Using the idea of neural networks to argue against essentialism

Jorrit Jan Walinga

Geesteswetenschappen, Utrecht University

7-2-2021

Begeleider: Erik Stei

tweede beoordelaar: Colin Caret

7,5 CE

Abstract

This paper gives an argument against essentialism by assuming a computationalist and physicalist stance of the mind and then showing by using the analogy of neural networks how concepts cannot have the properties ascribed to essences nor can they be the result of any direct understanding of essences. The essentialist view is then shown to be a redundant and unlikely assumption. An alternative explanation is given for the phenomena essentialism attempts to describe, and some potential implications for the future of philosophy, the interpretation of AI and some contemporary political debates are noted.

Introduction

The question of this paper is “can essentialism be attacked using the idea of neural networks”. And a sub-goal is to use the resulting argument to form ideas that can help AI research. So the belief that is attacked in this paper is essentialism. This is the idea that things (or even events, higher abstractions, etc.) have an inherent property or properties which give them their identity. The idea is that an apple X is in the same category as apple Y because they share some essence. And if substance loses certain properties it can cease to have the identity it once had. It is as intuitive and widespread as it is wrong, which I will attempt to demonstrate. The opposing explanation, which I will defend, of the phenomena essentialism attempts to describe is this: we generalize stimuli, and/or our internal abstractions, to concepts/abstractions which we then project on our experienced world (grouping continuous sensory data in stable, discrete information) and then think they exist mind-independently. This is relevant for AI because essentialism is about categorization - namely the prescribing of supposedly objective categories, but it’s not necessarily about all categorizations - and since a lot of what AI does, and I am mostly talking about neural networks here, is fundamentally also about categorization there are certain points of contact. How we see the categorizing of AI can shape how we see our own categorizing, and vice versa.

The modes of thinking assumed in this work are computationalism and physicalism. The first is seeing the mind as the result, or rather the first-person perspective of, the information processing in the brain. Neurons exchange and store information chemically, and in how they are structured, which produces our conscious experience. Physicalism is the view that there is no substance other than what is known as the material world. These are both popular positions so I will not attempt to defend them here, nor is that this purpose of the paper. The point is that when taking these positions and combining them with what we know of the mind it follows that essentialism is wrong, which can be shown using neural networks to visualize this.

The strategy of attacking essentialism is as follows: I cannot prove a negative (non-essentialism), but by showing how the phenomena essentialism attempts to describe is best described (a more coherent theory with more predictive power) by a non-essentialist account essentialism can be shown to be redundant and unlikely, going by Occam's razor, until better proof is given. Some philosophers defending essentialism will also be discussed. It is important to be precise about in what way essentialism will be shown as redundant, because it is of course a metaphysical claim, not one of mental phenomena. The problem is that we cannot say anything about how the world actually is without interpreting it subjectively. And as I will argue, the type of interpretation itself that we do gives rise to a feeling of essentialism. That does not mean essentialism as a claim about the world is false, but it would be like seeing a rainbow and then, while fully knowing how it is an illusion, also claiming that there is in the real world a giant colourful arch present (or the ‘real rainbow’ in some other way) existing at the same moment and time as the illusion of a rainbow. Such a claim is not necessarily in contradiction with what we know of the world, but it clearly lacks evidence. This alternative explanation of the ‘illusion of essentialism’ is quite similar to some eastern philosophies like the Buddhist school of thought called Nagarjuna’s Madhyamaka, where the main arguments rest on the logical incoherence of various essentialist and realist ideas (regarding objects, causality, the self, etc.).[1]

The focus on computationalism specifically is important for the argument because by seeing how our concepts, so also the concepts regarding essences/identity, get formed, stored and used it is possible to question some fundamental assumptions in essentialist thinking.

[1] Garfield, Jay L. (translation and commentary). *The fundamental wisdom of the middle way: Nāgārjuna's Mūlamadhyamakakārikā*. Oxford University press. (1995)

Physicalism is used to deny certain metaphysical entities also assumed in essentialist thinking.

The core argument

First we will examine the claims of essentialism further. Because the idea is not only that things have essential properties but also that we observe those properties or have them a priori so that we know how to categorize that thing.

