# Responding to the Watson-Sterkenburg debate on clustering algorithms and natural kinds

In *Philosophy and Technology* 36, David Watson discusses the epistemological and metaphysical implications of unsupervised machine learning (ML) algorithms.[1] Watson is sympathetic to the epistemological comparison of unsupervised clustering, abstraction and generative algorithms to human cognition and sceptical about ML's mechanisms having ontological implications. His epistemological commitments are that we learn to identify "natural kinds through clustering algorithms" (Watson, 2023a, p. 6), "essential properties via abstraction algorithms" (Ibid., p. 12), and "unrealized possibilities via generative models" "or something very much like them." (Ibid., p. 6; p. 12; p. 16). The same issue contains a commentary on Watson's paper in which Tom Sterkenburg fiercely opposes the epistemological claim that clustering algorithms can identify natural kinds. Sterkenburg argues there's nothing about clustering arguments *themselves* that enables natural kind identification (Sterkenburg, 2023, p. 3). But Watson was entitled to respond: he held that universally as well as existentially quantified readings of Sterkenburg's counterclaim leave his original epistemological claim unaffected (Watson, 2023b, p. 2). What's at stake in the Watson-Sterkenburg debate is whether it is possible for clustering algorithms to identify natural kinds and whether they underly human identification of natural kinds. Following the tendency of artificial intelligence to trump human cognitive capacities, the ethical significance of the debate is that if Watson's claim is true, then, optimised algorithmic natural kind identification will probably become superior to human natural kind identification, i.e. we will have reasons to defer to algorithms to partition the world for us.

My contribution to the Watson-Sterkenburg debate is twofold. First, I argue Sterkenburg's criticism of Watson's claim is too severe because it denies *any* clustering algorithm to identify *any* natural kind. Secondly, I argue Watson's reply overestimates both clustering algorithms' and humans' access to *natural* kinds and his claim has to be restricted accordingly to 'We identify ~~natural~~ *kinds* via clustering algorithms, or something very much like them'.

Section (1) reconstructs Watson's discussion of unsupervised learning clustering algorithms. Subsequently, (2) Sterkenburg's commentary on Watson and Watson's response are expounded.

[1] A conceptual primer on ML is in order. ML can be defined very broadly as algorithms that self-improve based on a performance measure, without the outcome that corresponds to the goal of the task related to the performance measure being programmed into the algorithm itself (El Naga & Murphy, 2015, p. 5). Central to ML is that instances (data points that are part of a dataset) are repeatedly inputted into the algorithm, after which the algorithm iteratively improves on itself so as to be able to reproduce the desired output with respect the inputted instance, which are often quantitively encoded into vectors. This computational process of self-improvement is termed *learning* and the computational processing of a single input is often called an *experience* or an *observation.* Learning proceeds by means of a mechanism of adjusting all points where the algorithm performs a computation so that the results of subsequent computations better match the desired output based on the quantified distance between the desired output and the actually computed output. In feedforward neural network architectures this is called 'backpropagation'. If the algorithm learns long enough, its accuracy increases and the training of the algorithm is completed. Now the algorithm should be able to generalise well (hasn't overfitted) when presented with novel instances of data that weren't part of the dataset used for training (Ibid.).

Lastly, (3) I criticise both Sterkenburg and Watson, consider objections to my arguments and (4) conclude.

## 1. Unsupervised learning and the epistemological clustering claim

Watson addresses the dearth of philosophically scrutinous treatments of unsupervised learning as compared to the abundance of work on supervised/reinforcement learning. Roughly, he espouses a functionalist philosophy of computation and cognition that deems implementation in computer or brain irrelevant and hence marries the different types of unsupervised learning to human cognition achieving similar goals (Watson, 2023a, p. 12).

Before turning to clustering algorithms, I explain the difference between supervised and unsupervised learning. In supervised learning the training data are instances labelled with classifications (output space) that represent the algorithm's desired output for the example. The intuition behind *un*supervised learning is that data examples are not labelled and the algorithm itself has to find the classificatory lines (El Naga & Murphy, 2015, p. 5; p. 8). Watson specifies that for an unsupervised learning task, the training data matrix $\mathbf{X}$, containing to-be-observed samples, represents a certain number to-be-discerned features of a target system. To learn the task with respect to the target system, the algorithm takes an observation vector $\boldsymbol{x}$ out of $\mathbf{X}$ that represents a point in the data's feature space. Processing vector after vector, the algorithm discerns the probabilistic structure $\boldsymbol{P}$ present in the data domain (Watson, 2023a, p. 3).

