of the body could be practically indistinguishable between human and humanoid. However, it is not scientifically implausible. Rather, what seems mainly unbelievable is that someone should actually want to create such a machine, which requires near-perfect simulations of all kinds of human body functions including perspiration, respiration, intoxication, sleep, growth, decay, and puberty. There are many other types of scenarios to consider, but this one lets us assume a fairly realistic level of human ignorance.

3.4. The Genealogy of Our Psychological Concepts

Some people think it easy to infer from the apparent conceivability of such a scenario that machines can acquire the sort of mindedness that is typical of human beings (i.e., human mindedness). But it is not that easy. The most obvious objection is that it is simply an illusion that the humanoids possess human mindedness: the humans in the scenario have lived under this illusion for so long that it is built into their concepts, and so these concepts are fundamentally preventing them from questioning the illusion, but from the outside we can see that those humanoids are just very advanced automata. This is essentially Searle's response, which contrasts this kind of machine with the case of non-human animals. Searle thinks we find it natural to ascribe mental states to non-human animals, because 'we can see that the beasts are made of similar stuff to ourselves' (1980 [2], 421). He explains: 'Given the coherence of the animal's behavior and the assumption of the same causal stuff underlying it, we assume both that the animal must have mental states underlying its behavior, and that the mental states must be produced by mechanisms made out of the stuff that is like our stuff' (421). We might make similar assumptions about a humanoid robot, he says, but once we found out, for example, that its head contained no human brain but something very different, we would give up the assumption that it had to have mental states and conclude, rightly in his view, that it was just 'an ingenious mechanical dummy' (421).

Searle is right about some of our likely reactions in such a case—after all, many people have wanted to agree with him on this point—but he is wrong about what inferences this kind of discovery really licenses. In my view, the natural reaction that Searle describes and recommends is by and large attributable to what I have described earlier as the sort of emotional bias in humans towards their own kind that culminates in bioism and the

Philosophies **2023**, 8, 89 15 of 21

biological objection (i.e., the claim that human mindedness requires being alive, so machines cannot have it). Now the way to see that this bias misleads us here is to realise how our psychological concepts are in fact the same as those in the alternative history recounted above, at least insofar as their genealogy is the same up until the time of the discovery that there are both carbon HUMANS and silicon HUMANS.

Up until that point, carbon HUMANS would have lived by the same assumption as actual humans, namely, that they live in a homogeneous society such that they would be surprised to find that some members have brains in their heads and others something very different. Of course, nowadays we know many things about the human brain, but there was a time when we knew next to nothing, and our psychological concepts in fact developed in virtual ignorance of the brain's existence; hence, the two sets of psychological concepts evolved in ways that are essentially identical. From this, I submit, it follows that even if they are not in fact the same, the concepts can at least be regarded as being identical for present purposes. Consequently, it would be wrong for us today to deny that a humanoid robot may possess human mindedness merely on the ground that they do not have a human brain. On the contrary, if we cannot detect any difference between humans and humanoids without opening their heads, and all other observable patterns of activity are the same, we will have every reason to believe that they share a single form of life, namely the human form of life, and are thus perfectly likeminded.

It may be objected to this argument that our concepts are really different, because our psychological concepts have actually always, at least in their primary function, only referred to beings with brains. This is probably true. Even so, such a difference in the extension of our concepts would in itself not prevent them from being applicable to other beings whose existence was previously unknown to us. Of course, someone who believes that our psychological concepts have only ever been applied to beings with brains may find it relatively difficult to apply them otherwise.

