UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Concept and Feature Change in Scientific and Deep Neural Net Representations

Permalink

https://escholarship.org/uc/item/13t3k5v6

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Author

Votsis, Ioannis

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Concept and Feature Change in Scientific and Deep Neural Net Representations

Ioannis Votsis (ioannis.votsis@nulondon – www.votsis.org)
Northeastern University London

Abstract

Scientific representations and their constituent concepts change over time to reflect improvements in our understanding of the world. Similar improvements in understanding lead to changes in DNN-procured representations and their features. In this paper, we investigate whether useful methodological practices in concept change and in feature change carry across the two types of representations. We argue that there is indeed considerable potential for methodological cross-pollination and offer some examples of how such benefit may be derived.

Keywords: artificial neural nets; concepts; conceptual change; deep neural nets; featural change; features; models; theory change.

Introduction

Arguably, the main aim of science is to provide adequate, and in ideal circumstances correct, representations of the natural, social and human worlds.[1] Scientific representations take many forms, e.g. theories, hypotheses, models, laws and principles. Here we emphasise scientific representations as extra-mental artifacts, not as mental states like beliefs (Dennett 1982; Fodor 1975; Simon 1978). What all representations share is that they are means through which we conceptualise the world. Their constituent parts can thus be thought of as concepts and relations between concepts. For example, the special theory of relativity, relates the concepts of mass and energy, the Lotka-Voltera model relates the concepts of predators and prey, and the Sapir-Whorf hypothesis relates the concepts of language and cognition.

Over time, scientific representations tend to change. The classical representation of physical systems, for example, was supplanted by relativistic and quantum representations. Any change of representations (typically) involves a change of concepts (and their relations). The classical concept of mass, roughly the amount of matter in a body (later understood as a measure of its resistance to acceleration), was reconceived in relativistic physics as two concepts: rest mass, which is invariant across all reference frames and observers and relativistic mass, which is not.

Representations also exist in other disciplines and human endeavours. This includes art, architecture and engineering, among others (Frigg 2022; Frigg & Nguyen 2020). More recently, representations in machine learning (ML) and deep neural nets (DNNs) have been garnering increased attention. Like scientific representations, DNN-procured models aim to provide adequate, and in ideal circumstances correct, representations of any target system for which we can gather data. This includes playing games, detecting patterns in images and inferring the 3D structure of proteins from amino acid sequences. Given the overlap in aims between scientific representations and DNN-produced models, it should be unsurprising that the latter have also been employed in science, i.e. as scientific representations. To avoid confusion, we hereby label scientific representations that have not been acquired via DNNs or indeed other AI methods 'classical'.

Like classical scientific representations, models produced via DNNs (whether deployed for scientific or non-scientific purposes) are also made up of constituent parts. Instead of concepts and their relations, they are made up of variables (also known as features), parameters (also known as weights or connections – think of coefficients) and the functional relations between them. The task of neural net practitioners is to optimise model parameters by minimising the difference between the predicted output values and the actual values measured. Typically, the variables are manually designed and selected by humans, but there are also possibilities to extract and select them automatically. Like classical scientific representations, DNN-produced models undergo changes over time. For example, large language models such as BERT, Gemini and GPT, undergo changes during the training phase, but also via fine-tuning and via successive iterations (e.g. GPT-3.5 vs. GPT-4). As the constituent parts of a DNN-produced model include variables, parameters and functional relations, any change in such a model means a change in its variables, parameters or functional relations.

Having motivated the various surface level similarities between classical scientific representations and DNN-procured models, the question arises whether there are useful methodological practices that carry across the two 'domains'. The present paper tries to answer this question by identifying such practices and by considering in what ways they can have a positive impact on the respective domains.

The next section identifies useful methodological practices in classical scientific representations, particularly those that concern conceptual change. Directly following it is a section that identifies such practices in the context of DNN-produced models, with a focus on feature change. Succeeding those sections is a section that discusses the potential for methodological cross-pollination between the two contexts. The paper concludes with a summary of the main points.

