Machine learning and essentialism

Kristina Šekrst
Sandro Skansi
University of Zagreb, Croatia

Abstract

Machine learning and essentialism have been connected in the past by various researchers, in order to state that the main paradigm in machine learning processes is equivalent to choosing the "essential" attributes for the machine to search for. Our goal in this paper is to show that there are connections between machine learning and essentialism, but only for some kinds of machine learning, and often not including deep learning methods. Similarity-based approaches, more connected to the overall prototype theory, spanning from psychology and linguistics, seem more suited for pattern recognition and complex deep-learning issues, while for classification problems, mostly for unsupervised learning, essentialism seems like the best choice. In order to illustrate the difference better, we will connect both paths to their sources in other disciplines and see how human psychology influences our decision in machine-learning modeling as well. This leads to a philosophically very interesting consequence: even in the setting of supervised machine learning, essences are not present in data, but in targets, which in turn means that the categories which purport to be essences are in fact human-made, and hand-coded in the targets. The success of machine learning, therefore, does not give any substantial evidence for the independent existence of essential

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 171–196

properties. Our stance here is to state that neither the existence nor the lack of "essential" properties in machine learning can lead to metaphysical, i.e., ontological claims.

Keywords

essentialism, machine learning, accidental properties, similarity-based approach, pattern recognition, modal necessity.

Essential and accidental properties: introduction

The purpose of this paper is to show that the existence of essential-like features in machine learning, or the lack of them, cannot provide an ontological commitment.[1] Researchers have connected machine-learning practices with essentialist and anti-essentialist stances, but we feel that such claims ignore that both "essentialist" and "anti-essentialist" paradigms in machine learning are both influenced by human psychology and have no real consequence on the verification of whether there *are* essential properties in nature or not.[2]

The outline of the paper is as follows. First, we will give a brief overview of what philosophical essentialism is and mention the scarce research on (anti-)essentialism in machine learning. Next, we will provide insight into the basics of machine-learning paradigms, namely supervised and unsupervised learning. The notion of essential properties is often connected to supervised learning, but we would like to

[1] The authors would like to thank the anonymous reviewers for their detailed analyses and insights.

[2] It is necessary to distinguish between ontological commitments regarding nature and ontological commitments that are necessary in every AI system (we call this difference an ontological gap). The former are the subject of this article, and the latter were analyzed by Krzanowski & Polak (2022a; 2022b).

find out why, so we will connect it to psychological essentialism and the development of human epistemological stances. We will notice that prototype theory seems closer to human understanding, and it can be seen as present in both supervised and unsupervised paradigms, even though they might be a better fit for the latter. The notion of a *feature vector*, as a collection of properties, will be connected to psychological prototypes. Last, we will observe how in both supervised and unsupervised learning, the human factor involved guides the learning, but this cannot be equated with the existence or non-existence of essentialism. Namely, some supervised tasks are better for some real-life or mathematical problems, while some unsupervised tasks are better for others. The question of essentialist-like or anti-essentialist paradigm here is just a question of using the right tool for your problem, and not an ontological consequence.

Philosophical essentialism

An *essential* property of an object is a property that an object must have, while an *accidental* one is the one the object happens could have, but that it could lack. That is, in modal terms,[3] we are talking about *necessity* and *possibility*,[4] respectively (Robertson Ishii and Atkins,

[3] Standard modal characterizations have been disseminated with the works of Ruth Barcan Marcus and Saul Kripke. Kripke's work on semantics is taking the truth of a formula relative to a possible world, since its truth value depends on what is true in accessible world. See Barcan Marcus (1993) for a modality synthesis and Kripke (1972) for Kripke semantics.

[4] There are, of course, differences between logical and (meta)physical possibilities. Something might not be a logical contradiction, but still be (meta)physically impossible, i.e., not conforming to the laws of nature, for example, a man travelling faster than the speed of light. The exact details of such classifications, especially between physical and metaphysical possibilities, are a matter of debate.

2020): an object is going to possess the essential property in all possible worlds, but for an accidental one, there is a possible world in which an object lacks such a property. *Essentialism* is a standpoint in which (at least) some objects have (at least some) essential properties (Robertson Ishii and Atkins, 2020). For example, an essential property of Socrates is to have originated from his parents but is not essential for him to have brown hair. An essential property of a dog is certainly not brown hair since there are dogs of other colors. In philosophy, some essential properties are not a matter of much debate. For example, a dog had to have some biological origin. *Canis canis* is also a dog. But in order to pursue the matter further, there might be various objections to candidates for essential purposes.[5] Often, a dog is considered a "wolf-like descendant", where various breeds might not conform to this ideal, along with the notion of "having an upturning tail".