The idea that we have the concepts corresponding to essences a priori is first of all not in line with how we seem to learn concepts, namely based on experience and abstracting further from those concepts. But secondly there is no causal explanation for how we would get those objective a priori concepts of the right properties or essences in the first place. If certain brain structures that stand for those concepts are predetermined by the genes it is a priori but not objective, since these genes are merely selected in a somewhat random process (not without incentives for certain outcomes of course) which could easily have been different because their primary reason for existing is how well they help an organism survive and reproduce and not to perfectly mirror reality itself. A similar attack, using the idea of evolution to argue against realist notions, can be found in Nietzsches 'The gay science'[2] and in Donald Hoffman his work 'The case against reality'[3]. And there does not seem to be a way we necessarily need to form certain concepts because of the structure of existence itself.

So if our understanding of the essential properties or essences is not based on inherent categories in our mind, which we learn to connect to, then the only alternative - within the

[2] Friedrich Nietzsche (1882). De vrolijke wetenschap. Reprint and translation by Hans Driessen, Paul van Tongeren, Uitgeverij VanTilt, 2018. p. 141 - 145

[3] Donald Hoffman. The case against reality: why evolution hid the truth from our eyes. New York : W.W. Norton & Company (2019)

world we assumed in the beginning of this paper - is that we gain the understanding of the essential properties or essences through experience.

By seeing the mind as an information-processor we can argue against this option: Concepts are the first-person perspective side of a process that can be described as information-storage in the brain. How the concepts are formed is similar to how a neural network categorizes, which we will now use as an analogy, in that there is training data which only/mainly triggers a specific subset of the structure it feeds into. Neural networks are virtual nodes which feed information forward through connections from an input-layer (that which needs categorization) to an output-layer (the categories). Weights are assigned based on training data in an attempt to replicate to some extent how we would (want to) categorize the future data. There are of course many differences between our concept-formation and how a standard neural network categorizes (respectively: fluid vs static structure, complex vs simple node-behaviour, interconnected vs lineair categories, etc.) but the important point of comparison is that (as good as) continuous input gets mapped and generalized to discrete states. I describe experience as continuous and unorganized here because a third-person description of the sensory data alone does not include a justification for dividing and grouping it in the ways we do, an idea famously first brought to life by Hume. It has also been argued that both biological neural networks (the ones responsible for what Kahneman describes as system 1 reasoning) and artificial neural networks work by blindly finding a ‘direct fit’ to a complex space of input without needing more abstract, interpretable rules to guide the process.[4]

So here is the problem for gaining understanding of the essential properties through experience: If our ‘training data’ of (whatever we refer to as) thing X was slightly different then it must be the case that our concept of thing X is different. Because like a neural net our

[4] Uri Hasson,Samuel A. Nastase,Ariel Goldstein. (2020) Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Elsevier, Neuron 105.

categorizations depend on the data we get. And of course for each individual their concept of thing X has to be different since no two individuals can possibly have the same training data in practice. And although different training data might in theory lead to the same categorizations, because of the uncountably infinite number of possible ways for the creation and being of a concept this will in practice never be so. (There are infinite ways of placing the necessary parts of the neural substrates for concepts in space, or even as a network when disregarding space entirely there are an infinite amount of possibilities as well since context matters.) And more generally: every aspect of our categorizations depend on factors which are changeable and have no objective standard we can directly compare them to.

How do we decide which individual system is better at categorizing the 'real thing', knowing we are systems ourselves with our own subjective categorization? There is no point of view without an interpreting observer from which we can judge what systems are better at recognizing the 'real thing X'. There is only the intersubjective judgement of other systems and their labelling. We don't teach a neural network to categorize the world as it truly is, but we teach it how we ourselves find it useful to partition experience.

Levels of similarity

But there is more to the categorization than the social aspect. I'll explain why it still seems like our concepts generally have the same meaning:

Although all meaning is only located in the subjective mind, there are various stages of intersubjectivity/similarity to other minds. At the first level, based on how necessarily shared parts of the mind are, we have the limitations the shared world sets for the systems that 'produce' the mind. These systems, currently known only as 'brains', which are the third-person perspective descriptions of the processes we call 'minds' (and their subconscious), exist in the same sort of world, which limits how they can function in the first place.