The objector may reply that modern neuroscience has since taught us that human mindedness is necessarily dependent on the human brain; moreover, that this has come to be reflected in our psychological concepts, as can be seen from Searle's and many others' intuitive judgement that human-like behaviour without a brain cannot constitute human mindedness. In response, I want to say three things. First, if human mindedness is the particular sort of mindedness typical of human beings (as we have defined it), then brain science will certainly not have taught us that this mindedness cannot be acquired by a nonhuman being, although it has taught us many important ways in which human mindedness is normally dependent on the brain. Second, the more plausible explanation of people's constant tendency to agree with Searle is what I have described as the emotional bias that culminates in bioism.³⁶ Third, whatever brain science has taught us has not had anything like the alleged effect on our psychological concepts. To see this, try the following thought experiment. Imagine our actual world suddenly turned out something like the possible world we have been considering. Imagine, for example, that your best friend's head pops open mid-conversation over dinner, and it is all wires inside, but your friend stays calm and very much their usual self, and begins to explain the situation to you. If you thus found out that your best friend had always been a machine, would you infer that your friend has never been truly conscious, intelligent, sentient, etc.? What if it turned out that many others, including many others close to you, had also always been machines? I believe Wittgenstein got this kind of case exactly right, when he wrote:

Can't I imagine that people around me are automata, lack consciousness, even though they behave in the same way as usual? ... Just try to hang on to this idea in the midst of your ordinary intercourse with others—in the street, say! Say to yourself, for example: "The children over there are mere automata; all their liveliness is mere automatism." And you will either find these words becoming quite empty; or you will produce in yourself some kind of uncanny feeling, or something of the sort. (Wittgenstein 1953/2009 [26], 420)

Philosophies **2023**, *8*, 89 16 of 21

The point is that we cannot imagine, or at least cannot sustain the thought, that others around us are mere automata. On the contrary, it is just as Fuchs said: 'We perceive others from the outset as embodied participants in a common form of life' (Fuchs 2022 [9], 10, my italics). In other words, we normally perceive others as being likeminded. Moreover, we do so regardless of whether the other has a brain inside their head or not (and, to this extent, regardless of whether they are in fact a machine or not). Anyone still in doubt about this may finally ask themselves what direct evidence they have that the people they know best, including themselves, have a brain inside their head. Thus, if today we had the chance to check what is inside everyone's heads, we would surely be surprised to find anything but brains; yet if we found that no one had a brain inside their head (or anywhere else), this in itself would have virtually no effect on our psychological concepts or our application of them (except in neuroscience, obviously).

It follows that our psychological concepts have in this respect remained the same as what they were a long time ago, and are thus in all relevant respects the same as those in the alternative history we have been considering, in which they clearly apply to humans and humanoids alike. And, so, it follows that machines can acquire human mindedness, that is, they can be intelligent, conscious, sentient, etc. in precisely the way that a human typically is all of these things.

4. Conclusions

In an attempt to provide a satisfactory, albeit (for reasons explained in Part 1) indirect answer to the biological objection—according to which human mindedness requires being alive, so machines cannot have it—I have tried to set the highest possible bar for what may reasonably be counted as acquiring human mindedness. In agreement with some of the most prominent defenders of the objection, namely Hubert Dreyfus and Thomas Fuchs, I have argued that it requires the sharing of the human form of life. Contrary to their position, I proceeded to show that machines can indeed share this form of life, and that this is so regardless of whether machines can be alive or not. To this end, I tried in a first step to make plausible the conception of a form of life that is at the heart of the later Wittgenstein's thought, which is also referenced by Dreyfus and Fuchs. According to this conception, a form of life is a form that is naturally abstracted from life (i.e., from something that is alive), yet its manifestation does not necessarily require something that is alive. In a second step, I argued that on this conception the human form of life turns out to be so very specific that no known non-human animal may share any part of it—which is what I think may just satisfy a desire to affirm some sort of human exceptionalism in some of my opponents—yet, I argued, it is possible nonetheless for a machine to share the human form of life and thus to acquire human mindedness.

There remain two important reasons why it might perhaps be thought that I set the bar for human mindedness too high or too low respectively, which I would like to address by way of conclusion.