Concept Change in Scientific Representations

We earlier broached the subject of changes in classical scientific representations and their constituent concepts (and

[1] Some philosophers of science suggest that practicing scientists prefer less accurate/correct/true representations on epistemic grounds, e.g. Potochnik (2017) and Elgin (2017).

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

relations), but did not venture to answer the question why those changes take place. The primary reason behind such changes is the demand to improve the adequacy of those representations and concepts with respect to the empirical phenomena. It should be clear that no (classical or other) scientific representation is perfectly adequate. Indeed, far from it, most such representations tend to be rather poor in this regard, at least when they are first posited. That means that those same representations and their constituent concepts (and relations) need to be adjusted to boost adequacy. How exactly scientists go, or ought to go, about making those adjustments is a subject studied by philosophers and historians of science (Andersen, Barker & Chen 2006; Kitcher 1993; Kuhn 1962) as well as psychologists and cognitive scientists (Barsalou 1992; Carey 2000; Nercessian 2008), among others. Various recommended methodological practices have emerged as a result. In this section, we identify some key practices, citing concrete examples along the way.

The first commendable methodological practice has to do with the import of mathematising scientific representations (Redhead 2001). The precision afforded by mathematisation ensures scientific representations stick their neck out for potential empirical falsification and refutation (Popper [1959] 2005). This in turn allows representations to be more carefully scrutinised and assessed, increasing the likelihood of progress towards understanding the world. Quantitative representations also mean that the concepts at stake get operationalised into measurable variables. For example, the concepts of predators and prey are operationalised by population density variables in the Lotka-Voltera model.

The second commendable methodological practice has to do with the import of trying out different variables and mathematical relations. Gopnik (2020) makes a similar point when she asserts that we must allocate enough time and resources for the exploration of scientific hypothesis spaces. Even if we assume that experimental scientists have chosen the right variables and measured them both accurately and precisely – a big ask by any account – it is non-trivial to decide how exactly to mathematically relate these variables. As an example of choosing a wrong variable, consider the angular velocity of a deferent (i.e. the theoretical centre of a planet's epicycle) in Ptolemaic astronomy. This helped motivate another wrong choice, namely the mathematical relation that a deferent's angular velocity is uniform with respect to the equant (i.e. another theoretical point in space). Bad choices such as these prevented Ptolemaic astronomy from ever settling on a stable empirically adequate account of planetary motions. It wasn't until Copernicus that the equant was justly eliminated from astronomy, and Kepler that the relevant variable was replaced with area speed. The latter allowed the formulation of the second law of planetary motion: the line joining a planet and the Sun sweeps out equal areas in equal time intervals. Even this mathematical relation, however, became a hard-won victory only after Kepler switched from a circular to an elliptical conception of orbits, which demonstrates once again the importance of trying out different variables and different mathematical relations.

The third, fourth and fifth commendable methodological practices concern solutions to the problem of concept inadequacy. Let us start with the third. This is the realisation that some concepts are empty. That is, they do not correspond to anything real. We already saw some examples (the equant, epicycles and deferents) in Ptolemaic astronomy. Other putative examples from the history of science, include the phlogiston, the caloric and the ether. All of these concepts are now defunct, often used as cautionary tales of how not to do science. The moral of these stories is that scientists must be prepared to move away from concepts that repeatedly lead nowhere, or, as Lakatos (1968) once put it, when scientific research programmes become degenerative.