A *clustering algorithm* discerns subgroupings in a dataset of whatever kind. Unsupervised clustering algorithms have to find categories to divide $\mathbf{X}$ based on a sequence of observations $\boldsymbol{x_1}\ldots\boldsymbol{x_n}$. As examples, Watson discusses '*k*-means' and 'hierarchical clustering'. In short, *k*-means algorithms are given a number *k* as quantification of the desired partition of $\mathbf{X}$, which Watson calls a 'prototype'. The algorithm statistically approximates a mean, "a hypothetical datapoint, a sort of Platonic ideal against which all others are compared" (Watson, 2023a, p. 4), from $\boldsymbol{x_1}\ldots\boldsymbol{x_n}$, by minimising a distance measure to the nearest centre relative to cluster variance. It determines statistical regularities shared by observed examples at the level of partition exemplified by the chosen value for *k*.

Hierarchical clustering differently divides $\mathbf{X}$ into subgroups, as the algorithm does not take *k* as an input, but instead "recursively partition[s] the remaining samples as *k* increases" (Watson, 2023a, p. 5). Simplified, it partitions $\mathbf{X}$ in two at $k$=1 and at $k$=2 those two categories are split into proper subsets of the initial two and so on for increasing *k*'s. The *optimal* partition, in terms of the granularity needed for the feature space, is determined via additional heuristics (Watson, 2023a, pp. 5-6).

As the Watson-Sterkenburg debate only focuses on clustering, I refer to the footnotes for discussion of *abstraction*[2] and *generation* algorithms.[3]

Watson argues that clustering algorithms' data partitioning entails an epistemological claim with respect to natural kinds, namely that "We learn to identify natural kinds via clustering algorithms, or something  very much like them" (**EC**) (Watson, 2023a, p. 6). A natural kind is "a grouping that reflects the structure of the natural world rather than the interests and actions of human beings" (Bird & Tobin, 2023) that "should permit inductive inferences and participate in natural laws" (Watson, 2023a, p. 6). $H_2O$ is such a distinct type of entity existing in the natural world as identified objectively by scientific procedures. Watson is a pragmatic with respect to natural kinds: clustering can succeed at identifying them at "varying levels of abstraction" where "[e]ach solution is the right answer to a different question" (Watson, 2023a, p. 8). He dismisses the ontological claim that 'ideally clustering algorithms have to find natural kinds', because this presupposes that natural kinds are methodologically reducible to effective computations. As such, they would all have to be successfully computable by an algorithm that terminates (algorithms that do not terminate cannot compute anything, as they remain circular (Turing, 1936, p. 233)). Watson recognises no a priori

[2] An unsupervised algorithm can be said to engage in abstraction when it learns simplified representations of **X** such that these represent a higher level pattern that is present in the underlying data in which **X**'s essential properties are 'embedded' (Watson, 2023a, pp. 9-10). An example of this is the neural network autoencoder. The general idea behind an encoder model is that an input is compressed into a vector embedding of certain dimensions. This is the latent informational space that represents the pattern, or essential properties, of the input. In a standard architecture, this latent informational space is given to a decoder model which translates the latent information into an output of a certain modality that contains or is constrained by the essential properties captured by the informational vector (see Roy, 2020).

[3] The generative category of algorithms are tasked with creating synthetic images, sounds, text-strings etc. of a quality indistinguishable from human-made images, sounds and text (Watson, 2023a, p. 13). Generative models come in different forms, but all are trained to "learn a probability distribution over the space of features that make up a possible world. In so doing, they delimit a horizon of possibility that constitutes a form of knowledge unto itself" (Watson, 2023a, p. 16). What Watson means to say with this phrasing is that generative unsupervised algorithms build a very complex statistical model gleaned from training data through which an input of a certain modality is rendered into an output of the desired quality. Encoder-decoder models can also be used generatively and then they are called 'autoregressive encoders'. But another paradigmatic example of a generative algorithm is a *generative adversarial network*, or 'GAN' for short. Explained in an implementation-agnostic fashion, a GAN is a type of neural network that combines a *generator*, a neural network that generates candidate outputs, and a *discriminator,* a neural network that evaluates the output of the generator. The output the generator produces is a 'fake' of a certain modality that is based upon 'noise', a random statistical distribution. This is paired with an observation ***x*** from the training dataset and subsequently the discriminator has to decide which image is the 'real' one. The game-theoretic relation between the discriminator and the generator is one where the generator has to fool the discriminator. The GAN goes through cycles of presenting produced outputs until the discriminator is fooled by the generator. Technically, the likelihood of the discriminator being incorrect is maximized and this is used for performing gradient descent (a form of backpropagation) on the generator. The generator henceforth performs better at fooling the discriminator and has learnt to generate outputs of a quality on par with the 'real' when presented with novel observations (Brownlee, 2019).