First, it might be thought that I set the bar too low, because my argument for the view that machines can share the human form of life, and thus acquire human mindedness, primarily concerned our concept of mindedness rather than mindedness itself, but the conditions for conceptual possibility are likely less strict in this case than those for metaphysical possibility. My general response to this kind of objection proceeds along the same lines as the point I borrowed from John Haugeland in response to the charge of anthropocentrism, which was to say that our psychological concepts have naturally evolved to be human-centred but that, whilst more objective ones may perhaps be desirable, these are in an important sense the only concepts we have, and hence our necessary starting points. Analogously, insofar as our concept of mindedness, and thus indirectly our psychological concepts as a whole, are a necessary starting point for the kinds of fundamental questions I have undertaken to answer in this essay, it seems only right that they should receive the amount of critical attention I have given them. Of course, even if my account of the workings of our psychological concepts is right, our very concepts may still be flawed in

Philosophies **2023**, *8*, 89 17 of 21

some relevant way. This is the reason why I have sometimes qualified my claim as being that whatever human mindedness requires, machines can *probably* fit the bill. I do believe that our concepts have stood the test of time in all relevant respects, and that we currently do not have a better theory available that would indicate otherwise.³⁷ Of course, perhaps we will discover in the future that sharing the human form of life and, hence, human mindedness requires being alive, or that it requires the proper utilisation of a human brain; but, again, perhaps we will discover that the requisite sort of machine *can* be alive, and that they *can* utilise a human brain in the requisite way.

Second, it might be thought that I set the bar for human mindedness too high, because it seems to require a humanoid robot with near-perfect simulations of many all-too-human functions including perspiration, intoxication, sleep, and puberty as well as certain direct limitations (for example, on memory and on processing speeds) that computing technology could easily overcome, while any superhuman robot would seem to fall short. However, this is precisely what I intended. Fuchs is quite right to point out that the vast majority of computers do not really compute anything, in the sense of the human practice originally referred to as 'computing', which is being taught in schools using Arabic numerals and pen and paper.³⁸ One of the most important questions that humans are asking themselves today about the future of AI is how much a machine could be like them, especially in psychological terms. In response to this question, I have presented an argument to the conclusion that machines can indeed be perfectly likeminded when compared with their human counterparts—that is, they can be intelligent, conscious, sentient, etc. in precisely the way that a human typically is—because machines can share the human form of life and thus acquire human mindedness; more specifically, if we cannot detect any difference between humans and humanoids in society without, say, opening their heads, and all other observable patterns of activity are the same, we will have every reason to believe that they share a single form of life, namely the human form of life, and are thus perfectly likeminded (in other words, we will have just the same reasons to believe that there exists qualitative identity between the mental states of another and my own in the case of a humanoid as in that of a human).

Let me end on a slightly less theoretical note. It might still seem improbable that we are actually going to create these kinds of machines, and much more probable that we are going to create machines (e.g., care robots) manifesting artificial forms of life that may resemble the human form of life closely—much more closely, for example, than that of any known non-human animal—yet are clearly not identical to it. However, there are some fairly obvious long-term considerations, including human enhancement and, above all, the seemingly endless demand for technological innovation, that strongly suggest that the human form of life and humanoid forms of life are on converging trajectories.

Funding: This research was funded by the Humanities and Social Sciences Talent Fund, Peking University (grant number 7101302578); the Fundamental Research Funds for the Central Universities, Peking University (grant number 7100604441); and the Chinese Ministry of Education (grant number 22JJD720007).

Acknowledgments: Earlier versions of the material published here have been presented over the past two years at Kyoto University, National Taiwan University, Northeastern University London, Peking University, Shanghai Jiao Tong University, Shanxi University, Technische Universität Berlin, and the University of Tokyo. I would like to thank both the participants and the organisers of these events for many useful discussions.

Conflicts of Interest: The author declares no conflict of interest.

Notes

- See also Dreyfus 1967 [4] and 2007 [5].
- See especially Fuchs 2020 [6], 2021 [7], 2022 [8], and 2022 [9].