Having said this, there is a fine line to be drawn between empty concepts that play no positive role in the adequacy of scientific representations, and those that are only seemingly empty as they do play such a role. In recent years, philosophers of science (e.g. Ladyman 2011; Votsis & Schurz 2012; Worrall 1989) have argued that all three of the above posits (the phlogiston, the caloric, and the ether) fall under the latter category. As an illustration, consider the case of the caloric. The caloric was posited to be a special kind of matter, i.e. distinct from ordinary matter, pertaining to heat phenomena. Although now abandoned, its proponents correctly identified some relations that play a positive role in the adequacy of the corresponding modern scientific representation, which cites kinetic energy flow (vs. caloric flow) in its explanations of heat phenomena. One such relation concerns the thermal expansion and thermal contraction of bodies. To wit, a body thermally expands when (caloric / kinetic energy) is added because the internal pressure (the repulsive force between caloric particles / the collisions of ordinary matter particles with the body's boundaries) is increased and so the volume needed by the body to accommodate it also increases. Similarly, a body thermally contracts when (caloric / kinetic energy) is removed because the internal pressure (the repulsive force between caloric particles / the collisions of ordinary matter particles with the body's boundaries) is decreased and so the volume needed by the body to accommodate it also decreases. The moral is that even seemingly empty concepts may hide within them some truth that needs to be preserved across successive scientific representations.

The next two commendable methodological practices, the fourth and fifth on our list, have to do with the fact that oftentimes concept extensions are not quite correctly circumscribed, as they include either false positives or false negatives or both. Knowing that concepts, especially those in their nascency, are likely to include false positives and/or negatives means that it is important to try out different extension adjustments. One example from the history of science that involves both false positives and false negatives is the concept 'planets'. As is well known, planets were thought of as wandering stars by the ancients, including Ptolemaic astronomers, as they move across the sky against what appears to be a backdrop of fixed stars. To be precise, the extension of the ancient concept of planets consisted of

Jupiter, Mars, Mercury, Saturn and Venus, as well as the Moon and the Sun. The last two were understandably included in that extension in that, in the eyes of an Earthly observer, they also move across the sky relative to the fixed stars. Our modern conception of planets has both added objects to that extension, e.g. Earth, Uranus and Neptune (the false negatives under the old conception), and removed objects from it, e.g. the Moon and the Sun (the false positives under the old conception). On the modern conception, planets are roughly speaking large approximately spherical celestial bodies that orbit around a star but are not themselves stars.

An associated methodological set of practices involves the splitting or the merging of existing concepts. Carey (2000) calls such cases 'differentiation' and 'coalescing or integration' respectively. As an example of differentiation, she points to Galileo's transformation of the concept of speed into the concepts of average velocity and instantaneous velocity. Our own examples above include the splitting of the concept of mass into rest mass and relativistic mass, as well as the splitting of the concept of wandering stars into planets, satellites, and stars. As an example of integration, Carey points to Galileo's transformation of the Aristotelian concepts of natural motion and violent motion into a single concept of motion. Our own examples above include the merging of (revised versions of) the concepts of ordinary matter and caloric matter into the concept matter, as well as the merging of (revised versions of) the concepts of fixed and wandering stars into the concept stars.

The reason why we do not classify the operations of merging and splitting of concepts as separate methodological practices is because they are effectively produced via expansions and contractions of the corresponding extensions. It is also worth noting that additional expansions and contractions are often needed, beyond those required by the simple operation of merging and splitting concepts. That's because the concepts to be merged or split are themselves revised, as indicated above. Take, for example, the modern concept of stars. This does not simply merge the Ptolemaic concepts of fixed stars and wandering stars since the latter also include modern day planets and the Moon. Instead, it only merges a revised version of the concept of the wandering stars, namely one that includes only the Sun, with the fixed stars. The concept of the fixed stars themselves is also revised to, among other things, include stars unseen with the naked eye, and exclude what look like stars but turn out to be galaxies.[2] Needless to say, our modern conception of stars, sees them not as fixed on a celestial sphere and hence as being equidistant from the Earth, but as free floating and populating the universe at vastly different distances from the Earth.