And, in the next level, there are ‘attractor points’ for how minds will be likely to think. If we assume a type of computation will be used that utilizes concepts, or an equivalent to it, then we can say some manners of concept-creation and combining of the (systems of) concepts will prove more useful and give more evolutionary benefit than others. And concepts themselves are already likely since it is for example very useful to generalize similar information to the same kind of states so that the same kind of action plans can be used, then experience can be used for the future.

The next level in the causes of similarities is that of the specific genes we share. These genes roughly cause the same development of neurons and how they interact with each other. In this layer we can find the heuristics for concept-creation like ‘the principles of grouping’ Gestalt psychologists have studied (the biases for seeing stimuli as part of the same object), although these heuristics might also be a consequence of more basic rules almost inevitably leading to these more complex rules when exposed to the kind of world we live in. This is a large difference between our biological concept-creation and the categorization of neural networks, yet these genomic biases are very likely to exist in humans because of our relatively rapid learning for specific tasks.[5]

With the succeeding level we cross the divide between a priori and a posteriori when considering the impact of shared environments. This distinction is not absolute and there is always influence of the lower levels to the higher, as Tuomos also suggests in his description of a priori reasoning causing the formation of a posteriori knowledge.[6] When two systems already roughly share the same manners of concept-creation then it becomes much more

[5] Zador, A.M. *A critique of pure learning and what artificial neural networks can learn from animal brains*. Nat Commun 10, 3770 (2019)

[6] Tahko, Tuomas. (2011). A Priori and A Posteriori: A Bootstrapping Relationship. Metaphysica.

likely that the same stimuli from environments result in similar concepts formed from them, when those stimuli are also nested in a similar context.

After that we find the effects of the cultural/social sphere. In every interaction we, through context, gain understanding of how others have attached meaning, what (systems of) concepts, to common labels (like words or objects) and then decide, often subconsciously, whether to adjust our own meaning to theirs or if we 'should' convince them of our meaning.

There is a noteworthy difference between 'simple concepts' and 'abstract concepts'. The former is the set of concepts formed from what we call 'objects' and their properties, and the concepts that are created because of the configurations and relations between them. The latter are the concepts made from other concepts (and other inner states). Since abstract concepts have no direct stimuli which can help in creating shared meaning they diverge quite quickly and more extremely with each extra layer of abstraction. This divergence in meaning is of course the reason this paper was necessary in the first place.

We can use these 'levels of similarity' to explain the family resemblances between our ideas, we can explain the amount of resemblance and we can give justifications for adopting one (system of) concept(s) over the others. As an example of that last point: When we want to argue that 'alcohol' should be included in the concept of 'drugs' we can appeal to the bias for concept-creation that drives us to see members that share certain basic properties as the same (here the property of inducing an altered state of mind). Opponents can appeal to the bias of basing a concept purely on the associations (caused by its real-world usage). If neither bias is preferred then the consequences of preferring one bias over the other can also be evaluated.

Continuing the core argument

So if there is no objective (meaning necessarily shared) understanding of what we call ‘things’, and other groups of categories we imagine to be outside of the mind, before or after learning, then what actually is it that we have understanding of? It is always only of our subjective concepts, of ‘things’ in this case, which we see as existing in our sensory world (which is imagined as outside the mind) as permanent objects because the stimuli we receive consistently triggers the same concepts.

The only refuge for essentialists is now claiming that next to our world of subjective categories there is a world of the ‘real categories’ which just happens to (sometimes) run parallel, but these ‘real categories’ have no influence on our understanding of the world nor do they impact us in any other way, so why assume them? The assumption itself cannot even come from the ‘real categories’ but only as an abstraction of our subjective categories which we imagine to be independently existing. Furthermore, the idea of an essence binding substance together into a thing which is part of a certain category is contrary to physicalism because nowhere in our models of the world in physics is anything like an essence to be found. ‘Essences’ seem to be the result of built-in category mistakes; we cannot help, which becomes apparent by our use of language, to see these mental phenomena (the concepts and their similarities) as physical phenomena (things and their essential properties/essences). As if seeing the split between them needs constant effort. There is still the possibility that there still are objective, discrete categories but since they are not necessary for a description of the world and there is no causal relationship between them and our mind as of yet this assumption can be dismissed.