The former is the reason why there are various kinds of essentialism in philosophy. Standard Aristotelian essentialism also deals with necessities, and in his categories, he was researching properties that all the members of the category have in common, without which, they cannot be members of that category.[6] One of the most famous criticisms comes from Wittgenstein,[7] who observed the debate from a linguistic angle and stated that words can mean innumerable

[5] A concise description of the debate is provided by Cartwright (1968, p.615): "What are the essential attributes of, say, Dancer's Image? No doubt it will be counted essential that he is a horse and accidental that he was disqualified in this year's Kentucky Derby. But what of the attribute of being male, or of being a thoroughbred, or of not being a Clydesdale stallion? Here, I suppose, essentialists may disagree. Indeed, a reasonable essentialist might well take the position that these are hard cases that admit of no clear decision."

[6] For more details on Aristotelian essentialism, see Aristotle (2014) and Matthews (1990).

[7] See Cohen (1968) for more details.

things depending on their use, paving the way to modern pragmatics. Probably the most common standpoint takes into account that both minimal and maximal essentialism apply. *Maximal essentialism* states that all of any given object's properties are essential to it, while *minimal essentialism* presupposes that there are no limits to the ways a given object might have been different from its current actual state, and the only essential properties seem to be trivial ones, like "being *F*" or "being non-*F*" for any property *F* and "being self-identical" (Robertson Ishii and Atkins, 2020). For the purpose of this paper, we will consider the most common stance as our starting point: maximal and minimal essentialism both hold. The mentioned doctrine that at least some objects have at least some essential properties is the most common one (Robertson Ishii and Atkins, 2020).[8]

Previous work on machine learning and essentialism dealt with various types of machine learning under the same hood, connecting them often to essentialist ideas. Works of Pelillo (2013), Pelillo and Scantamburlo (2013) seem to be the most prominent ones. Tunç (2015) follows Pelillo's (2013) ideas but mostly focuses on epistemology and inductive inference, emphasizing abstracting, idealization, and theoretical variables in machine-learning research. Duin (2015) provides an anti-essentialist approach in pattern-recognition systems, claiming that in most of the applications in pattern recognition, there is no known, small set of essential features (a notion we agree with). Our goal is to show how various cases of essentialist-like and non-essentialist-like stances can be seen manifested in machine-learning choices, but that does not mean we are talking about real essentialist or anti-essentialist ontology.

[8] Explicitly stated by Mackie (2006). For more details about various types of essentialism, see (Robertson Ishii and Atkins, 2020).

Machine-learning basics

Machine learning, as a part of artificial intelligence and computer science, is a field of approaches and methods that use data in order to improve their performance on some problems. Artificial intelligence can be seen as a certain type of philosophical engineering (Skansi, 2018, p.vii): we want "to build machines that can think, [. . .] understand the meaning, act rationally, cope with uncertainty, [. . .] handle and talk about objects". In a nutshell, we are replicating standard philosophical concepts. It is no wonder that philosophical concepts are deeply embedded in their methods as well but may seem hidden underneath technical layers.

In machine learning, data is usually split into *training* and *test* data, the same way a student learns methods and approaches to some problems and gets previously unseen ones in an exam. Such an approach, compared to learning in the presence of a supervisor or a teacher, is called *supervised learning*: an algorithm learns from *labeled* data and is able to predict outcomes on previously unseen data. For example, if we had a dataset consisting of various pictures of animals, and we wanted to train the algorithm to recognize cats, we would want it to be able to somehow point out what is *essential* for an animal to be classified as a cat. An important part of supervised learning is therefore the act of *classification*: a certain object of interest possesses or does not possess certain property, i.e., it is or is not a member of a class. A certain image of a dog might be marked as 98.6% dog if it is very close to all of the properties that seem to be essential for classifying a picture of an animal as a dog. However, a cat might have some properties, such as four legs and a tail, but that would be a low percentage. In another class of problems, there are *regression* problems, in which the algorithm is predicting continuous

values. For example, given previous real-estate prices in a certain area, predict the prices for the next couple of years. Here, we would be dealing with real numbers instead of binary Boolean classifications.[9]

To summarize, a supervised machine-learning algorithm receives a set of training data points (a point in space where the axes are the properties given) and labels (row vectors), and in this phase, the algorithm creates a hyperplane—a decision boundary that helps classify its data points—by adjusting its internal parameters (Skansi, 2018, pp.55–56). This phase is the training phase that receives inputs as row vectors with corresponding labels (called training samples). In the next, predicting phase, the trained algorithm takes a number of row vectors, but this time without labels and creates the labels with the hyperplane (Skansi, 2018, p.56). In other words, "the learner receives a set of labeled examples as training data and makes predictions for all unseen points, [. . . a scenario commonly] associated with classification, regression, and ranking [i.e. ordering items to some criterion] problems" (Mohri, Rostamizadeh and Talwalkar, 2018, p.6).