reason for this,[4] but claims they do "reduce murky debates about identity and essence to formal procedures for grouping elements together based on precise notions of similarity and difference" (Watson, 2023a, p. 7). Watson analogises clustering to human mental and linguistic partitioning of the world and based on functionalist presuppositions he contends the underlying processes are the same or at least very much alike. Specifically, with respect to language "[d]ividing one's perceptual field into a collection of things with names that persist under some range of conditions was a key step in the development of complex concepts for individuals and collaborative communication for groups" (Watson, 2023a, p. 8).

To recapitulate, Watson's **EC** holds (a) natural kinds can, pragmatically speaking, be discerned via clustering algorithms and (b) human identification of natural kinds likely proceeds via clustering functionally equivalent to unsupervised clustering algorithms, because the workings of human thought and language resemble them.

## 2. Sterkenburg on Watson and Watson on Sterkenburg

Sterkenburg[5] argues there are two problems with **EC**'s core premise: "*clustering algorithms identify natural kinds*" (**EC***) (Sterkenburg, 2023, p. 2). Firstly, for clustering algorithms to work, they require beforehand specification of a distance metric: a rule for how to compare the quantified difference between data instances. The "real work", according to Sterkenburg, thus is already done *before* the clustering algorithm does anything, which makes **EC*** trivial because the 'right' *k* or heuristic to identify the natural kinds in the underlying data is already given to it (Ibid.). Additionally, further choices to make "them suitable for identifying natural kinds" (Sterkenburg, 2023, p. 4) are needed. Although agreeing with Watson's natural kind pragmatics, Sterkenburg thinks Watson doesn't adequately dissolve this methodological vacuity. Secondly, he claims clustering algorithms lack *success criteria*. He backs this by citing a mathematical proof that ideal clustering algorithms are impossible[6] and statest it's logically unclear what distinguishes a clustering algorithm from any arbitrary function that partitions a numerical space (Ibid.).

In short, Sterkenburg thinks **EC*** is vacuous because (i) when clustering algorithms would identify natural kinds it is because of the right *external* configurations and (ii) the concept of clustering algorithm is mathematically opaque and doesn't admit of ideal solutions or success criteria.

However, Watson seems unfazed by this critique. Firstly, he evaluates Sterkenburg's **EC***, universally and existentially quantified. On the universal quantification he deems it "bogus" because clustering algorithms often end up with all kinds of farfetched, "baroque" categories and concludes it's implausible to suppose that *all* clustering algorithms identify natural kinds (Watson, 2023b, pp. 1-2). On the other hand, Watson says Sterkenburg denies **EC+** (*some* clustering algorithms identify

[4] Here Watson is oblivious to the arguments made by proponents of the physical variant of the Church-Turing Thesis, which holds that all of reality is in principle effectively computable. See (Piccinini, 2007) for discussion of this claim.

[5] While Sterkenburg disagrees with the absence of philosophical discussion of unsupervised learning that Watson purports, he agrees on the fact that philosophical research on unsupervised learning is mostly concerned with their output models rather than with the philosophical significance of the "underlying learning mechanism" (Sterkenburg, 2023, p. 1).

[6] Watson himself already references the impossibility result that Sterkenburg's second argument relies on (Watson, 2023a, p. 4).

natural kinds) by saying that the "real work" of determining distance and evaluation of the resulting clusters is external to the algorithm itself (Ibid.),[7] which leads to the dilemma that we *either* have to expand the notion of what an algorithm itself is by including "ex-ante and post-hoc modelling choices" *or* we have to exclude "auxiliary" steps that define what measure to minimise (this figures in supervised learning as well) from it. Watson concludes, contra Sterkenburg, that *both* have to be treated as part of the algorithms, because otherwise the concept of algorithm itself becomes vacuous (Watson, 2023b, p. 3). Watson's originally intended to say that "clustering algorithms (or something very much like them) are an *essential component* of any effort to identify natural kinds" (Ibid.), which remains true under both minimal and maximal interpretations of the scope of the algorithm. Secondly, Watson rejects Sterkenburg's lack-of-success-criteria objection because (i) the statistical theory of unsupervised learning remains underdeveloped and (ii) Sterkenburg overlooks the general concepts that do define the hypothesised clusters, namely stability and generalisation (Watson, 2023b, p. 4).[8]