And to be absolutely clear: if we only have concepts to use when reasoning about what is or isn’t essential to (for example) things then we have neither mind-independent standards nor anything actually essential to all. Since these concepts are mere generalizations of stimuli the boundaries are arbitrary, yet they feel ‘solid’ because we lack a meta-understanding of our own concepts; we don’t know how our concepts could be different.

Identifying and replacing essentialist thinking

To show the improved explanatory power and coherence of this alternative to essentialism let’s examine the classic thought experiment ‘the ship of theseus’: A ship begins to deteriorate with the years, so it is replaced piece by piece. Is it the same ship after having everything replaced as it was before this process?

An essentialist might suggest that it is the structure, a continuity across time, the substance it’s made out of, etc. that is/carries the essential property for the ships identity, but in all cases what is being described is why the sensory information is or isn’t linked to the same concept across time. There is no objective rule to be found which states how this sensory information ought to be categorized, there are only the biases for specific manners of computation inside of us which determine the answers given. It would be very strange to postulate that existence itself prescribes one bias over the other here. There is no plausible causal explanation for that whatsoever.

An answer we can give with the new non-essentialist perspective to the question if it is the same ship or not could simply be “depends on who you’re asking” or more generally “depends on what system and its method of categorization we use”, and then we can either move on to more productive questions or ask why we might prefer one way of categorizing over the other.

Then there is the issue of ‘natural kinds’ which this version of non-essentialism needs to have an answer to. The idea is that arbitrary boundaries between concepts referring to man-made (like a clock or a car) or man-dependent (like colours, since they depend on how we see light) things might be granted but surely the fundamental constituents of existence itself (like elemental particles) are not arbitrarily defined/interpreted.

To this I would answer that natural kinds are also mind-dependent, existing only as concepts, but there is something about our world that causes us to be able to form concepts that trigger extremely consistently. This consistency in our experience is what causes us to see something as independent, not anything we can get from the world directly. The problem with essentialism in general is that it is a description of how we refer to the world, not what we actually refer to (as this is not directly knowable). And there is still nothing like the concept itself, as a discrete identifier, in the world itself as far as we can know with our models. The models cannot be said to be anything but useful abstractions.

And this problem for natural kinds, our discrete categories not necessarily having discrete equivalents in the world, underlines an assumption of a lot of metaphysical thought I have been implicitly attacking until now and this is also relevant for essentialism itself: The assumption that how we think is in any way directly related to how the (rest of the) world itself is. I would like to attack this assumption but there usually isn’t any reasoning given, so I will have to make do with quickly explaining my anti-realist position regarding this.

That there is this split in interpretation and how the world itself is, and that the former does not necessarily neatly map onto the latter, flows naturally from again seeing the mind as an information processor. In the pre-modern perception of the world where we were seen as units of being, souls, with no internal processing other than our consciousness directly observing the world it could be thought that our observing was an unmediated consequence of how the world itself was.

But when considering how that sensory information is actually transported and transformed to our experience this assumption doesn't hold, we can shortly consider the visual processing but it is equivalent in principle to all other senses: When light hits the retina and activates receptor cells those cells already combine and divide activation inside the eye itself (to points, changes, colours), this information is transported along the optic nerve where it is again transformed (combining the input from both eyes), where in the visual cortex points get made into lines, then into moving lines, into more complex shapes, etc. etc. all the way up to (presumably sub-symbolically represented) neural substrates of concepts. With the help of these information-transformations these concepts can trigger consistently even when the sensory input coming from what we perceive as the same object differs immensely. Examples like visual illusions, and altered states like deliriums, Charles Bonnet syndrome and the psychedelic experience also show that at least during those events we observe our mental model of reality, not reality itself (well, that is the easiest explanation if we want to maintain people share a common reality). The idea that we only experience what our brains can construct, in its effort to match bottom-up information with top-down predictions, is also the central idea in the widely influential theory of predictive processing.[7]

Needless to say that our interpretation of the world does not neatly nor necessarily correspond to how the world itself is, because these information transformation processes - which again have no objective standard to compare their functioning to, usefulness alone is their reason for existing - assign activation to abstractions which are never directly found in experience or our scientific models of the world. We build a mental model which values

[7] Clark, Andy. *Surfing uncertainty: prediction, action and the embodied mind.* Oxford University Press (2016).

coherency and predictive power, and from there the correspondence is automatically assumed.