Another type of machine learning, *unsupervised learning*, handles various datasets without any explicit instructions or labels. That is, the learner receives unlabeled training data and makes predictions for all unseen points, and "since, in general, no labeled example is available [. . .], it can be difficult to quantitatively evaluate the performance of a learner" (Mohri, Rostamizadeh and Talwalkar, 2018, p.7). Unsupervised learning encompasses a broad definition of learning without labels or targets, but this broad definition begs the cognitive question of how we learn without feedback (Skansi, 2018, p.70). In the previous case, in order for the computer to learn what is a dog,

[9] As a side note, most of the algorithms do not predict using Boolean outcomes such as 0 or 1 for not being or being a dog, but as a matter of a percentage. In such cases, we are effectively talking about fuzzy intervals.

we had to correctly label dogs or provide a list of properties in other supervised examples. Here, a computer is seemingly on its own: for example, neural networks[10] tend to automatically find structures in the data by analyzing useful features. Data is often grouped into *clusters*, and then it is easy to see the outliers, anomalies (for example, for fraud detection), associations (for recommender systems), and similar connections.

We might start to notice something interesting here. First, if we are telling the computer while we are labeling the data that something is or is not a certain kind of object, we are effectively taking a certain essentialist stance. Intuitively, there seems to be something *essential* in all of the properties that make a cat a *cat*. In various cases of supervised learning, we might list a number of features that we could consider important. For example, my algorithm might be tracking pointy ears, four legs, two eyes, and fur. But such an algorithm might recognize dogs and rabbits as well but miss some dogs without pointy ears. And we are not even starting to talk about three-legged dogs and similar "obviously" accidental properties. Second, it all boils down to starting human decisions. This seems like a trivial claim, and from a description of supervised learning, it is rather intuitive. Blaming it on the data might seem like a common excuse in machine learning, but recently, AI ethics has dwelled on the questions of initial data handling and responsibilities.[11] However, why did we choose some features on top of others? The answer might lie in human psychology.

[10] See Skansi (2018) for an introduction to deep learning.

[11] A famous example of an accidental algorithmic breach of ethics includes machine-learning racism in tagging black people as gorillas. See Zhang (2015).

Psychological essentialism

Gelman (2004, p.405) describes how once children learn a new fact about one member of a category, they generalize the fact to other members of that category, even if they look substantially different. By four years of age, children display subtlety and flexibility when they make category-based inductive inferences. For Gelman (2004, p.405), properties seem to be "fixed at birth", demonstrated by the following experiment. A child might learn about a newborn kangaroo that was switched at birth, and then went to live with goats. The child was then asked whether the animal would be good at hopping or climbing, or if would it have a pouch or not. Turns out, preschool children typically reported that it would have been good at hopping and have a pouch, something that seems *inherent* to kangaroos even for children. Such an understanding seems to appear by about six years of age, and it might be as early as four years of age: the time when children reason about animals, plants, and social categories (Gelman, 2004, p.406).

By the age of two, children view *causes* as vital to what something is (Gelman, 2004, p.406). This is interesting from a philosophical standpoint. *Causal essentialists* hold that a property essentially bears its causal and nomic relations (Gibbs, 2018, p.2332). Such a stance constrains what is possible and rules out possibilities where a property bears causal and nomic relations differently from how it actually bears them (Gibbs, 2018, p.2334). It seems that the notion of a cause and similar notions of origins seems to be closely tied to our early-childhood understanding of such relationships. There are some intriguing mistakes here: Gelman (2004, p.406) mentions that children sometimes can be more "nativist" than adults. For example, five-year-olds claim that a child switched at birth will speak the language of their birth parents rather than adoptive ones. We know that this is

not the case, but it is intriguing to see how an essentialist "feeling" might not always be correct if we take cognitive development as our guideline. Causality is central to children's categories, claims Gelman (2004, p.406), since it provides consistent domain-specific causal explanations for the properties that members of a category share. That is, category membership is stable over transformations (a dog cannot be transformed into a cat), and internal properties seem to be salient to young children. In a way, this is how computers behave as well: learning from observations and from their parents and other people. In the case of supervised machine learning, that is a combination of a prelabeled dataset and learning from data.

Here, the notion of a *feature* comes in handy as an individual measurable property. In character recognition, features might be shapes and pixels, and in voice recognition, frequency, noise, and strength. In computer vision, we might be talking about blobs, i.e., regions in images that differ in properties from the rest of the surrounding regions, for example, in color or brightness. Basically, it is a collection of information used for future problem resolution. In the case of classification, this may be compared to children learning about classes, memberships, and categories. But, how to describe a dog, say, using words? What are some essential properties the algorithm would be searching for? For example, a type of face, number of various body parts, color, etc. Children learn to recognize various members of the class and then generalize and use this knowledge in novel situations, i.e., previously unseen examples of that class. We want the computer to follow a similar process. In order to generalize well, a good selection of features needs to be selected. In the next section, we will observe how such a process is followed in machine learning and how the question of feature selection has important philosophical consequences.