To sum up, Sterkenburg criticised **EC** for its vacuity because of its reliance on external configurations and lack of robust criteria for identifying natural kinds (and succeeding in general). Watson retorted to Sterkenburg's **EC*/EC+** that the external configurations are part of the algorithm, otherwise that concept itself becomes empty; and Sterkenburg's lack-of-success-criteria objection is premature. Watson leaves us with a restated, but supposedly unscathed, **EC**: clustering algorithms are an *essential component* to any identification of natural kinds.

## 3. The explanatory and instrumental value of clustering algorithms and the Wittgensteinian approximation of natural kinds

In this section, I contribute to the Watson-Sterkenburg debate by arguing (1) against Sterkenburg that his criticism is too severe and (2) against Watson that his reply still overestimates the possibility that clustering algorithms can be compared to human thought and language with respect to the identification of *natural* kinds.

(1) *Defending the scientific and explanatory value of clustering algorithms.* Sterkenburg undermined the theoretical basis for **EC***, but this misses the point of comparing clustering algorithms to human natural kind identification. His criticism is too severe because it denies *any* clustering algorithm to identify *any* natural kind. Watson conflates human cognitive partitioning and scientific partitioning of the world into natural kinds and Sterkenburg does not call him out for this. Pressing this distinction however, the reason that Sterkenburg's conclusion is too strong becomes that under the pragmatic realism about natural kinds he agrees upon with Watson, it is an *empirical* question what

[7] In Watson's words: "The interesting question is whether we can simultaneously accept that natural kinds exist and deny that any clustering algorithm could ever identify one" (Watson, 2023b, p. 2).

[8] Here Watson refers to section 5 of his original paper, where he writes that "According to Mayo, our belief in some hypothesis *h* is justified only to the extent that *h* has passed severe tests, i.e., tests that should detect flaws or discrepancies from *h* with high probability. Practitioners rarely bother to subject unsupervised learning models to the same scrutiny as their supervised or reinforcement learning counterparts, as standards for such tests are not well developed. While familiar ML notions such as "loss" or "regret" may not apply in these settings, alternative desiderata such as stability and generalization do. More importantly, they are testable. Resampling procedures like bootstrapping and subsampling provide readymade tools for evaluating the robustness of clusters and abstractions" (Watson, 2023a, p. 2).

gets to be determined as a natural kind, rather than a property of an algorithm. The charge he levels that 'there is no ideal clustering function to do the identification work' is hence targeted at the wrong domain, because even if the functions are nonideal, they can still facilitate the inductive processes involved in both human cognitive partitioning and scientific partitioning of approximating the contours of natural kinds that do not involve the logical proofs distinctive of statistics. Watson himself already problematised the ontological claim that clustering algorithms are to identify natural kinds under ideal circumstances, because of scepticism of those ideal circumstances ever obtaining in practice. So while Sterkenburg is right to question the validity of the claim that clustering algorithms identify natural kinds, this fails at targeting Watson's real claim and overlooks the application of clustering algorithms to the nonideal identification of natural kinds that figures human cognitive partitioning and scientific partitioning of the world.

It can be objected that forsaking the requirement of mathematically ideal solutions for identifying natural kinds is throwing out the baby with the bathwater and inductive generalisation and stability simply do not have the same logical functions as mathematical proofs have. This is to say that turning the algorithmic problem away from ideal solutions/clear success criteria is no viable escape route, because this trivialises the entire endeavour of including technical research into algorithms in the philosophical debate. I respond that this is indeed an undesirable move from a mathematical perspective, as the mathematical perspective demands to follow proofs. However, it is not the lack of ideal solutions that is doubted here, but rather the view that because of such a lack, clustering algorithms completely lose their (i) *explanatory value* for human cognitive identification of natural kinds and (ii) *instrumental value* for scientific endeavours that seek to approximate natural kinds. If duly recognised, the logical limitations prohibit admitting of clustering algorithms implying any metaphysical conclusions (which, to recollect, Watson's rejection of the ontological claim was already conscious of) but does not hamper stressing their success and utility for explaining cognitive and scientific identification of natural kinds.