The sixth and final commendable methodological practice concerns the importance of introducing new concepts. To be clear, by 'new' we do not mean that such concepts have no anchor in something we understand, interact with, see or imagine. Rather, and generally speaking, they are concepts that are considerably less tethered to existing ideas. Indeed, no claim is made here that there is a clear cut off point of what counts as a new concept. There are certainly some clear instances, and some clear counter-instances, of new concepts. Clear counterinstances, i.e. clear instances of concepts that are not new, include the modern concepts of planets and stars, as they are continuous in non-negligible ways with their Ptolemaic predecessors. Clear instances of new concepts include those first introduced as unseen common causes of seen correlations. The very identification of a correlation between seen things raises the question whether the things are themselves cause and effect or whether they are both effects of a yet-to-be seen common cause. Many discoveries and conceptual innovations in science proceed from the assumption that there is indeed such a common cause. It is thus useful to provisionally posit and subsequently attempt to observe and measure them. Examples from the history of science abound and include the postulation of the planet Neptune in physics to explain correlated disturbances in the orbits of other bodies in the solar system (most notably Uranus), and the postulation of common ancestors in biology to explain correlations between morphological characteristics like eyes in various species. Obviously, not all postulations are successful. The postulation of the planet Vulcan to explain the precession of the perihelion of Mercury is a good example where the presumed common cause turned out to be fictional. Still, common cause attempts to explain phenomena and, more generally, the introduction of new concepts are important tools in the theoretician's toolbox. Indeed, the explanation of Mercury's precession was only made possible by the introduction of the concept of mass as curvature of space in the general theory of relativity.

Before we bring this section to a close, it is worth highlighting that many of the aforesaid methodological practices are not unique to science, but can also be found, at least in some recognisable form, in everyday life. This is not news to many developmental and evolutionary psychologists. As Carey notes "Just as the concepts of person and animal change in fundamental ways throughout childhood, so do a host of interrelated concepts [in science] also undergo conceptual change" (p. 16).

Feature Change in DNN Representations

In the introductory section, we outlined the constituent parts of models procured via DNNs. In this section, we briefly explain the production process to put the representational capacities of these models into perspective. Moreover, since DNN-produced models strive to provide adequate, and in ideal circumstances correct, representations of their target systems, we identify various recommended methodological practices, whose aim is to improve their adequacy.

DNNs are made up of three layers (input, hidden and output) of neurons and their connections. What marks DNNs from other artificial neural nets is that they possess *several*

[2] Some galaxies, e.g. the Andromeda Galaxy, can be seen with the naked eye, though it appears like a single star.

hidden layers of neurons, i.e. their layers are 'deep'. In a typical neural net architecture, neurons at the input layer receive one instance of values at a time from all the input variables of a training dataset. They then pass these values on as outputs to neurons in the first hidden layer, which receive them as inputs. Which neurons in the hidden layer are receiving these values depends on whether there is a connection between the given input layer neuron and the given hidden layer neuron. These connections are differentially weighted. Each input for a hidden layer neuron value is multiplied by the weight of the corresponding connection. The output value of a hidden layer neuron is sequentially determined by: (1) the sum of all the weighted values of input layer neurons connected to it, and (2) the chosen activation (e.g. the sigmoid) function which takes that sum and transforms it, often in a nonlinear fashion. The output value of the given neuron in the first hidden layer then gets passed on as an input to one or more neurons in the next hidden layer, or, if there aren't any, the output layer. Either way, the process is repeated by calculating the sum of the given neuron's weighted inputs and applying an activation function. When this calculation takes place at the output layer neuron(s), a prediction or classification ensues. This can then be compared to actual measured values (in supervised learning) or to some other 'targeted' values (in unsupervised learning) to establish the loss function, roughly, how badly the values in the output match the desired (actual or 'targeted') values. An apt algorithm (e.g. backpropagation) is then deployed to minimise the mismatch by adjusting the connection weights to achieve a better fit. The whole process, i.e. from the feeding of input variable values into the input layer neurons to the calculation of output values at the output layer neurons and their comparison to desired values, is repeated numerous times until a model with good accuracy and good generalisability (i.e. external validity) is generated.[3]