Features and prototypes

As mentioned, *features* tend to be measurable properties that are successful in discriminating and differentiating between different categories of data. For example, in face detection (Bishop, 2006, p.3), we aim to find features that are not only fast to compute but also preserve useful discriminatory information enabling faces to be distinguished from non-faces. The study of feature selection finds its practical needs in machine learning, where a learning algorithm constructs a description of a function from a set of input/output instances through interaction with the world. Machine learning is more concerned with non-continuous features, while pattern recognition deals with continuous ones. It is not the same, for example, to classify something as a dog or not, compared to finding a face or another pattern or blob inside an image (Liu and Motoda, 1998, p.2). Liu and Motoda (1998, p.2) state that many forms of representations for machine-learning functions are available, including first-order logic, which is interesting from a philosophical standpoint, or weighted networks, but they have focused on features since they are 1) *primitive* 2) *convenient* 3) *independent* 4) *widely used* 5) *reasonably general*, i.e., powerful for many applications.

The first condition is the most important one for a metaphysics approach, and they define it as "the basic units for defining a problem, a domain, or a world to be observed, and do not require much effort from human experts to design them". Taken into account that feature selection tasks often fall into the hands of non-metaphysicists, there is a hunch of an innate human ability to generalize and select something that might, at least in the layman's sense of the word, seem essential for the object in question. Features are also called *attributes*, *properties*, or *characteristics* and can be discrete, continuous, or complex

(Liu and Motoda, 1998, p.3). For example, a dataset consisting of various hairstyles might have a feature of [color], which would take color names or RGB codes as its discrete value, [hair_length] may be a continuous numerical value in centimeters or inches, while there might be a Boolean [is_dyed] with true or false discrete values.

Trying to describe a certain object by finding whether it has or has not some constitutional properties, along with describing them, is not a novelty of machine learning. The same formal approach was popular with the advent of structural linguistics. Since phonology studies its basic units—phonemes, morphology analyzes morphemes, and syntax inspects sentence elements such as subjects, objects, and phrases, it was natural to try to find a basic unit of meaning that would make semantics an equal member of the formalized grammatical discipline ensemble.

Semic analysis was the first approach that aimed to find minimal units of meaning, which later developed into *componential analysis* within the standard structuralist framework. In particular, Pottier (1964) analyzed various types of chairs in order to find out what are the minimal features needed in order to distinguish between them. For example, they might have a back side or not, might have arms or not, can be fixed or folding, can have one seat or several seats, etc. One can see that we are already dealing with both discrete and continuous values here. A classic example in the componential analysis is how to describe various words for human beings in various stages of their lives, taking into account their gender. A *man* can be described as [-woman] and [+adult] or [+man] and [+adult]. Here we are dealing with Boolean man/female and adult/not adult, which does not reflect the fuzzy values of such categories, but structuralist linguistics was extremely focused on binary oppositions. Next, a *woman* would be [+woman] and [+adult] or [-man] and [+adult], a *girl* would be

[+woman] but [-adult] or [-man] and [-adult], while a *boy* could be described as [-woman] and [-adult] or [+man] and [-adult].[12] Such an approach has been developed and changed but is still used in semantics, which, as one of its tasks, analyzes the internal structure of a word by finding distinct and minimal components of meaning (Palmer, 1981, p.108). In such a framework, we might differentiate our *dog* from a *wolf* by finding distinct features. Both are certainly [+animal] and [+canine], but we might add [+domesticated] to the dog and [-domesticated] to the wolf. Such choices often seem arbitrary and there is no consensus on what the most basic properties of objects or classes of objects are, and it would also seem necessary to connect not only machine-learning feature selection with philosophy but linguistics and psychology as well.

We have mentioned that a strict binarist approach may seem inadequate in many cases. Departing from a standard Aristotelian notion of fixed categories, Eleanor Rosch (1973) introduced the *prototype theory* in which there is a graded degree of belonging to a certain category: some members are more central than others. For example, whatever essential properties of a bird might be, it seems somehow intuitive that in this—perhaps arbitrary—category there are some *more prototypical members* or examples than others: a sparrow is a more prototypical bird than an ostrich or a penguin. But this seems culturally anchored in both time and space, an apple is a more prototypical fruit in Europe, but other fruits might be better examples in Africa, such as bananas.

In machine learning, a feature does not have to be a binary Boolean, it can also be seen and created as a certain prototype. For ex-

[12] Such a method was formed on the basis of Prague structuralist school dealing with phonology. A phoneme has a set of discrete properties, for example *b* would be [+voiced], while *p* would be [-voiced], but both would be [+labial] plosives.

ample, in image recognition, there is a need to give the best prototype for a category. In the case of supervised learning, if we are training our models to recognize birds, and we are only using edge-case birds, we are not using the most generalized and best prototype or a versatile dataset consisting of central and edge-case members. The majority of images presented in a labeled training dataset would be close to being a prototype of the category. If we wanted to recognize apples, a rotten or a half-eaten apple would not be a prototype but would be a wanted member of the class, and if we wanted to recognize cats, a one-eyed cat without ears would not be a prototypical image, but we would somehow like to get the essentials with our prototypes in order to also include this as a result. In this case, we would expect percentages stating the probability that something is a dog or a cat to be higher for prototypical members, possessing all the necessary features, and less for edge-case or less prototypical members of a category.