(2) *Restricting* ***EC*** *to ~~natural~~ kinds.* Watson was right to question the exclusion of algorithm configurations as 'external' to its workings, but he wasn't able to parry the vacuity charge. Firstly, because his revised **EC** is still rigged by the clause 'or something very much like them'. The clause makes **EC** indeterminate by virtue of allowing *anything* involved in the human cognitive process that resembles clustering algorithms to make the claim true. This is nontrivial, but too wide.

Secondly, he is overly optimistic that we identify *natural* kinds via clustering algorithms; this is problematic because it presupposes human cognition neatly partitions reality into things with names for language to latch onto. Taking a Wittgensteinian perspective on natural kinds however urges one to be sceptical of the conflation of (a) *how* language is used with (b) the robust *mapping* of some words on observable essences in reality. Steffen Borge argues that the Wittgensteinian view of language use is at odds with the standard view of natural kinds as words that neatly correspond to a fixed category of entities in reality. The latter doesn't follow from the former, because it may be so that the utility underlying the use of a word in a linguistic community is not determined by the non-linguistic features of the thing the word can be used to refer to (Borge, N.D., p. 105). To name an example, the word 'water' did not come to have its standard use because people naturally 'cognitively clustered' it to be $H_2O$. Rather, people wield the word 'water' to refer to a variety of fluids that often contain a lot more than just $H_2O$. Even though this might seem as a silly example that comes about by Watson's conflation of cognitive/scientific partitioning, Watson's comparative picture is complicated by it, because it shows that the standing criteria for something being a noncontroversial natural kind are too high *not only* for unsupervised learning algorithms, *but also* for non-scientifically organised human cognition. The Wittgensteinian view that the way words are used in everyday language latches onto a complicated net of overlapping conceptual similarities and dissimilarities (Wittgenstein, 1953/2009, §67) does give us reason to suppose that human language and thought relate to observable patterns in the world. But accepting this, it then becomes unreasonable to say that *if* clustering is involved in this, it can involve identifying *natural* kinds. Only the weaker claim follows that human cognition identified *real patterns*, or, to put it differently, 'just' *kinds*. So for it to be realistic, leaving the functionalist premise conditionally accepted, Watson's **EC** has to be restricted to **EC-**: 'We identify ~~natural~~ *kinds* via clustering algorithms, or something very much like them'.

However, we can object to this counterframing on equally Wittgensteinian grounds, namely by employing the Joscha Bach's framing of deep learning as an essentially Wittgensteinian endeavour:

> In the *Philosophical Investigations* [Wittgenstein] basically expresses his despair about dealing with imagery. In the original *Tractatus* [...] he hopes that he can define operations over the pictures, but eventually he figured out that there was no way in which grammatical language was able to generate pictures from patterns, at least not anywhere that he could see. And this was the same problem of symbolic AI and this is the problem that deep learning is solving. It deals with the messiness of reality with unprincipled reasoning, that is able to give you results even if the result is not logically exactly true and it does this by approximating arbitrary functions and when you give it patterns, these functions are able to predict these patterns and find regularities in them and this works so well that you can use these pattern matching algorithms, these function discovery algorithms/function approximators on grammatical language and you find better solutions than you do with analysis with Chomsky grammars. (Bach, 2022)

On Bach's view, deep learning approaches like unsupervised learning clustering algorithms are actually deeply in line with the later Wittgenstein's view on human language use, due to the fact they function as *approximators* rather than as logical picture producers.

If I do not want to reject the Wittgenstenian affiliation, the bullet my argument has to bite is that unsupervised clustering algorithms hit at the core of how human cognition and language identify natural kinds, as they do so via processes of experiential approximation, which is emulated by clustering algorithms. Now the grounding that enabled my argument to distinguish scientific or ontological *natural* kinds from the 'real pattern' *kinds* identified by human cognition is taken away.

But I need not lose hope. One can counter this undesirable conclusion by interpreting this Wittgensteinian deep learning discourse differently, namely by seeing it as an existential proof that the centralisation of experiential approximation empties the adjective 'natural' in the phrase 'human identification of natural kinds' of all its meaning. This happens because if we accept that human cognition uses functional equivalents of clustering algorithms to structure linguistic reference to things in the world (in accordance with Bach's Wittgensteinian framing of deep learning), then it follows that all this amounts to is the *human construction of a net of linguistic identification via approximation over the world*; a net that is so deeply socially and experientially constructed through cognitive approximation, that there is no reason left to connect it logically to *natural* kinds independent of scientific/ontological observers.