Philosophies **2023**, 8, 89 18 of 21

Fuchs actually writes: 'Might future humanoid robots not only simulate life but actually come alive?' (2022 [9], 10). It is a bit strange to say 'come alive' here, although it is clear enough what he means. Perhaps phrasing the question in this way already betrays Fuchs's view that machines cannot be alive.

- ⁴ See especially Fuchs 2021 [7], 35–41 and 2022 [9], 12–14.
- ⁵ See, for example, Luisi 1998 [10], Machery 2012 [11], and Mariscal & Doolittle 2020 [12].
- 6 See Varela & Maturana 1972 [13].
- See, for example, Bains 2004 [14], Peng 2015 [15], and Malaterre et al. 2022 [16]. A recent critical review concludes: 'Greater diversity of possible stable silicon molecules could be exploited by hypothetical sulfuric-acid-based life . . . Silicon could be widely used as a heteroatom component in sulfuric acid biochemistry' (Petkowski et al. 2020 [17], 24).
- See, for example, Kagan et al. 2022 [18], Kamm & Bashir 2014 [19], and Smirnova et al. 2023 [20].
- See also Turing 1950 [1], 443, 448, and 443, respectively, and, on Turing on racism in this connection, Sunday Grève & Yu 2023 [21], 24. See also my note 35 below.
- ¹⁰ See, for example, Calvo et al. 2017 [22], Tero et al. 2010 [23], and Trewavas 2014 [24].
- See, for example, Fuchs 2018 [25], 212, note 2 and 2022 [9], 10, note 8.
- References to Wittgenstein's posthumously published *Philosophical Investigations* [26] are to sections.
- To be clear, Wittgenstein does not claim greater fundamentality for either the biological or the sociological. See also Floyd 2016 [27] and unpublished [28].
- This is precisely the kind of statement and argumentation that one finds in authors advancing the biological objection. Once again, Fuchs's account is a useful example:

Intelligence in the true sense of the word is tied to insight, overview and self-awareness . . . And the prerequisite for consciousness, in turn, is . . . a living organism. *All experience is based on life*. The notion of an unconscious intelligence is a wooden iron. . . . As nonsensical as it would be to attribute to the clock a knowledge of time, it is equally nonsensical to attribute to an 'intelligent robot' an understanding of language or to an 'intelligent car' the perception of danger. (Fuchs 2022 [8], 257)

His starting point is the general view that (human) mindedness requires being alive and machines cannot be alive, i.e., the biological objection. From this he infers that a machine cannot think etc., specifically that it is a logical or conceptual mistake ('nonsensical') to try to suggest otherwise.