Essentialist paradigm(s) in machine learning

It seems intuitive and obvious that supervised machine learning incorporates some kind of essentialism. That is, we are either given discrete or continuous features in datasets that are used for our predictions, usually whether something is a member of a class or not. But there are other kinds of machine learning, and we must not ignore the notion of unsupervised learning. We have already mentioned unsupervised learning, in which a model tries to establish regularities, clusters, or patterns in previously unseen data. This can be compared to the process of human learning at an early age, in which a human being tries to generalize the already acquired knowledge. Consider this, even if you are getting an unlabeled dataset of weird alien creatures, you will

most certainly be able to connect similar ones together in groups or do classifications, even if you do not know what *is* actually in the background of your dataset. We would like the computer to do the same. For example, if we trained our models on a certain map, they might recognize landmasses, developed areas, forests, or wetlands and group them together, by finding similarities between them. In non-visual data, you might be presented with some numbers, say bank transfers, and you might connect the usual activity into groups, while the outliers might be suspicious.

Pelillo and Scantamburlo (2013) were one of the pioneers of trying to connect machine learning with metaphysics. For them, the majority of traditional machine learning techniques are centered around the notion of a "feature", which we have observed. However, they note that there are numerous application domains where either it is not possible to find satisfactory features, or they are inefficient for learning purposes. Such examples might include cases when experts cannot define features in a straightforward way (e.g., protein descriptors vs. alignments), cases when data are highly dimensional (e.g., images), situations when features consist of both numerical and categorical variables (e.g., person data, like weight, sex, eye color, etc.), or in the presence of missing or inhomogeneous data.

In his overview of pattern recognition, which is mostly unsupervised, Pelillo (2013) states that features are *essential* properties. He reports Watanabe (1985) stating that "under all works of pattern recognition lies tacitly the Aristotelian view that the world consists of a discrete number of self-identical objects provided with, other than fleeting accidental properties, a number of fixed or very slowly changing attributes. Some of these attributes, which may be called 'features', determine the class to which the object belongs. Pellilo (2013, p.2) reaffirms that the goal of a pattern recognition algorithm

is to discern "the essences of a category" and that we should talk about an essentialist paradigm in machine learning. We have already mentioned Rosch's (1973) work on prototypes, which Pelillo (2013, p.2) uses to illustrate the "multifaceted nature of real-world categories" and emphasizes that for anti-essentialist stances, relations are in focus. That, of course, does not have to be the case, the main idea for anti-essentialism is to claim accidentality: there are possible worlds in which the object has the property in question and possible worlds in which it does not. But he does emphasize that the feature-based aspect is a *reductionist* position since objects are seen in isolation and overlook relational or contextual information (Pelillo, 2013, p.1).

The notion of a *feature vector* is often used in machine learning: an n-dimensional vector that serves a purpose of a collection of features. For example, just as a red/green/blue combination will form a single color, a certain combination of features will be used in machine-learning tasks to better identify objects or predict values. Pelillo (2013, p.3) emphasizes that the community has focused on feature-vector representations, rather than on single, standalone features. In computer vision and pattern recognition, each object is described in terms of a vector of numerical attributes and mapped to a point in a Euclidean vector space, so that the distances between the points reflect the similarities and dissimilarities between the respective objects (Pelillo, 2013, p.3). Pelillo emphasizes the recent trend in *similarity-based techniques*, which are still not challenging the traditional paradigm but work with graphs or structural representations to find objects or values that seem to be closer according to some criterion (Pelillo, 2013, p.4). We have to note that such an approach is analogous to a prototypical relationship, where members are grouped around a prototype in a certain graph-like manner. A green apple is

more similar to the prototype of a red apple than a red strawberry, and if such connections would be shown as a weighted graph, then we would expect a less expensive traversal to a red apple.

Accidental properties in machine learning

The processes and disciplines of *feature selection* and *extraction* show us that there is a strong presupposition that something as essential as a feature exists. There is no doubt that machine learning today is still enveloped in a strong essentialist paradigm. In feature engineering,[13] a system automatically discovers representations needed for feature detection. For example, it finds close points (neighbors) in a graph and clusters data around, say, percentages. If feature engineering is an essentialist stance, what kind of essentialism is it? It seems that is not maximal, but also not minimal, we would expect it to lie somewhere in between, judging by its success factor.

Here, what is interesting is that, unlike in human-led feature selection, automated feature engineering may use features that a philosopher would deem completely accidental, but it would still do a great job in classification or similar predictions. That is, deep-learning feature engineering does not have to correspond to some natural kinds or essential properties: it is not really essentialism, but a certain kind of accidentalism.