So after this roundabout, we are still left **EC-** that it might well be that human cognition works along the lines of unsupervised clustering algorithms or something very much like them, but that the clusters human cognition ends up with can hardly be logically connected to *natural* kinds in any scientifically or metaphysically meaningful sense of the term.

## 4. Conclusion

In this essay, I surveyed the Watson-Sterkenburg debate on unsupervised learning algorithms identifying natural kinds and their relation to human language.

After reconstructing their theses and disagreements, I argued that Sterkenburg's criticism missed its target by employing a mathematical framing of the possibility of the identification of natural kinds. I considered the objection to this dismissal that this throws out the baby with the bathwater, to which I responded that from a formal mathematics perspective this may well be so, but that this should not stop investigations into (i) unsupervised clustering algorithms' explanatory value for how humans cognition identifies (natural) kinds and (ii) its instrumental value for scientific research into the approximating of (natural) kinds.

Against Watson I argued that even after his rebuttal of Sterkenburg's criticism, the comparison his epistemological clustering claim makes between algorithms and human cognition remained to strong and that it should be restricted to kinds rather than to *natural* kinds. I considered the objection by way of Bach that it follows from Wittgenstein's insights into the workings of language that deep learning approaches capture the workings of human cognition with respect to the identification of natural kinds especially well. My retort was that if this is true, then the adjective *natural* has become entirely vacuous in its intended sense of 'objective' and 'scientifically determinable'. Instead, I bent it around to support my claim that human cognition doesn't identify natural kinds at all, but rather cuts the world along experientially approximated lines that reflect the contingencies of human experience, rather than the dotted lines of noumenal prototypes. This conclusion is of implicatory importance for the ongoing debate on the ethics of comparisons between human cognition and algorithmic functions, as versions of Watson's claim could have indicated that algorithmic perception could in principle partition the world better than humans can.

## 5. Bibliography

Bach, J. (2022, November 2). *From Language to Consciousness (Guest: Joscha Bach).* [Video]. YouTube. https://www.youtube.com/watch?v=ApHnqHfFWBk.

Bird, A. and Tobin, E. (2023). Natural Kinds. *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), Zalta, E. N. & Nodelman, U. (eds.). https://plato.stanford.edu/archives/spr2023/entries/natural-kinds/.

Borge, S. (N.D.). Wittgenstein and Natural Kind Terms. 103-108. https://www.academia.edu/10244050/Wittgenstein_and_Natural_Kind_Terms.

Brownlee, J. (2019, July 19). *A Gentle Introduction to Generative Adversarial Networks (GANs).* Machine Learning Mastery. https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/.

El Naqa, I. and Murphy, M.J. (2015). What Is Machine Learning? In: El Naqa, I., Li, R., Murphy, M. (eds.) *Machine Learning in Radiation Oncology*. Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1.

Piccinini, G. (2007). Computationalism, The Church–Turing Thesis, and the Church–Turing Fallacy. *Synthese* 154: 97-120. https://doi.org/10.1007/s11229-005-0194-z.

Roy, A. (2020, December 12). *Introduction To Autoencoders*. Towards Data Science. https://towardsdatascience.com/introduction-to-autoencoders-7a47cf4ef14b.

Sterkenburg, T. F. (2023). Commentary on David Watson, "On the Philosophy of Unsupervised Learning," *Philosophy & Technology*. *Philosophy & Technology* 36: 63. https://doi.org/10.1007/s13347-023-00663-2.

Turing, A. M. (1936). On Computable Numbers, with an Application to the *Entscheidungsproblem*. *The London Mathematical Society*, 2-42: 230-265. https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf.

Watson, D. S. (2023a). On the Philosophy of Unsupervised Learning. *Philosophy & Technology* 36: 28. https://doi.org/10.1007/s13347-023-00635-6.

Watson, D. S. (2023b). Reply to Tom Sterkenburg's Commentary. *Philosophy & Technology* 36: 69. https://doi.org/10.1007/s13347-023-00674-z.

Wittgenstein, L. (2009). *Philosophical Investigations*. (Anscombe, G. E. M., Hacker, P. M. S. and Schulte, J., Trans.). Wiley–Blackwell. (Original work published in 1953).