Passing comments on neural nets and the Shivasutras

Luke Smith

Abstract

The rise of neural nets and resurgence of Connectionism have reinvigorated some interesting questions about the "interpretability" of scientific models. Must a good model be understandable in intuitive terms? I argue that there are generally two distinct goals in scientific modelled which are mututally reinforcing, but shouldn't be confused.

In that light, I'll also detail an early run in interpretability this in linguistics in Panini's early works on Sanskrit grammar. Specifically, Panini employed a notably "uninterpretable" model of Sanskrit morphophonology in the Shivasutras, grouping phonemes together with only questionable phonetic or phonological relation.

These apparently partially random groups, however, communicate very abstract relationships between sounds in the Sanskrit language. They allow Panini to postulate a very economical rule system, but also indirectly embed the workings of deeper Indo-European morphology, like the ablaut system (while Panini himself was unaware of this).

Two types of models

There are two main goals of scientific modeling: understanding and accuracy. To understand something is to have some kind of intuitive idea of the mechanism behind something, or why it works the way it does and how it interacts with other systems. But to be accurate is something different:

you don't necessarily need to understand the mechanism, just predict what it does and predict it well. To be clear, accuracy is *not* a lesser goal, as *understanding* something very well doesn't mean at all that you can predict it at all.

We can group scientific theories into two main categories based on this distinction. There are referential models, which aim at "understanding", and formal models, aiming at accuracy. The difference is not necessarily categorical, but more or less all scientific theories fall into one category or another.

We can describe the difference visually below:

	Formal Models	Referential Models
Goal:	Accuracy	Understanding
Object of study:	Output of a system	Mechanism of a system
Metric of value:	Use for prediction	Use for "making sense of it"
Example:	Statistical modeling	Psychological modeling

Which category a particular model falls into may depend on the motives of the analyst. Form example, formal logic is typically used as a formal model, but there is debate as to if it can be interpreted as psychologically real in some sense (see, for example, Partee (1979)).

Neural Nets are a particularly interesting formalism in that they notionally straddle both. Most implementations of Neural Nets are for practical purposes, eschewing understanding for the accuracy of formal modeling, but at the core, as their name suggests, Neural Nets are originally meant to be analogies for the actual mechanism of cognition.

Even critics of "connectionism" like Fodor and Pylyshyn (1988) acknowledge that part of the justification for these types of models is the potential cognitive reality of them (a reality defended by Oaksford and Chater (1994) and others). Neural Nets can thus have a kind of a dual identity, in that they are formally effective, but are sometime inside or outside of intuitive

understanding, giving us the general problem of interpretability.

How the types of models mutually reinforce

Now the categories of the human mind are useful for human life, but not necessarily for understanding the machinations of the external world; this presents an interesting "problem" for what I've called "referential models". This fact has even been noticed in the popular press; Richard Dawkins, for example, introduces the concept of "Middle World", meaning that relatively narrow band of reality that the mind mind has evolved to understand.

Humans have common-sense intuitions about the physical qualities of objects of about our size, how they fall or interact, but taken a level up, looking at the "macro" world of elliptical orbits and black holes, our "intuitions" about that realm are mostly learned and haphazard. Taken a level down, into the "micro" subatomic or quantum world, and the universe is only more confusing.

As a metaphor, it's often been said that humans do not really "climb" gracefully like other animals do, but ascend rocky inclines using a series of inelegant controlled falls. The same can be said about human cognitive life. We don't really understand quantum or intergalactic space at an intuitive level, but rely on a series of tactically employed metaphors to predict how the world outside our intuition works. This is true of the "macro" and "micro" worlds, but also true of our inner (cognitive) world.

With all that said, the domain of referential models is actually quite limited, as not very much in the universe makes sense to people. Still, *formal* modeling can actually help to expand this domain.

If we create an exhaustive formal model of a phenomenon, we expect that "intuitive" concepts have some reality in that model, for the mere fact that they seem to motivate the data. In the course of formal modeling as well, there might be some priors or mechanisms that are non-intuitive as well.

These categories may be the inexplicable nodes in a neural net or an apparently idle *ad hoc* principle posited for raw data solvency.

I would argue that the next step in the cycle is to take the formal model and evaluate if those unexplainable points are in fact, simply psychologically/empirically/referentially real things at a level of abstraction.