Namely, sometimes, features even outside deep learning that generate best models are often surprising and maybe even lucky correlations.[14] A famous example is a system (Lapuschkin et al.,

[13] For more details about feature engineering, see Zheng and Casari (2018).

[14] Some would argue that such processes might fall under the umbrella of *unexplainable AI*, if we are dealing with multiple layers within deep neural networks, but in

2019) performing horse recognition that learned to cheat by looking for the copyright watermark in horse images instead of finding some horse-essential features.[15]

When it is led by humans, that does not mean that there is an omniscient metaphysicist in computer engineers deciding what is essential and what is not. There are two important problems in machine learning. The first one is *underfitting*, the case in which a model is too general and does not fit the data property. For example, if we were doing dog recognition, from the training set, our underfitted model would consider that necessary features would be to have pointy ears and tails. In this case, we might recognize cats and rabbits too. An *overfitted* model has the opposite problem, it too closely responds to training data, and it is too specific. Basically, as if you only knew how to solve problems that appeared in your homework, but you are unable to solve the same problem when the numbers are replaced with other numbers. Our model might only recognize white and fluffy dogs with grey spots on their backs. Such a case might also be a result of bad feature engineering in the first place. Using automated feature engineering actually reduces the overfitting of your models, taking into account the standardized method of figuring out which one of your selected features might cause problems for your model to be too specific. We might imagine a case in which that also might seem like

---

the worst cases, "unexplainable" is not *impossible* to test or retrace, just *not easy*. We deem that the problem is not in unexplainability, but usually in the human inability to comprehend the data or the wrong (perhaps "accidental") approach taken.

[15] There are various legends and "folk tales" stating variations of a tank story, in which Russian tanks were photographed during daytime, unlike British tanks, so the AI system used that to its advantage. Most of such stories are farfetched but they do serve a purpose of illustrating a *possible* way an AI system might come to the right conclusion using the wrong method.

an essential property, but not for machine-learning purposes. Properties chosen or discovered might not be relevant or essential but make the model perform well.

Machine learning or essentialism?

Our previous conclusion might imply two separate things. Either there is an anti-cybernetic stance in which human learning that encompasses a certain kind of innate essentialist knowledge is a different process in machine learning, or that, for practical purposes, knowledge of essential properties is not a necessary prerequisite for everyday classifications and predictions. The latter seems more intuitive. It does seem that a similarity-based approach, mimicking the prototypical relationships found in psychological and linguistic research, may work well in various human and machine usages, along with a combination of properties (features) together with their relations (cf. feature vectors). For some machine-learning tasks, pure essentialism, often a binary or Boolean one, works best. We believe that essentialism and anti-essentialisms are not binary choices a computer scientist or a philosopher must make in order to describe how processes are being generated and run in machine learning paradigms today, but it is a matter of choice *for a specific type of task*. There is no essentialism equated with machine learning, but there is both essentialism and anti-essentialism for specific tasks. For some classification tasks and simple pattern recognitions, essentialist features are often the best choice, and for others, systems will work better with combinations of these properties. For unsupervised learning and pattern recognition, prototypical systems, i.e., similarity-based approaches, perform better. A philosophical take here is that, at least in machine learning, there is

no ontological obligation towards either of these stances, but rightful usage for rightful tasks. The choice of your machine-learning system, and therefore, a supervised or unsupervised approach, will depend on the type of task in question: *what performs better*. It is just a matter of technical performance that has no metaphysical consequences of the existence of essentialism or anti-essentialism.

From a psychological standpoint, Gelman (2005) has shown that essentialism is present in our everyday choices and is a reasoning heuristic readily available to both children and adults. As human beings, we seem to be hard-wired to search for parts and underlying structures. She claims that preschool children and adults from a variety of cultural contexts expect members of a category to be alike in a non-obvious way. That is, we treat "certain categories as having inductive potential, an innate basis, stable category membership, and sharp boundaries" (Gelman, 2005). It is no wonder that essentialist research has emerged as a metaphysical position. However, often, in our everyday practice, we are proven wrong, and that goes for our early childhood as well: Gelman's (2004) example of children being more nativist than adults. If essentialism might not always be the right choice for humans in various contexts, then the characterization of machine learning as an "essentialist" paradigm only reflects our inner psychological phenomena.

In philosophy, such an idea is present in the stance of conventionalism. Conventionalism seeks to expose conventions likely to be mistaken for truths (Ben-Menahem, 2006, p.2). This relativistic view is close to our claim that both supervised and unsupervised learning are plagued with human psychological categories that do not say anything about the possibility of objective categories, but only that we might or might not interpret conventions in various ways, even in essentialist and anti-essentialist terms.

