Barry Smith

Model-induced Escape

A spam filter works like this. Your email provider, call him "google", collects data pertaining to how you deal with incoming emails. Each email is stored in the email system as a long string (call it "s") of 0s and 1s. Depending on whether or not you press "spam" when you open the email, the system will create a tuple of the form <s, 0> for "not spam", and <s, 1> for "spam". The resultant set of tuples (call it "tup") then provides the email provider with the information it needs to block those future emails which are *like* the emails which users earlier identified as spam.

To do this the email provider uses a kind of mathematics that was known to mathematicians such as Bošković, Legendre and Gauss already in the 18th century, but which has only recently come into wide usage via modern statistical learning whose powers have been disclosed thanks to the availability of massive computing power in today's computers.

Before statistical learning, mathematicians were constrained to do their work using only what we shall call "explicit mathematics", for which are required only a pencil, paper, and a waste basket. (Philosophers, it is said, manage without the waste basket.) Also required is a human brain, which enables mathematicians to think out explicitly what it is that they want to say, and then record the result on paper.

Modern statistical learning, in contrast, allows what we can call "implicit mathematics", which is a kind of mathematics that allows data about what is going on in the world to drive the creation of algorithms in a way which does not require any intervention of human beings. This implicit mathematics is the basis of practically all current work in artificial intelligence, and it works like this. First (in a manner which still looks very much like explicit mathematics) there is created inside the computer what is

called the "neural net training algorithm". This algorithm is then fed with data drawn from the real world. The latter are called "sample data" because they represent a *sample* from a much larger body of data. In the email filter example described above, this body of data would be something like the set of all the emails that will be received in the future on the email system in question.

Implicit mathematics is what takes place when this sample data is fed into the neural net software in a process that is called "training". This process is "implicit" in the sense that it performs its wizardry without human intervention. It yields as output what is from the mathematical point of view a gigantic polynomial function, which might involve billions or trillions of parameters.

Considering again our spam filter example, when this function is applied to each new incoming email, it yields as output either "1" (for what it filters out as spam) or "0" (when it lets the associated email through to the user).[1] In a case like this, artificial intelligence works well.[2] It can do so because it can learn the pat-

---

[1] We note in passing that for any specific algorithm of this size and complexity, it must remain a mystery to human beings how it yields its output. And since algorithms of this sort make up the bulk of contemporary AI, talk of "explainable AI" is at best misleading.

[2] Other even more impressive AI successes, for example in the field of protein folding prediction, show the tremendous power of contemporary AI. But it is noteworthy (though still seldom noted) that these successes are achieved only along certain narrow lanes, which means that they fail when it comes to emulating, for example, human intelligence. Jobst Landgrebe and I seek to explain why this is so in our *Why Machines Will Never Rule the World, Artificial Intelligence without Fear* (Abingdon, UK: Routledge, 2022), where we show that the now standard approach to the creation of AI software which we have described in the foregoing will work only in domains where it is possible to obtain sample data that are *representative* of the entire body of data from which the samples are drawn. Representative sample data do not exist wherever the numbers of variables governing the behavior of a system are large and wherever the system acquires over time new elements and new interactions – including new types of elements and interactions. These conditions, described at greater length in the book, hold in all domains where humans are involved, including medicine, finance, climate, agriculture, war, mating behavior, and many more.

terns characteristic of spam as they are encapsulated in tup, and it can, with a high degree of reliability, accept or reject emails arriving in the future according to whether or not they exhibit those patterns.

𝒥

But it can work well only for a while. And here entereth the problem of model-induced escape. For the evil authors of spam do not sleep. They are always and continuously seeking new ways to generate emails which will get through existing spam filters, in a process which gives rise to an arms race between (machine-assisted) authors of spam and (machine-assisted) authors of spam filters. In course of time, every source of spam emails will begin to escape the model encapsulated in any given spam filter.

Something similar arises where the creators of AI systems attempt to write algorithms that will enable them to predict, for example, the future price of oil. Here again, if ever such an algorithm were deployed in the market, other market participants would before long adjust their behaviour in ways that would start to falsify those predictions.

It is, similarly, impossible to produce a vaccine against the influenza virus which will be effective against this virus over the long term, because the virus itself mutates to evade the antibodies generated by each successive vaccine: a case of viral-induced escape.

𝒥

Which brings us to Nyíri.