Models produced via DNNs tend to be exceedingly complex. According to some estimates (Briganti 2024), the latest GPT models have hundreds of billions of parameters. It's not clear exactly how many variables – henceforth: features – are employed in such models, but what is clear is that the number is unacceptably high. AI theoreticians and practitioners are under to reduce the high dimensionality of (i.e. the high number of features in) datasets. On top of this complication concerning the sheer number of features, there is also the issue of the intricate functional relations that link the features. The form of these relations may be comprehensible, but a systematic understanding of how input gets transformed into output seems out of reach for the limited human mind. No wonder then that the resulting representations are notoriously difficult (some even say impossible) to decipher, earning them the sobriquet 'black boxes' (Bender & Koller 2020; Harnad 2024).

For our purposes we will focus on methodological efforts to reduce this complexity so as to make the DNN-produced models easier and less computationally costly to train and run, but also, hopefully, to make them more transparent, interpretable, and explainable. Some of these methods, e.g. pruning (Frankle & Carbin 2018; LeCun, Denker & Solla 1989), aim to simplify the DNN itself by removing connections between neurons or neurons themselves, without loss of accuracy. Other methods, e.g. feature selection and feature extraction, aim to simplify the datasets upon which the DNN-produced models are trained. Since the comparison between features (in DNN-produced models) and concepts (in classical scientific representations) is more intelligible than the comparison between the said neurons and concepts, the discussion that follows is restricted to the application of feature selection and extraction methodological practices.

What are feature selection methods? Such methods are employed to reduce the set of input features to a proper subset that makes the most significant contributions to the successful training and generalisability of the model. There are three kinds of feature selection methods: wrapper, filter and intrinsic methods.

Wrapper feature selection methods are computationally expensive as they involve training several models, each with a different subset of input features. The subset composition varies according to the specific method (e.g. forward selection, backward elimination, recursive elimination) employed to generate it. At the end of the process, the input features of those models that perform best are selected as the ones making significant contributions.

Filter feature selection methods assess the contributions of individual input features to the output variable, typically by employing statistical measures (e.g. Pearson's correlation, Spearman's rank and mutual information). The choice of statistical measures depends on factors like whether the values are numerical or categorical. The input features that come out on top in these statistical assessments, e.g. those that are most highly correlated with the output variable, are then selected as the ones making significant contributions.

Intrinsic (a.k.a. embedded) feature selection methods are so called because they are built into the normal training pipeline for models. Well-known examples of such methods include L1 (also known as Lasso) and L2 (also known as Ridge) regularisation. These methods penalise parameters with high values, thereby indirectly reducing or eliminating the impact of features that are multiplied by those parameters. Among the main benefits of regularisation is overfitting avoidance.

Besides feature selection methods, there are also feature extraction methods. The latter are employed to identify and obtain features from raw data. As with feature selection, feature extraction can be employed to help reduce data complexity, thereby speeding up training and improving performance by focusing on the most relevant features. They also help to simplify the model, rendering it more likely to be transparent, interpretable and explainable. Unlike feature selection, feature extraction methods do not pick out existing features, but rather automatically create new ones, or, more

[3] We here omit discussion of the validation and test phases of training neural nets for expedience.

accurately, transform existing into new features. In cases of medical image processing, for example, feature extraction may lead to the discovery of hitherto unknown characteristics that matter for diagnostic/prognostic purposes (Kuan 2017).

There are several distinct kinds of feature extraction methods, including autoencoders, component analyses (principal component, kernel principal component, independent component and linear discriminant analysis), and feature hashing. Autoencoders are special artificial neural nets that learn to compress unlabelled data from high to lower dimensional representations (encoding) that can then be used to reconstruct the original data (decoding). Component analyses transform the original features into a new set of fewer features. Principal component analysis, for example, linearly transforms the data so that they are framed within a new coordinate system, where the greatest variances are explicated by the new features (i.e. the coordinates). Feature hashing allows the conversion of categorical into numerical data, thereby reducing the relevant features by mapping them to a fixed-size vector. In all feature extraction methods, the transformation of features attempts to preserve the most important information present in the original dataset.