As we have shown, the dichotomy should have never been the one about the differences in learning by humans or machines since these epistemic differences do not exist. The first reason is simple: machine learning is modeled after human learning, and only after the initial modeling is fine-tuned to make it computationally feasible. It is "essentially" the same by design. The differences are, again by design, accidental and purely due to different hardware/wetware. The second reason is more cybernetic in nature: if we are to develop a learning theory, it should be able to be as general as possible. Today one would never accept a psychological theory that only explains fear in adults or anxiety in women. Even though we might need to limit our theory in such a manner until further research is conducted, we would never accept this to be a completed theory. A theory of learning which would explain learning in children but not adults would likewise be incomplete and unacceptable except as a work in progress. This theory would be expanded to adults, people with disabilities, and to different cultures. After all, this is supposed to be a general theory of learning. Even though limiting the theory to humans might sound appealing, one could speculate that there will be more than a handful of researchers interested to see how such a theory applies to apes or dogs. Xenobiologists might take an interest too, as could AI researchers. Social scientists and cultural anthropologists might be also tempted to see if such a theory can describe models of societal learning or cultural integration. The point here is that the cybernetic call is a very natural force in scientific expansion and research, one that is to be expected, and one we had seen in a number of fields, perhaps the most recent and interesting one being social physics (as a branch of social network analysis). The insights gained in this fashion not only have huge practical benefits, but they do tend

to encompass a basic scientific curiosity, which no philosophy of science can avoid: "they say X and Y are not connected, but what happens if I use X on Y?".

The true dichotomy still present is a wholly different one. In fact, it is the same one that René Descartes described half a millennium ago: do the categories present in my mind have objective validity?[16] The easiest way to a positive answer is essentialism, which claims that the categories in our minds are formed via the essential properties present in the world. And machine learning, a new paradigm where machines are finally intelligent enough, is believed by many to show exactly this. If machines can learn the same things we do, then obviously the categories used are not intrinsically human. If machines can learn this by crunching data obtained from the world, then the categories are in fact present in that very data as essential properties. Machine learning is, on this account, simply a family of algorithms capable of extracting not just information from data, but essential properties as well. As we have shown, this view is wrong, since: (i) this could in theory hold true only for supervised learning, and more importantly (ii) supervised learning is defined via its use of targets or labels which are *man-made*. Since they are man-made, this means that human annotators bring in their categories "cat/dog", "animal/non-animal", "happy/sad", etc., and connect this to actual data, e.g., pixel values, or numeric data. The machine-learning algorithm then extracts this connection and applies it to previously unseen data. But the essential properties are not the ones discovered by the algorithm, they are brought in by human annotators, and do not have to reflect the "real" ontology at all. Even in the case of unsupervised learning, the features are being clustered and interpreted by humans, bringing again their own categories into play.

[16] See Descartes (1641; English translation: 1991) for more details.

Essentialist and anti-essentialist stances are both present in supervised and unsupervised learning, but we have pinpointed a couple of claims. First, supervised learning is easily connected with essentialism, but we wanted to pinpoint that it does not bear an *ontological commitment* to the existence of such features. Even though the view itself that humans are creators of essential features in machine learning might seem trivial, it does not say anything about ontology, but it says a lot about human psychology. Second, we might talk about the anti-essentialist stance in unsupervised learning (as Duin (2015) does), but this again is a strong ontological claim. Our goal was to show that unsupervised-learning approaches follow the prototypical learning and categorization model, inherent to human psychology, which also might be something the model creators are bringing to the model itself. The choice of supervised or unsupervised methods, which some might equate with essentialist or essentialist stances, actually does not exist since the choice depends on the problem we want to solve. Machine-learning systems do not discover anything about background ontology, but they do show us human epistemology and psychology present in seemingly competitive stances.

Bibliography

Aristotle, 2014. Categories. In: Jonathan Barnes, ed. *The Complete Works of Aristotle: The Revised Oxford Translation, One-Volume Digital Edition*. 6. print., with corr. Vol. 71:2, *Bollingen series*. Princeton, N.J: Princeton University Press, pp.25–70.

Ben-Menahem, Y., 2006. *Conventionalism : From Poincare to Quine* [Online]. Cambridge; New York: Cambridge University Press. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=529339&lang=pl&site=ehost-live> [visited on 13 January 2023].

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*, *Information science and statistics*. New York: Springer.

Cartwright, R.L., 1968. Some Remarks on Essentialism. *The Journal of Philosophy* [Online], 65(20), pp.615–626. https://doi.org/10.2307/2024315.

Cohen, M.F., 1968. Wittgenstein's anti-essentialism. *Australasian Journal of Philosophy* [Online], 46(3), pp.210–224. https://doi.org/10.1080/00048406812341181.

Descartes, R., 1641. *Renati Des-Cartes Meditationes de prima philosophia, in qua Dei existentia et animae immortalitas demonstratur.* [Online]. Paris: Michael Soly. Available at: <https://gallica.bnf.fr/ark:/12148/btv1b86002964> [visited on 25 August 2021].