An example of a philosopher who assumes this direct metaphysical access and concludes essentialism from it is David Oderberg. In ‘Real essentialism’ he starts his defense of essentialism with “It is a metaphysical truth that the world contains both unity and plurality. There is a multiplicity of things and they all have features in common. In one sense, everything in the world is united to everything else, …”.[8] Why does Oderberg assume here that claims about his experiential world are translatable one to one to claims about the world itself? This is clearly an explanation of (the abstractions of) how he experiences reality. As shown, the intuition of direct correspondence does not seem to be coherent with what we now know about how the brain functions, and this intuition is not backed up in any way. And all conclusions Oderberg derives from this assumption afterwards cease to have believable support. An explanation for how the subjective experience of essences implies the objective experience of essences is lacking.

Putnam on the other hand does address the problem of a potential internal notion of meaning, with his thought experiment, and gives an argument for meaning being external.[9] A twin earth is imagined which is identical except for water not being H2O but XYZ (just assuming that something like that would be physically possible), and the word ‘water’ refers to XYZ on the twin earth. Putnam reasons that on both worlds the mental state accompanying ‘water’ is the same, while its meaning is different. And when the microstructure is examined someone from the original earth, when visiting the twin earth, would say they mistook XYZ for water, while the meaning ‘should’ not be the same.

What Putnam describes is not ‘meaning’ in a metaphysical sense directly affecting how we think when it is discovered that the microstructure of XYZ is different from H2O,

[8] Oderberg, David. *Real essentialism.* Routledge (2007). p. 44

[9] Putnam, Hilary. (1973). *Meaning and Reference.* The Journal of Philosophy, 70(19), 699-711.

but rather something in the world causing stimuli (consistently) in such a way that we change our concepts. A sort of hierarchy in meaning appears in our mind, we cannot go back to the previous way of thinking where H2O and XYZ were seen as equivalent, but that does not mean one interpretation of the world grasps the true meaning. We just tend to see mental models that produce ones from which there is no going back as inferior to the new models. When the measuring devices used to see what sort of microstructure 'water' has on the twin earth are actually affected, only on the twin earth, in such a way that they see H2O as XYZ then we would again say (intuitively speaking) that the earlier misconception did grasp the true meaning all along. But notice that the only thing we can truly say, when we are not working from the "objectively right viewpoint" one assumes in a thought experiment but when imagining the world from the perspective of the imagined inhabitants, is that one way of interpreting the world is replaced by one that at that moment(!) feels more coherent (with what we already know and the interpretation of incoming information). And the "objectively right viewpoint" used in these kinds of thought experiments only gives the illusion that one kind of interpretation must always prevail over others because the description of the world, what the whole world is in the thought experiment, does exist out of concepts (which can actually be equal to the concepts of the imaginined inhabitants). The actual world does not and a view of the world without subjective interpretation does not exist. And to add to that, there can be no referent found for 'meaning' in our models for the world outside of the mind.

With both Oderberg and Putnam there seems to be some kind of (intuitive, automatic) category mistake happening, where we imagine that the world itself is made out of concepts (which we then call essences, the (parts of) the 'real identity') and thereby trying to see our subjective interpretation as universal. But when looking at how information has to be processed we can identify only subjective concepts in the cases where we feel as though there have to be essences.

A final example

Before discussing the further implications let’s go over the whole argument with a specific example: When we observe what we call an apple, how do we know how to categorize it? If there is an essence, or essential properties, which make the apple objectively part of that category then we would either know that essence a priori or through experience. There is no known method by which we know categories a priori and it seems more likely that our genes merely provide the basis for more general concept-learning and not concepts themselves. But if genes did contain information on what concepts to form then these concepts are not directly informed by the essences of things but the genes responsible have had to prove themselves in evolution by their usefulness and not necessarily how they create concepts that perfectly represent the world.