Finally, it is worth pointing out that feature selection and feature extraction methods are closely related to dimensionality reduction techniques. In fact, some researchers see feature extraction as a proper subset of dimensionality reduction techniques (de-la-Bandera et al. 2020). As the name suggests, dimensionality reduction involves the reduction of dimensions or features for the same reasons mentioned earlier, e.g. computational efficiency, etc. Of particular importance to us here is the so-called 'curse of dimensionality'. This is the observation that there is an inverse relationship between the number of dimensions (features) and the generalisability of the model. That is, other things being equal, the more dimensions in a model, the less generalisable the model is expected to be.

Methodological Cross-pollination

A good place to begin the discussion of whether there is potential for methodological cross-pollination between the two contexts, i.e. classical scientific and DNN-produced representations, is by specifying similarities and differences in the methodological practices they employ to change concepts or features. Once we achieve that, we can then move on to consider if some of the methodological practices that work well in one context may be successfully transplanted into the other context.

The first similarity between methodological practices to note is rather trivial, but it cannot be ignored as some researchers find this practice highly objectionable. Both classical scientific and DNN-produced representations are often mathematically expressed. In fact, the latter cannot but be mathematically expressed as the methods involved in producing them are thoroughly mathematical: all input gets numerically vectorised and processed through a sequence of summation and activation functions. Classical scientific representations, on the other hand, are not always so expressed. Indeed, there is still considerable resistance in producing mathematical representations in the social sciences and the humanities, not least because such representations are thought by some (e.g. hermeneutics advocates like Habermas 1965) to go beyond what is possible in describing behaviour. Such resistance has led to the creation of filter bubbles within and across disciplines (Maree & Maree 2020), with some researchers choosing to cut off all communication with their colleagues along these battle lines. The difficulty of mathematising representations in the social sciences and the humanities notwithstanding, we hope that our thoughts on the import of mathematisation are clearly laid out above.

The second similarity to note is the importance of trying out different mathematical relations between the concepts (strictly: the corresponding variables) or features at issue. In both contexts, the sky is the limit as to which mathematical relation may be tried out, as, *in principle*, any of them may lead to fruition. The universal approximation theorem vouches for the richness of mathematical expressibility via DNN-produced representations. This states, roughly, that any feedforward neural net with sufficient complexity can approximate any continuous function to any desired degree of accuracy.

The third similarity to note concerns the adequacy of a representation, which crucially depends on the choice of concepts or features. This is true in both contexts. We have already seen how some concept choices, e.g. Ptolemy's angular velocity of the centre of a planet's epicycle, impeded progress in the history of science. In the context of DNN-produced models, the importance of such choices becomes obvious in cases where feature selection and extraction methods are applied. When existing features offer nothing to write home about vis-à-vis the model's generalisability, such methods are employed to weed them out. Indeed, sometimes entirely new features are needed – that's where feature extraction methods become exceptionally handy.

The fourth noteworthy similarity is that both concepts and features can be reshaped to increase the adequacy of their respective representations. The operations that can be brought to bear in remoulding concepts are contraction (to decrease the false positives) and expansion (to decrease the false negatives). In the case of features, the reshaping is carried out by feature extraction methods. These allow new features to be extracted from raw data, features that seek to preserve the essential information found in that data.

The fifth and final similarity worthy of note is that new concepts and features are sometimes needed to overcome failures in representational adequacy. In the case of concepts, we have seen how the method of seeking a common cause can lead to new concepts that denote hitherto unobservable, but hopefully existing, things (posited to explain correlations between observables). In the case of features, as we already saw, feature extraction methods can lead to new features that seek to capture the essence of the kinds of patterns and trends already present in the raw data.