Let's look at a tangible example in linguistics.

Levels of abstraction and the Śivasūtras

Morphophonology has many different levels at play. We're generally familiar with the obvious fact that sounds have acoustic and articulatory traits which may come to bear, as well as language specific phonological variants and such. But there is sometimes a deeper relationship between sounds that at least *superficially* defies intuitive or empirical explanation, but on closer inspection make total sense.

We can take the earliest example from the Śivasūtras of Pāṇini (n.d.). The Śivasūtras are a simple list of sounds of the Sankstrit language categorized by Pāṇini in a way that may seem partially arbitrary to the trained linguist.

The goal of these sūtras is sheer *lāghava* (economy) in Pāṇini's rule formulation. That is, instead of saying sounds $x, y, z...$ turn to $a, b, c...$ in some given context, we can say that sounds of Line X turn into sounds of Line A, etc. This is not dissimilar to the idea of us saying that something [-voiced, +consonantal] becoming [+voiced] in between [-consonantal] sounds, although Pāṇini typically abstained from dirtying his hands with things so unwieldy as features directly.

We can reproduce the entirety of the Śivasūtras in the fourteen lines below.

1. a i u Ṇ
2. ṛ ḷ K
3. e o Ṅ
4. ai au C

5. ha ya va ra Ṭ
6. la Ṇ
7. ña ma ṅa ṇa na M
8. jha bha Ñ
9. gha ḍha dha Ṣ
10. ja ba ga ḍa da Ś
11. kha pha cha ṭha tha ca ṭa ta V
12. ka pa Y
13. śa ṣa sa R
14. ha L

Note that the letters rendered in majuscule at the end of each line are Pāṇini's abbreviations for the categories. Note also that there are some sounds, *h* for example which are included on multiple lines. This is intentional.

Now some lines show us what any linguist could identify as roughly "natural classes". Line 7 contains all the nasals. Line 13 contains sibilants. Line 11 contains mostly aspirates with some mysterious stragglers. We can see how Pāṇini can refer to these groups without saying the features directly.

Still there are some lines which seem to utterly vex the idea that these are natural classes. As said above, Line 11 is mostly aspirates, but not entirely. Some murmured consonants appear on Line 8, while others on Line 9.

This might seem like a decision made out of phonetic ignorance, but it's important to remember that the phonetic knowledge of Sanskrit grammarians was, in short, complete. They knew a voiced sound (nāda) from a voiceless one (śvāsa), and their places of articulation, aspiration and the rest.

The Śivasūtras are interesting because they present a level of abstraction deeper contemporary modern phonology. These categorizations can be thought of as being morphophonemic or historical in nature, because they unify sounds with very indirect and abstract relationship.

For example, why are the low and high vowels categorized in Line 1, while the mid vowels are categorized in Line 3? On superficial phonetic or phono-

logical terms, or in modern SPE-like notation, there's no clear reason why *a* should be with the high vowels rather than with *e* and *o* which are phonetically closer.

Now good Sanskritologists will know that Late Proto-Indo-European had all five of the cardinal vowels: */i, e, a, o, u/. Indo-Iranian languages, however are distinct in that */e, a, o/ all *merge* into /a/ in Sanskrit and sister languages. The Sanskrit /e/ and /o/ are of secondary origin: from the PIE diphthongs *ay and *aw, while the Sanskrit diphthongs /aj/ and /aw/ come from the PIE long diphthongs *āy and *āw.

Thus Lines 1, 3 and 4 all have historically distinct origins. If this were merely the case, it's hard to understand how Sanskrit grammarians uncovered this correlation, but the three categories actively alternate at an abstract level in the language at in different verbal and nominal paradigms, based on the PIE ablaut system.

As an example, the PIE root *(s)tew- yields the verb *tudéti "strike". The /u/ vowel is a result of the syllabification of /w/ due to loss of the /e/ vowel to the proterokinetic movement of the stress to the thematic vowel. The first syllable of *tudéti is treated as "zero grade" in the PIE ablaut system, and that *tud formant may vary with the full grade *tewd or the lengthened grade *tēwd in different verb forms or derived words.