When we go by the idea that we have learned to categorize the apple through experience we can liken this process to how a neural network categorizes input. Training data, usually labelled in real-time in the case of humans or labelled beforehand and in one package in the case of neural networks, is used to make a system activate a distinct part (a category/concept). In this way a sort of general rule for information processing leads certain activation consistently to the same states.

First of all nothing about this process causes the exact same concepts to be formed in practice for humans; there are differences in the training data (regarding the apple itself, no experience of an apple in any way is exactly the same as other experiences of apples), differences in the context (the environment and related concepts), differences in how neurons react (and are thereby likely to connect and have certain ‘weights’), and one can even imagine how slight changes in brain chemistry caused by a different diet can impact concept-formation. The point is: no two concepts are exactly the same. Our intuition tells us they are, the computational view of the mind informs us they are not.

So unlike the essence of an apple which is supposed to be unchanging and mind-independent, our conception of it is subjective and changeable. And this conception is created through a continuous stream of experience which we only afterwards divide into concepts. There is at least no direct connection to the essence/essential properties of the apple which informs us how to categorize it. And experience itself lacks the prescriptive element of essences. (This is an important part because it shows how essentialist assumptions of stimuli being caused by pre-arranged groups having an objective identity is an unnecessary addition, because we do not get information about how stimuli has to be grouped from the stimuli itself.)

Now a parallel world of essences next to our concepts can be imagined, which does either not influence how we categorize our experience at all or has some indirect influence on experience through the substance of the apple.

But where does such an essence exist and how does it impact experience? Or what decides what essential properties of the apple are? First of all these essentialist ideas seem redundant, because the consistency and similarity between our categorizations can already be explained with a detailed account of subjective concept-creation alone. Secondly they assume more than our best scientific models of the world can account for by postulating something metaphysical which makes some properties both mind-independent and essential and/or by imagining a metaphysical essence which binds substance into a thing and makes it part of a category. And thirdly they provide no causal explanation for the phenomena they describe.

The more coherent explanation is that the apple is a concept which is experienced as existing independent of the mind, and because the concept is activated by sources we identify as multiple objects we see the similarities between them also as independently existing (as the essences/essential properties). This can be generalized to all concepts, since the principles for their formation necessary for the argument is the same for all.

Implications and uses for AI

An effect this non-essentialist view has on artificial intelligence is that human judgements, as in the manner of categorizing the world, are not seen as necessarily more important/useful than those of artificial intelligence because the source and validity of the judgements is roughly the same, with both not seeing the ‘actual state of the world’ and only make models that fit the data well. And arguments against strong AI that use intentionality as the defining difference cannot be made, because the idea of intentionalism (the ‘aboutness’ of thought, that thought has content because it refers to something) is in contradiction with the non-essentialist view. For a thought to be about an apple, referring to something other than the thought itself, there needs to be an independently existing category of ‘apple’. This has been shown to be unlikely.

But more importantly being non-essentialist about the origin of our concepts gives the freedom to examine our categorisations without chasing after a metaphysical source that in all likelihood will not be found. Because, again, although essentialism is a metaphysical theory it is also implying a prescriptive mode of thinking regarding our concepts. That they should correspond to the ‘actual identities’ of objects. So not focussing on what our concepts ought to be, assuming a prescriptive and objective standard (the essences), but rather working descriptively in how we come to form them computationally. This in turn can help us in creating AI that creates categories similarly to humans.

A central problem of AI is that, because the computations simulate the continuous computation of system 1 (using the terminology of Kahneman[10]) and not the discrete and interpretable computation of system 2, its decisions are not necessarily explainable and it is often treated as a 'black box'. But using the earlier discussed 'levels of similarity', which can of course be further examined and expanded upon (as it stands now it is a theoretical starting point, not a final guideline), it is possible to incorporate these to a degree in the structure and training of AI. And as these levels of similarity help make categories between humans similar and interpretable it will also do so for AI, potentially even letting the AI itself connect shared labels (i.e. words) to its own automatically generated concepts.