Besides similarities, there are some differences between the practices in the two contexts. These differences, we argue, act as opportunities for cross-pollination.

The first such difference is that, practically speaking, the choice of mathematical relations in either context is constrained in various ways. In classical scientific representations that choice is cognitively bounded. For example, throughout history scientists have been inclined to posit simple mathematical relations such as linear and inverse square equations. Such an approach has served us well in the past, as simple relations have turned out to be empirically adequate, but it may not continue to serve us as well in the future. For all we know, the next big step in understanding the world may require massively complex mathematical relations. The application of DNNs to science, their admittedly serious transparency, interpretability and explainability problems notwithstanding, may thus prove propitious, as they are clearly capable of constructing massively complex mathematical relations. At the very least, we have one good reason to mimic the DNN approach in liberating our choice of mathematical relations. On the other hand, DNN modelling tends to be unnecessarily profligate in expressing mathematical relations (Cf. Lemos et al. 2023). That's at least partly why all the aforesaid methods are imperative, including those that aim to simplify networks and those that aim to simplify datasets. Indeed, one may argue that knowing how simple most classical scientific representations are should inspire and put pressure on DNN theoreticians and practitioners to devise networks that lead to substantially more parsimonious models.

The second major point of divergence between the two contexts concerns the ways they go about selecting concepts or features. In the case of concepts, it is not always easy to see which ones (strictly: the corresponding variables) contribute to the adequacy of a scientific representation. Part of the problem has to do with the fact that it is hard to manually make various calculations. The same is not true of features, as feature selection and feature extraction methods automate these calculations by default. Indeed, sometimes a combinatorially exhaustive approach is adopted (e.g. some wrapper methods are like this) that guarantees success at the expense of speed. But besides directly applying ML and DNN methods to the context of scientific representation production, this approach offers another potential pathway to success. Sometimes a shift in scientific representations and their concepts requires brute force, and sometimes feature selection and extraction inspired heuristics may do the trick.

The third major point of divergence between the two contexts is that it is unclear if features can be empty like concepts. That's because concepts, unlike features, are not directly measurable. It's only when concepts become operationalised into variables that we can be sure that they are not empty, even though they may still be mistargeting, i.e. measuring something other than what we intend to observe. Features are, by definition, measurable properties, and, as such, do not face the emptiness challenge. They do, however, also face the problem of mistargeting. The question then arises whether any of the methodological practices discussed earlier can help with mistargeting. It seems plausible, at least prima facie, that practices analogous to feature extraction methods can be applied to concepts to shift their sights towards the right targets by reshaping them in the ways outlined above. Having the right concepts (or concepts fairly similar to the right ones), i.e. those that trace the contours of natural categories relatively faithfully, is paramount to being able to successfully predict and manipulate the world around us. As such, it is well-worth trying to analogously apply selection and extraction methods to concepts, as a conceptual engineering of sorts.

The fourth and final point of difference between the two contexts is how they go about reshaping concepts and features. In the case of feature extraction, the reshaping is achieved by trying to preserve variance in the data, since that variance is conjectured to reflect real patterns and trends. Even so, feature extraction methods seem to ignore another potentially fruitful way to reshape features, namely via systematically expanding or contracting the set of things to which those features apply. That's something that DNN practitioners and theoreticians can learn from concept change in the context of classical scientific representations. They need a method that performs analogous expansion and contraction operations with respect to feature extensions. We see no reason why such a method could not prove to be a useful supplement to existing feature extraction practices. Indeed, we provisionally propose the names 'feature expansion' and 'feature contraction' for these new methods.