Descartes, R., 1991. Meditations on First Philosophy. *The Philosophical Writings of Descartes, vol. 2* (J. Cottingham, R. Stoothoff and D. Murdoch, Trans.). Cambridge: Cambridge University Press, pp.1–63.

Duin, R.P., 2015. The dissimilarity representation for finding universals from particulars by an anti-essentialist approach. *Pattern Recognition Letters* [Online], 64(C), pp.37–43. https://doi.org/10.1016/j.patrec.2015.04.015.

Gelman, S., 2004. Psychological essentialism in children. *Trends in Cognitive Sciences* [Online], 8(9), pp.404–409. https://doi.org/10.1016/j.tics.2004.07.001.

Gelman, S.A., 2005. *Essentialism in Everyday Thought*. Available at: <https://www.apa.org/science/about/psa/2005/05/gelman> [visited on 12 January 2023].

Gibbs, C., 2018. Causal essentialism and the identity of indiscernibles. *Philosophical Studies* [Online], 175(9), pp.2331–2351. https://doi.org/10.1007/s11098-017-0961-y.

Kripke, S.A., 1972. Naming and Necessity. In: D. Davidson and G. Harman, eds. *Semantics of Natural Language* [Online], *Synthese Library*. Dordrecht: Springer Netherlands, pp.253–355. https://doi.org/10.1007/978-94-010-2557-7_9.

Krzanowski, R. and Polak, P., 2022a. Ontology and AI Paradigms. *Proceedings* [Online], 81(1), p.119. https://doi.org/10.3390/proceedings2022081119.

Krzanowski, R. and Polak, P., 2022b. The Meta-Ontology of AI systems with Human-Level Intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.197–230.

Lapuschkin, S. et al., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* [Online], 10(1), p.1096. https://doi.org/10.1038/s41467-019-08987-4.

Liu, H. and Motoda, H., 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA: Kluwer Academic Publishers.

Mackie, P., 2006. *How Things Might Have Been: Individuals, Kinds, and Essential Properties* [Online]. 1st ed. Oxford: Oxford University Press. https://doi.org/10.1093/0199272204.001.0001.

Marcus, R.B., 1993. *Modalities: Philosophical Essays* [Online]. New York: Oxford University Press. Available at: <http://catdir.loc.gov/catdir/enhancements/fy0638/91048105-t.html> [visited on 12 January 2023].

Matthews, G.B., 1990. Aristotelian Essentialism. *Philosophy and Phenomenological Research* [Online], 50, pp.251–262. https://doi.org/10.2307/2108042.

Mohri, M., Rostamizadeh, A. and Talwalkar, A., 2018. *Foundations of Machine Learning* [Online]. 2nd ed., *Adaptive computation and machine learning*. Cambridge, MA: The MIT Press. Available at: <https://cs.nyu.edu/~mohri/mlbook/> [visited on 13 January 2023].

Palmer, F.R., 1981. *Semantics* [Online]. 2nd ed. Cambridge: Cambridge University Press. Available at: <http://archive.org/details/semantics00palm> [visited on 13 January 2023].

Pelillo, M., 2013. Introduction: The SIMBAD Project. In: M. Pelillo, ed. *Similarity-Based Pattern Analysis and Recognition* [Online], *Advances in Computer Vision and Pattern Recognition*. London; Heidelberg; New York; Dordrecht: Springer, pp.1–10. https://doi.org/10.1007/978-1-4471-5628-4_1.

Pelillo, M. and Scantamburlo, T., 2013. How Mature Is the Field of Machine Learning? In: D. Hutchison et al., eds. *AI*IA 2013: Advances in Artificial Intelligence* [Online]. Vol. 8249. Cham: Springer International Publishing, pp.121–132. https://doi.org/10.1007/978-3-319-03524-6_11.

Pottier, B., 1964. *Vers une sémantique moderne*. Strasbourg: Klincksieck.

Robertson Ishii, T. and Atkins, P., 2020. Essential vs. Accidental Properties. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Winter 2020. Stanford, CA: Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/win2020/entries/essential-accidental/>.

Rosch, E.H., 1973. Natural categories. *Cognitive Psychology* [Online], 4(3), pp.328–350. https://doi.org/10.1016/0010-0285(73)90017-0.

Skansi, S., 2018. *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence* [Online], *Undergraduate Topics in Computer Science*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73004-2.

Tunç, B., 2015. Semantics of object representation in machine learning. *Pattern Recognition Letters* [Online], 64(15), pp.30–36. https://doi.org/10.1016/j.patrec.2015.03.016.

Watanabe, S., 1985. *Pattern Recognition: Human and Mechanical*. New York: John Wiley & Sons, Inc.

Zhang, M., 2015. *Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software*. Available at: <https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/> [visited on 13 January 2023].

Zheng, A. and Casari, A., 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. 1st ed. Beijing; Boston; Farnham; Sebastopol; Tokyo: O'Reilly.