In practice this means for example creating the causes for 'principles of grouping' in object-recognition neural networks. When a neural net now learns to detect pictures of cats it might use information in the picture which is not directly part of what we would see as the object 'cat'. All training examples could have (if we assume colour is used) a light blue, green or brown background (belonging to perhaps popular environments cats are photographed in) which is then seen as evidence by the neural net, as a subset of the structure (which we might call a 'property') consistently activates because of these sort of stimuli. It might become even so dependent on very unintuitive factors that a single pixel can cause an undesirable prediction.[11] We ourselves of course do not see the background of a cat as part of its identity, because we first group stimuli together as objects. When an artificial neural network would do so in the same manner we do then the properties, the sub-categories the concept of cat exists out of, would have a higher likelihood of being similar to the properties we would ascribe to the concept of 'cat'. If this is done with enough aspects of categorization then AI will no longer be a black box but largely explainable in our language.

[10] Kahneman, Daniel. *Thinking fast and slow.* Penguin books (2012) p. 19-97

[11] J. Su, D. V. Vargas and K. Sakurai, "One Pixel Attack for Fooling Deep Neural Networks," in *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828-841 (Oct. 2019)

Philosophical implications

This view also does imply that we have to give up on some ideas in philosophy, while it opens up others, when it is fully realized and accepted. Both negative and positive consequences stem from our concepts losing their immediate prescriptive nature.

The first negative-seeming consequences are that it becomes much harder to form a believable defense for or against the use of certain meanings of words. “This is what this word must mean to be in line with how reality is.” is, at first glance, a much more convincing idea than “this is how we might want to use the meaning of this word, because it is more coherent with X and satisfies bias for meaning Y.” Another effect of adopting the non-essentialist stance is that the contemporary debates around gender, life, species, etc. would take drastic turns. To name some examples: Currently the most common attempts to say whether we should or shouldn’t consider transgender people as he gender of the opposite sex go back to “being a man/woman is essentially having property X or Y and we should change our meaning to that”. With a non-essentialist perspective the debate would center around arguments discussing the use of adopting one meaning over the other. Since we can now see that the current essentialist arguments are nothing more than people trying to universalize and prescribe their subjective meaning, and making no progress if the rhetorical part of the argument fails to work, I see the non-essentialist approach as more fruitful.

For the second example we can look at how life is defined in the debate around abortion. The most common argument I have encountered from both sides is that a fetus is or isn’t alive/fully human in the sense that it would be eligible for applying the maxim “killing (something which is alive/human) is wrong”. Aside from the moral objectivist position which can be denied this argument already presupposes the meaning of words as prescriptive for our behaviour instead of the mere tools we can begin to see them as from a non-essentialist perspective. Asking why we would assign a label and a specific meaning to that label that would prevent us from removing the fetus is where the debate can go. Valid reasons can be given to both sides but at least there is now a dialogue.

A third example is the essentialist views surrounding species. A regular retort against the teaching of evolution is that it is never observed that one species changes into the next or that there are always missing links between humans and early hominids. This again requires viewing what a species is in an essentialist manner and not allowing for the gradual change evolution brings.Only our biases and learned prototypes determine our interpretation of an organism belonging to one species or the other.

Because essentialism seems to be a hardwired tendency these sorts of examples of clashes of ideas with no way forward can be found everywhere, and by giving up a feeling of being objectively right, a certain accessibility of the ideas (essentialism is easier to grasp than non-essentialism) and a persuasive element of the arguments, progress can be finally made in these areas.