Conclusion

This paper took as its starting point some initial similarities between classical scientific representations and models procured via DNNs. They both: (i) aim to provide adequate, and in ideal circumstances correct, descriptions of their target domains, (ii) change over time to improve their adequacy, and (iii) have constituent parts, concepts and features respectively, that often change along with them. The paper then explored useful methodological practices that influence conceptual change in the context of classical scientific representations and those that influence featural change in the context of DNN-produced models. These explorations set the stage and allowed for a more meaningful comparison of the methodological similarities and differences between the two contexts, and, more crucially, the potential for cross-pollination. It was suggested that cross-pollination is indeed promising, and some rough examples were offered to demonstrate how practices found in the one context may bring benefit to the other. The upshot and hoped-for outcome is that by assisting theoreticians and practitioners to come out of their methodological shells, new insights may be gained in our attempts to advance our understanding of the world.

References

Andersen, H., Barker, P., & Chen, X. (2006). *The cognitive structure of scientific revolutions*. Cambridge: Cambridge University Press.

Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields, and contrasts* (pp. 21–74). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185-5198).

Briganti, G. (2024). How ChatGPT works: a mini review. European *Archives of Oto-Rhino-Laryngology*, 281(3), 1565-1569.

Carey, S. (2000). Science education as conceptual change. *Journal of applied developmental psychology*, 21(1), 13-19.

de-la-Bandera, I. et al. (2020). Feature extraction for dimensionality reduction in cellular networks performance analysis. *Sensors*, 20(23), 6944.

Dennett, D. C. (1982). Styles of mental representation. *Proceedings of the Aristotelian society* (Vol. 83, pp. 213-226). Aristotelian Society, Wiley.

Elgin, C. Z. (2017). *True enough*. MIT press.

Fodor, J. (1975). *The language of thought*. Harvard University Press.

Frigg, R. (2022). *Models and theories: A philosophical inquiry*. Taylor & Francis.

Frigg, R., & Nguyen, J. (2020). *Modelling nature: An opinionated introduction to scientific representation*. Cham: Springer.

Frankle, J., & Carbin, M. (2018). 'The lottery ticket hypothesis: Finding sparse, trainable neural networks'. *arXiv preprint arXiv:1803.03635*.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, 375(1803), 20190502.

Habermas, J. (1965). Erkenntnis und interesse. *Merkur*, 19(213), 1139-1153.

Harnad, S. (2024). *Language writ large: Llms, chatgpt, grounding, meaning and understanding*. arXiv preprint arXiv:2402.02243.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press, USA.

Kuan, K. et al. (2017). Deep learning for lung cancer detection: tackling the kaggle data science bowl 2017 challenge. *arXiv preprint arXiv:1705.09435*.

Kuhn, T. S. (1996 [1962]). *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press

Ladyman, J. (2011). Structural realism versus standard scientific realism: The case of phlogiston and dephlogisticated air. *Synthese*, 180(2), 87-101.

Lakatos, I. (1968). Criticism and the methodology of scientific research programmes. *Proceedings of the Aristotelian Society*, 69: 149–186.

LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. *Advances in neural information processing systems*, vol. 2.

Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., & Battaglia, P. (2023). Rediscovering orbital mechanics with machine learning. *Machine Learning: Science and Technology*, 4(4), 045002.

Maree, D. J., & Maree, D. J. (2020). The Methodological Division: Quantitative and Qualitative Methods. *Realism and Psychological Science*, 13-42.

Nersessian, N. J. (2008). Creating scientific concepts. Cambridge, MA: MIT Press.

Popper, K. ([1959] 2005). *The logic of scientific discovery*. London: Routledge.

Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Redhead, M. (2001). The intelligibility of the universe. *Royal Institute of Philosophy Supplements*, 48, 73-90.

Simon, H. A. (1978). On the forms of mental representation. In W. Savage (Ed.), *Perception and Cognition* (pp. 3-18). University of Minnesota Press.

Votsis, I., & Schurz, G. (2012). A frame-theoretic analysis of two rival conceptions of heat. *Studies in History and Philosophy of Science Part A*, 43(1), 105-114.

Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica*, 43(1-2), 99-124.