- The following two sections in the *Investigations* [26], 285 and 286, confirm this reading of the second paragraph of section 284. In particular, the objection Wittgenstein makes himself at the start of section 286—'But isn't it absurd to say of a *body* that it has pain?'—shows that the point he takes himself to have made previously is the kind of point I have said it is. See also the following passage in which Wittgenstein speaks explicitly, and (again) with approval, of a transition from quantity to quality: 'I think it is an important & remarkable fact, that a musical theme, if it is played <at> (very) different tempi, changes its *character*. Transition from quantity to quality' (Wittgenstein 2015– [30], MS 137, 72b(6); translated in his 1977/98 [31], 84). See also Peter Hacker's commentary on section 284, which notes: '"'From quantity to quality'": a Hegelian, and later a Marxist, term of art signifying an abrupt qualitative change resulting from a marginal quantitative change' (Hacker 2019 [32], 90). I am grateful to an anonymous reviewer for pressing me on this point.
- The replacement of the notion of a logical form with that of a form of life is the result of the sort of change Wittgenstein describes in the *Investigations* [26] as follows: 'The *preconception* of crystalline purity can only be removed by turning our whole inquiry around. (One might say: the inquiry must be turned around, but on the pivot of our real need.)' (108). Note that saying *form of life* is a key concept does not imply that the term 'form of life' is frequently used, which in fact it is not. But note as well that many of Wittgenstein's employments of the term, especially those in the *Investigations*, appear to be designed to lend it prominence and significance.
- See, for example, Wittgenstein 1953/2009 [26], 2, 5, 7, 19, 23, 25, 130, 244, 415, and 554 and 1953/2009 [33], ch. xii, sec. 365. See also Kuusela 2022 [34], 50–55 and Sunday Grève 2018 [35], 176–178. My own argument in Part 3 of the present essay is an example of this kind of method, which uses both real and imagined natural history.
- See, for example, Hacker 2015 [36], Hunter 1968 [37], and Moyal-Sharrock 2015 [38]. Cavell 1988 [39] sees Wittgenstein's position as itself being essentially ambivalent. For a useful review of the literature, see Boncompagni 2022 [40].
- As James Conant observes: 'The form here in question figures in Wittgenstein's conception of philosophy as a very abstract *logical* (or, as he later prefers to say, grammatical) category' (Conant 2020 [41], 646). Similarly, Joachim Schulte says: 'The word "form" is ... meant to allude to a certain ... shape, a pattern ... that life embodies' (Schulte 2010 [42], 138). See also Floyd 2016 [27] and 2018 [43], Laugier 2022 [44], Schulte & Majetschak 2022 [45], and Tejedor 2015 [46]. This kind of reading receives additional confirmation from Wittgenstein's occasional alternative use of the German expression 'Form des Lebens' to designate the same concept. For an instance of his using both German alternatives in the same context, see Wittgenstein 2015–[30], MS 115, 239(1).

Philosophies **2023**, 8, 89 19 of 21

As Schulte 2010 [42] (138–141) notes, forms of life including non-linguistic ones (more on which in the main text) may also, just like Wittgenstein's 'clear and simple language-games' (1953/2009 [26], 130), function as what Wittgenstein describes (in the same passage) as 'objects of comparison'.