Adopting non-essentialism also impacts epistemology. How it does this exactly, what the (perspectivist) alternatives would be and what the further implications of those are is beyond the scope of this paper, but the quick version is: The correspondence theory of truth, which also seems to be the standard intuition and comes ‘pre-packaged’ with essentialism, rests on the idea that there is a way things are objectively that can be similar to how we think. And if our thoughts/sentences map to how the world really is, concepts and their connections corresponding to real things and their relations, then the thought/sentence is objectively true. The version of non-essentialism defended in this paper denies there are objective things and properties altogether, and thus there is no way for our sentences/thoughts to directly correspond to the rest of reality. An internalist version of truth, and possibly even a perspectivist version, because a ‘platonic blueprint’/objective standard for concepts is missing, among other reasons, has to be adopted because what is actually going on is sentences/thoughts being coherent with our subjective model of the world. Despite it being some steps removed from the core objective of the paper I feel the need to mention perspectivism, because as Dummett notes an internalist version of truth, as required by such an anti-realist framework I use here, will have to deal with explaining how and/or to what extent meaning is shared and how to give justification for it.[12] The ‘levels of similarity’ I have used is a first step, but the kind of perspectivism I think it leads to is often attacked by stating that it is incoherent because to claim it is true goes against the central idea of not allowing objective truth. To preemptively defend the philosophical path non-essentialism might lead to I will give a short answer to this (as it is by far the most popular argument against perspectivism): This argument against perspectivism is the result of a failure in imagining the full implications of a perspectivist framework; a perspectivist would not claim their view is objectively true but only in their mental model of the world. And a justification can be given for it more as a rhetorical device which is more or less likely to convince based on what parts of that mental model and the logical structures for evaluating new additions is shared. I will not give a full defence and explanatory framework for perspectivism, but I hope to have shown that the avenues the argument of the paper can lead to are not dead ends. Also notice how I still have to use language which implies essentialism and the correspondence theory of truth, but I use language in this way as a shorthand for the more coherent framework underneath. An interesting area of further research is if and how language can be used in a way which does not imply essentialism, and if this is even practical.

[12] Michael Dummett. The Seas of Language. Oxford: Clarendon Press (1996). p. 476 - 477

Conclusion

Philosophy has generally taken the approach of using introspection to make claims about how we think. In the process of reconciling philosophy with the increasing knowledge of the workings of the brain, and using a physicalist viewpoint which is more in line with modern science, there are certain implications for where epistemology and metaphysics can lead. I have shown how when considering our concepts as ‘results’ of information transformation and storage processes, with the help of comparing the neural substrates responsible for concepts to neural networks, that there are conclusions that can be drawn regarding essentialism. Namely that it seems increasingly implausible as an explanation for our categorizations. And abandoning this common intuition which is at the core of much of our thought has far reaching but potentially extremely useful consequences. The argument and alternatives to our incoherent intuitions presented in this paper can help fix the confusions essentialism brings and can open up debates that appear to be stranded because of the unnecessary conflicts rooted in the falsely assumed universal prescriptivity of language, and non-essentialist thinking can help with opening up the research that can make artificial intelligence more understandable and therefore more widely applicable and easier to use.

References

Clark, Andy. *Surfing uncertainty: prediction, action and the embodied mind.* Oxford University Press (2016).

Dummett, Michael. *The Seas of Language.* Oxford: Clarendon Press (1996).

Garfield, Jay L. (translation and commentary). *The fundamental wisdom of the middle way: Nāgārjuna's Mūlamadhyamakakārikā*. Oxford University press. (1995)

Hoffman, Donald. *The case against reality: why evolution hid the truth from our eyes.* New York : W.W. Norton & Company (2019)

Kahneman, Daniel. *Thinking fast and slow.* Penguin books (2012)

Nietzsche, Friedrich (1882). *De vrolijke wetenschap.* Reprint and translation by Hans Driessen, Paul van Tongeren, Uitgeverij VanTilt (2018)

Oderberg, David. Real essentialism. Routledge (2007).

Putnam, Hilary. (1973). *Meaning and Reference.* The Journal of Philosophy, 70(19), 699-711.

Tahko, Tuomas. (2011). *A Priori and A Posteriori: A Bootstrapping Relationship.* Metaphysica. 12. 10.1007/s12133-011-0083-5.

Uri Hasson,Samuel A. Nastase,Ariel Goldstein. (2020) *Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks.* Elsevier, Neuron 105. Accessed via: https://doi.org/10.1016/j.neuron.2019.12.002

J. Su, D. V. Vargas and K. Sakurai, "One Pixel Attack for Fooling Deep Neural Networks," in *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828-841 (Oct. 2019) doi: 10.1109/TEVC.2019.2890858.

Zador, A.M. *A critique of pure learning and what artificial neural networks can learn from animal brains*. Nat Commun 10, 3770 (2019). https://doi.org/10.1038/s41467-019-11786-6