These three forms in PIE give us precisely the variants shown in the Śivasūtras. *tudéti appears as *t**u**dáti* "he strikes" in Sanskrit (Line 1); its equivalent form in the perfect aspect from the PIE full grade *tewd is the second syllable in *tut**o**da* "he has struck" (Line 3) (remember that PIE *e > Sanskrit /a/ and Early Vedic /aw/ > /o/; the first syllable is a reduplicant). The PIE lengthened grade *ēw by the same predictable sound laws yields the vowel in the Sanskrit aorist: *at**au**tsīt* (Line 4)

t**u**dáti "stikes" tut**o**da "has struck" at**au**tsīt "struck"

The end result is that Pāṇini can refer to this historical and morphophonemic variation, even though he is unaware of it *per se*, by referring to the appropriate lines of the Śivasūtras. So, a root vowel appears in its form on

Line 1 in the present, Line 3 in the perfect, and Line 4 in the aorist, etc. (this is simplifying, ignoring other rules in Sanskrit). At that, since the three vowel series on Lines 1, 3 and 4 have different origins in PIE, we can make reference to them as a holon when describing other linguistic facts that correlate with them, say syllable structure.

Regardless, this kind of abstract, but superficially uninterpretable relationship not only formed the basis of Pāṇini's grammar, but was the input to de Saussure (1879)'s later theory of extended ablaut to what he called *coefficients sonantiques*, which would turn into what we know as Laryngeal Theory.

The level of abstraction adopted in the Śivasūtras thus is somewhat deeper than typical phonology, and apparently transcends the actual nature of the sounds, but it has in it the potential to describe alternations that are beyond the understanding of even the person (Pāṇini) who posited the relations between sounds. Pāṇini can, thus, be accurate without understanding, while we, with thousands of more years of Indo-Europeanism under our belts, can have both, since we understand that the formal choices made by Pāṇini happen to actually reflect an aspect of the morphophonology that's really real for an understandable reason.

Relevance

Let's be clear about the relationship of the Śivasūtras to the wider point. It's often a concern that, say, a neural net which solves for some phonological problem will rely on nodes that "don't make sense" in the traditional sense. We would like it for there to be a node that "means" voicing or corresponds to a place of articulation or perhaps an acoustic feature.

After a point of wideness in the data, however, the levels of abstraction needed for an efficient net will bypass the more intuitive or obvious aspects of the mind. In my parlance, while a referential model wants to make things intuitive and provide the actual mechanism of reality, after a point,

referential modeling becomes implausible.

Formal modeling, with an eye only for data economy (*lāghava* in Pāṇini's terms) becomes the only option, but the story doesn't end there. Formal modeling gives us things like the Śivasūtras: the realization that there may be a useful level of abstraction greater than what we understand how. While the different lines of the Śivasūtras may seem as a quasi-arbitrary categorization, deeper inquiry, in this case into historical Indo-Europeanism, reveals that there is in fact a principled reality to the relationships.

Formal modeling thus provided an important clue as to where to lead our next intuitions. If we design a neural net to account for some data alternation, it might be worth investigating if those mysterious nodes we *don't* understand don't have an external reality at a level of abstraction we don't yet have a concept of.

As a reminder, the Śivasūtras are not meant to be a logically derived set of sounds from exhaustive theoretical analysis. We know now that there is a historical bases for some of them, but this was beyond the possible knowledge of Pāṇini. As their names suggest, their arbitrary and phonetically and phonologically unjustifiable classification is traditionally thought of to simply be a divine gift to Pāṇini from Shiva himself, nothing more, nothing less. Pāṇini's grammar was an unassailable empirical accomplishment, built on an arbitrary and occult foundation. If that hadn't been the case, misplaced skeptical rigor wouldn't have let it survived so long.

References

de Saussure, Ferdinand. 1879. *Mémoire Sur Le Système Primitif Des Voyelles Dans Les Langues Indo-Européennes*. Leipsick.

Fodor, Jerry, and Zenon Pylyshyn. 1988. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28:3–71.

Oaksford, Mike, and Nick Chater. 1994. "A Rational Analysis of the

Selection Task as Optimal Data Selection." *Psychological Review* 101.4:608–31.

Pāṇini. n.d. "Aṣṭādhyāyi."

———. n.d. "Śivasūtāṇi."

Partee, Barbara. 1979. "Semantics – Mathematics or Psychology?" In *Semantics from Different Points of View*, edited by Bäuerle et al. Springer-Verlag.