- See also Schulte & Majetschak 2022 [45].
- Wittgenstein indicates, at the end of section 35 of the *Investigations* [26], that similar cases include recognising, wishing, and remembering and, in a remark inserted between sections 35 and 36, the case of 'when one *means* the words "*That* is blue" at one time as a statement about the object one is pointing at at another as an explanation of the word "blue". The *Investigations* are in fact filled with instances of Wittgenstein's struggling against this kind of tendency to give superficial, oversimplifying psychological explanations, including a good deal of his influential discussion of following a rule. This focus is also apparent from the way the text continues immediately after the passage I quoted from section 23, including the beginning of section 24: 'Someone who does not bear in mind the variety of language-games will perhaps be inclined to ask questions like: "What is a question?" Is it a way of stating that I do not know such-and-such, or that I wish the other person would tell me . . . ? Or is it a description of my mental state of uncertainty?'
- See also, for example, the following passage from *Philosophy of Psychology A Fragment*: "The words "It's on the tip of my tongue" are no more the expression of an experience than "Now I know how to go on!" We use them in *certain situations*, and they are surrounded by behaviour of a special kind, and also by some characteristic experiences. In particular, they are frequently followed by *finding* the word. (Ask yourself: "What would it be like if human beings *never* found the word that was on the tip of their tongue?")' (1953/2009 [33], ch. xi, sec. 300).
- My own argument in Part 3 proceeds along similar lines to demonstrate that the way we are presently applying certain psychological concepts is practically independent of certain physical inner processes.
- See also, for example, the following passage from *Philosophy of Psychology A Fragment*: "Grief" describes a pattern which recurs, with different variations, in the tapestry of life. If a man's bodily expression of sorrow and of joy alternated, say with the ticking of a clock, here we would not have the characteristic course of the pattern of sorrow or of the pattern of joy' (1953/2009 [33], ch. i, sec. 2).
- For Wittgenstein's discussion of concepts having a family-resemblance type structure, see his 1953/2009 [26], 65–77 and especially 67. See also my 2024 [47], 49–53.
- See also Wittgenstein's discussion of hope at 1953/2009 [33], ch. i, sec. 1 and his discussion of fear at 1953/2009 [26], 649–650.
- See also my note 14 above.
- ²⁹ See especially Wittgenstein 1953/2009 [26], 243.
- The precise nature of this relation to the whole is an interesting question, to which I briefly return a few paragraphs down in the text. As regards the task of defining the whole, i.e., a given form of life, I have already explained that a form of life is an abstract pattern of activity that is typically lifted (abstracted) from life and thus typically has an organic origin and may change over time.
- Wittgenstein seems right to think that his imagined case does not warrant the judgement that the dog's acquired ability is any kind of simulation at all. His beginning the section with the question 'Why can't a dog simulate pain?' perhaps shows that he further held the view, which I think is mistaken, that dogs cannot possibly learn behaviour that manifests a pattern of activity sufficiently similar to the human form of life of simulating pain so as to constitute a form of life that it would be correct to call 'dog simulation of pain'.
- For a useful systematic and historical account of this tradition, see especially Boyle 2012 [48] and 2016 [49].
- ³³ See, for example, Wittgenstein 1953/2009 [26], 25, 491, 493, and 647–650 and 1953/2009 [33], ch. i, sec. 1.
- See especially Boyle 2012 [48] and 2016 [49], Conant 2016 [50] and 2020 [41], and McDowell 1994 [51]. As Matthew Boyle puts it at one point: 'An account of *our* [human] sort of perceiving and desiring must itself refer to the role of these capacities in supporting a specifically rational form of life' (Boyle 2012 [48], 424).
- Turing was already aware of this problem, including its implications for the perception of machines and of their other capacities: 'Strawberries and cream ... Possibly a machine might be made to enjoy this delicious dish, but any attempt to make one do so would be idiotic. What is important about this disability is that it contributes to some of the other disabilities, e.g., to the difficulty of the same kind of friendliness occurring between man and machine as between white man and white man, or between black man and black man' (Turing 1950 [1], 448). See also my note 9 above. Compare, e.g., Fukuda 2020 [52].
- ³⁶ See also Wittgenstein 1953/2009 [26], 156–158.
- For an account of some of my reasons for believing that our concepts have evolved to be good ones, see my 2024 [47].
- ³⁸ See Fuchs 2021 [7], 31.

References

- 1. Turing, A.M. Computing machinery and intelligence. *Mind* **1950**, 49, 433–460. [CrossRef]
- 2. Searle, J.R. Minds, brains, and programs. Behav. Brain Sci. 1980, 3, 417–424. [CrossRef]
- 3. Dreyfus, H.L. What Computers Still Can't Do: A Critique of Artificial Reason (Revised Edition of What Computers Can't Do, 1972); The MIT Press: Cambridge, MA, USA, 1992.