In 1982 Nyíri writes a paper demonstrating convincingly that there are strong signals of a conservative strain of thought in the writings of Ludwig Wittgenstein.[3] This has initially only a tiny

[3] J.C. Nyíri, "Wittgenstein 1929-1931: Die Rückkehr". *KODIKAS/CODE – Ars Semeiotica* 4-5/2 (1982), pp. 115–136, abridged version as: "Ludwig Wittgenstein as a Conservative Philosopher", *Continuity: A Journal of History*, 8, Spring 1984, pp. 1–23. See also his "Wittgenstein's New Traditionalism", *Acta Philo-*

effect. But then a more significant effect sets in as the authors of Wittgenstein secondary literature draw attention to features of Wittgenstein which cast the conservatism thesis in a negative light.[4]

We have here a case of model-induced escape which can be understood along the following lines. Nyíri advances a model of the philosopher Wittgenstein that is designed to help us understand the latter's output. (Whether or not the model proffers a true picture of the relevant strands in this output is here not important.) The operation of the Nyíri model generates a reaction, in the form of new proposed models of the Wittgenstein corpus and of the events in Wittgenstein's life. In these new models, features of the latter which were either hitherto unnoticed or noticed but set to one side as insignificant, are now brought to the fore. The Nyíri model, which initially seemed so attractive, now appears questionable (and this again independently of whether or not the claims on which it rests are true).

𝒥

There is no such thing as email spam. Rather there is a flow of constantly mutating spam patterns.

There is no such thing as influenza. Rather there is a flow of constantly mutating viruses.

There is no such thing (no such *obiectum philosophiae*) as

---

sophica Fennica 28/1–3 (1976), pp. 503–512.

In subsequent writings, Nyíri has also demonstrated the philosophical potential of a reading of Wittgenstein along these lines for example in "Conservatism and Common-Sense Realism", *The Monist*, vol. 99, no. 4 (October 2016), pp. 441–456, and "Towards a Theory of Common-Sense Realism", in András Benedek et al., eds., *In the Beginning was the Image: The Omnipresence of Pictures: Time, Truth, Tradition*, Frankfurt/M.: Peter Lang Edition, 2016, pp. 17–27.

[4] For example, in Cressida Heyes, *The Grammar of Politics: Wittgenstein and Political Philosophy*, or in David R. Cerbone, "The Limits of Conservatism: Wittgenstein on 'Our Life' and 'Our Concepts' ", chapter 2 of *The Grammar of Politics*, Ithaca: Cornell University Press, 2003, pp. 43–62.

Wittgenstein. Rather, there is a constantly mutating set of interpretations of a certain body of work – a body of work in which nowadays – and again thanks to Nyíri[5] – those places where Wittgenstein deploys images are featured in the foreground to a greater degree than in ages past.

Perhaps this is what makes philosophy so problematic when viewed from the perspective of *results*, or in other words of signs of progress commonly accepted across the discipline. As some of the more impressively comprehensive contributions to the *Stanford Encyclopedia of Philosophy* demonstrate, there are, in philosophy, just too many ways of inducing escape from any given putative discovery; too many dimensions along which an interpretative or definitional arms race can be triggered.

FOLLOW-UP NOTE

Some further areas of application of the ideas on model-induced escape advanced in the foregoing are:

There can be no best whisky: If the proposition that McX's whisky is the best whisky becomes generally accepted then this will have multiple consequences which will undermine it, for example the consequent increased demand will make this whisky appear more popular and thereby undermine its status as being somehow exceptional.[6]

The David Lewis Syndrome,[7] manifested when philosophers bring extraordinary dialectical ingenuity to bear on behalf of completely implausible philosophical theses, will in the long run undermine the David Lewis Syndrome, as the piling up of ever more implausible philosophical theses undermines the methods used to achieve them.

---

[5] J. C. Nyíri, *Meaning and Motoricity: Essays on Image and Time*, Peter Lang GmbH, Internationaler Verlag der Wissenschaften, 2014.

[6] See W. David Marx, *Status and Culture: How Our Desire for Social Rank Creates Taste, Identity, Art, Fashion, and Constant Change*. Penguin, 2022.

[7] See https://leiterreports.typepad.com/blog/2004/03/busy_freud_davi.html.

How the new philosophical scholasticism is establishing itself (from Korsgaard's 2022 Dewey Lecture[8]):

> Young people are expected to produce an absurdly large number of papers, preferably published in refereed journals, in order to get tenure, or even in order to get jobs. … The papers are supposed to be blind reviewed, and these days many referees for journals require that papers should respond to the extant literature on the topic, whether responding to the extant literature enhances the author's argument in some way or not. Because the sheer mass of the literature is growing exponentially, people draw the boundaries of their specializations more and more narrowly, both in terms of subject matter and in terms of time. The extant literature necessarily becomes the recent literature, which is a philosophically arbitrary category. Big, systematic philosophy of the sort we find in Kant and Aristotle, philosophy that is responsible to the ways in which one's views in one area fit in with one's views about everything else, has become nearly impossible, because someone trying to do that kind of work would supposedly have to know the literature in too many areas.

[8] Christine M. Korsgaard, "Thinking in Good Company", The John Dewey Lecture Delivered on January 13, 2022, at the One Hundred and Eighteenth Eastern Division Meeting of the American Philosophical Association.