Philosophies **2023**, *8*, 89 20 of 21

- 4. Dreyfus, H.L. Why computers must have bodies in order to be intelligent. Rev. Metaphys. 1967, 21, 13–32.
- 5. Dreyfus, H.L. Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philos. Psychol.* **2007**, 20, 247–268. [CrossRef]
- Fuchs, T. Verteidigung des Menschen: Grundfragen einer verkörperten Anthropologie; Suhrkamp: Berlin, Germany, 2020.
- 7. Fuchs, T. In Defense of the Human Being: Foundational Questions of An Embodied Anthropology; Oxford University Press: Oxford, UK, 2021.
- 8. Fuchs, T. Human and artificial intelligence: A critical comparison. In *Intelligence—Theories and Applications*; Holm-Hadulla, R.M., Funke, J., Wink, M., Eds.; Springer: Berlin, Germany, 2022; pp. 249–259.
- 9. Fuchs, T. Understanding Sophia? On human interaction with artificial agents. Phenomenol. Cogn. Sci. 2022, Online first. [CrossRef]
- 10. Luisi, P.L. About various definitions of life. Space Life Sci. 1998, 28, 613–622.
- 11. Machery, E. Why I stopped worrying about the definition of life... and why you should as well. *Synthese* **2011**, *185*, 145–164. [CrossRef]
- 12. Mariscal, C.; Doolittle, W.F. Life and life only: A radical alternative to life definitionism. Synthese 2018, 197, 2975–2989. [CrossRef]
- 13. Varela, F.J.; Maturana, H.R. De Máquinas y Seres Vivos: Una Teoría Sobre la Organización Biológica; Editorial Universitaria: Santiago, Chile, 1972.
- 14. Bains, W. Many chemistries could be used to build living systems. Astrobiology 2004, 4, 137–167. [CrossRef]
- 15. Peng, S. Silicon-based life in the Solar System. Silicon 2015, 7, 1–3. [CrossRef]
- 16. Malaterre, C.; Jeancolas, C.; Nghe, P. The origin of life: What is the question? Astrobiology 2022, 22, 851–862. [CrossRef] [PubMed]
- 17. Petkowski, J.J.; Bains, W.; Seager, S. On the potential of silicon as a building block for life. Life 2020, 10, 84. [CrossRef]
- 18. Kagan, B.J.; Kitchen, A.C.; Tran, N.T.; Habibollahi, F.; Khajehnejad, M.; Parker, B.J.; Bhat, A.; Rollo, B.; Razi, A.; Friston, K.J. In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron* **2022**, *110*, 3952–3969.e8. [CrossRef] [PubMed]
- 19. Kamm, R.D.; Bashir, R. Creating living cellular machines. Ann. Biomed. Eng. 2014, 42, 445–459. [CrossRef]
- Smirnova, L.; Caffo, B.S.; Gracias, D.H.; Huang, Q.; Pantoja, I.E.M.; Tang, B.; Zack, D.J.; Berlinicke, C.A.; Boyd, J.L.; Harris, T.D.; et al. Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish. *Front. Sci.* 2023, 1, 1017235. [CrossRef]
- 21. Sunday Grève, S.; Yu, X. Can machines be conscious? Philos. Now 2023, 155, 24–25.
- 22. Calvo, P.; Sahi, V.P.; Trewavas, A. Are plants sentient? Plant Cell Environ. 2017, 40, 2858–2869. [CrossRef]
- 23. Tero, A.; Takagi, S.; Saigusa, T.; Ito, K.; Bebber, D.P.; Fricker, M.D.; Yumiki, K.; Kobayashi, R.; Nakagaki, T. Rules for biologically inspired adaptive network design. *Science* **2010**, 327, 439–442. [CrossRef]
- 24. Trewavas, A. Plant Behaviour and Intelligence; Oxford University Press: Oxford, UK, 2014.
- 25. Fuchs, T. Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind; Oxford University Press: Oxford, UK, 2018.
- 26. Wittgenstein, L. *Philosophical Investigations*, 4th ed.; (1st ed., 1953); Anscombe, G.E.M., Hacker, P.M.S., Schulte, J., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 2009.
- 27. Floyd, J. Chains of life: Turing, *Lebensform*, and the emergence of Wittgenstein's later style. *Nord. Wittgenstein Rev.* **2016**, *5*, 7–89. [CrossRef]
- 28. Floyd, J. Lebensform vs. Kultur, aspect vs. technique: The emergence of Wittgenstein's mature style. Unpublished.
- 29. Haugeland, J. Artificial Intelligence: The Very Idea; The MIT Press: Cambridge, MA, USA, 1985.
- 30. Wittgenstein, L. Wittgenstein Source Bergen Nachlass Edition. In *Wittgenstein Source* (2009–); Wittgenstein Archives at the University of Bergen: Bergen, Norway, 2015.
- 31. Wittgenstein, L. *Culture and Value: A Selection from the Posthumous Remains*; rev. ed. (1st ed., 1977); Nyman, H., Pichler, A., von Wright, G.H., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 1998.
- 32. Hacker, P.M.S. Wittgenstein: Meaning and Mind (Part II: Exegesis §§243–427), 2nd, Extensively Revised ed.; Wiley-Blackwell: Hoboken, NJ, USA, 2019.
- 33. Wittgenstein, L. *Philosophy of Psychology A Fragment*. In *Philosophical Investigations*, 4th ed.; (1st ed., 1953); Anscombe, G.E.M., Hacker, P.M.S., Schulte, J., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 2009.
- 34. Kuusela, O. Wittgenstein on Logic and Philosophical Method; Cambridge University Press: Cambridge, UK, 2022.
- 35. Sunday Grève, S. Logic and philosophy of logic in Wittgenstein. Australas. J. Philos. 2018, 96, 168–182. [CrossRef]
- 36. Hacker, P.M.S. Forms of life. Nord. Wittgenstein Rev. 2015, 4, 1–20. [CrossRef]
- 37. Hunter, J.F.M. "Forms of Life" in Wittgenstein's Philosophical Investigations. Am. Philos. Q. 1968, 5, 233-243.
- 38. Moyal-Sharrock, D. Wittgenstein on forms of life, patterns of life, and ways of living. *Nord. Wittgenstein Rev.* **2015**, *4*, 21–42. [CrossRef]
- 39. Cavell, S. Declining decline: Wittgenstein as a philosopher of culture. *Inquiry* 1988, 31, 253–264. [CrossRef]
- 40. Boncompagni, A. Wittgenstein on Forms of Life; Cambridge University Press: Cambridge, UK, 2022.
- 41. Conant, J. Kant on the relation of a rational capacity to its acts. In *The Logical Alien*; Miguens, S., Ed.; Harvard University Press: Cambridge, MA, USA, 2020; pp. 574–647.
- 42. Schulte, J. Does the devil in hell have a form of life? In *Wittgenstein on Forms of Life and the Nature of Experience*; Marques, A., Venturinha, N., Eds.; Peter Lang: Bern, Switzerland, 2010; pp. 125–141.

Philosophies **2023**, 8, 89 21 of 21

43. Floyd, J. *Lebensformen*: Living logic. In *Language*, *Form*(*s*) *of Life*, *and Logic*; Martin, C., Ed.; De Gruyter: Berlin, Germany, 2018; pp. 59–92.

- 44. Laugier, S. La forme logique de la vie. Arch. de Philos. 2022, 85, 77–97. [CrossRef]
- 45. Schulte, J.; Majetschak, S. Lebensform. In Wittgenstein-Handbuch; Weiberg, A., Majetschak, S., Eds.; Metzler: Berlin, Germany, 2022; pp. 285–292.
- 46. Tejedor, C. Tractarian form as the precursor to forms of life. Nord. Wittgenstein Rev. 2015, 4, 83–109. [CrossRef]
- 47. Sunday Grève, S. Real names. In *Engaging Kripke with Wittgenstein*; Gustafsson, M., Kuusela, O., Mácha, J., Eds.; Routledge: New York, NY, USA, 2024; pp. 28–59.
- 48. Boyle, M. Essentially rational animals. In *Rethinking Epistemology*; Abel, G., Conant, J., Eds.; De Gruyter: Berlin, Germany, 2012; Volume 2, pp. 395–428.
- 49. Boyle, M. Additive theories of rationality: A critique. Eur. J. Philos. 2016, 24, 527–555. [CrossRef]
- 50. Conant, J. Why Kant is not a Kantian. Philos. Top. 2016, 44, 75–125. [CrossRef]
- 51. McDowell, J. Mind and World; Harvard University Press: Cambridge, MA, USA, 1994.
- 52. Fukuda, T. Cyborg and bionic systems: Signposting the future. Cyborg Bionic Syst. 2020, 1310389. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.