Sandra D. MITCHELL

Instrumental Perspectivism: Is AI Machine Learning Technology like NMR Spectroscopy?

The question, "Will science remain human?" expresses a worry that deep learning algorithms will replace scientists in making crucial judgments of classification and inference and that something crucial will be lost if that happens. Ever since the introduction of telescopes and microscopes humans have relied on technologies to "extend" beyond human sensory perception in acquiring scientific knowledge. In this paper I explore whether the ways in which new learning technologies "extend" beyond human cognitive aspects of science can be treated instrumentally. I will consider the norms for determining the reliability of a detection instrument, nuclear magnetic resonance spectroscopy, in predicting models of protein atomic structure. Do the same norms that apply in that case be used to judge the reliability of Artificial Intelligence deep learning algorithms?

Philosophers of science explore and explain how scientists acquire knowledge of nature. Most have agreed that we must give up oversimplified accounts of direct experience of "the given" (which is the English translation of the Latin *datum* or *date*) and overambitious requirements that scientific knowledge be restricted to claims that are universally true and exceptionless. As a result, many factors that enter into scientific practice have been exposed as relevant to our understanding of how knowledge of nature is constructed, how it is judged, and how it is used. For example, which observations are judged to provide reliable data? What features of phenomena are represented in an explanatory model? In which contexts and for what purposes will an explanatory model be adequate? To be sure, science is a product of human activity, both *causally*, through experience and experiment and *inferentially*, though logic, calculation, and simulation. What is investigated and how it is investigated, is shaped by decisions which are themselves dependent on and constrained by human pragmatic goals, like curing diseases, or understanding the expanse of the universe.

The question, "Will science remain human?" is posed in response to a worry that AI machines will replace scientists, and that something crucial will be lost if that happens. Stark examples driving this worry are found in the proliferation of deep learning strategies of AI: AlphaGo beating the world's Go champion, DeepMind's application to problems in healthcare (Fauw et al 2018), deep learning models for data reduction in high energy physics (Guest, Cranmer and Whiteson 2018) and bias in autonomous systems (Danks and London 2017).[1] But are these new technologies really different from what we have come to see as legitimate extensions or instrumental replacements of human capacities by what we now accept as less threatening machines? In this paper my strategy is to explore in what ways machine learning is similar to other scientific instruments, taking the results of instrumental engagement as providing a useful non-human perspective on the phenomena. If AI is understood instrumentally, then it is clear we use it (or not) for our own, human, purposes. But when should we use it, and when not? When should we trust it, or why not? I will suggest that the same norms that govern judgments of other scientific instrumental reliability should be used to warrant the use of AI. My argument is in support of the norms to be applied, rather than an account of the success of any particular use of AI in practice.

[1] AI, Machine Learning and Deep Learning are not identical. AI is a machine way of performing tasks that are characteristic of human cognition, but may or may not attempt to represent the way humans perform those tasks. Machine Learning is one set of practices to achieve AI, where the algorithm is not explicitly programmed to perform a task, but "learns" how to achieve a specified goal. Deep Learning is one form of Machine Learning that explicitly references human brains by using Artificial Neural Net structures, with many discreet layers (deep structure) of connected artificial neurons.

Ever since the introduction of telescopes and microscopes humans have relied on technologies to "extend" beyond human sensory perception in acquiring scientific knowledge. Simple instruments relying on lenses present mediated images to the human observer. This constitutes an indirect causal interaction between the scientist and the phenomena studied. Contemporary scientific experiments, like x-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), cryo-electron microcopy, and small-angle neutron scattering are used for predicting the three-dimensional structure of proteins, involve more complicated causal interactions in order to detect and process information about the target phenomena. Scientists trust these detection instruments, from simple lenses to elaborate experimental equipment, to reveal features of nature. Indeed, we must trust them more than unaided human detection.

Recently developed artificial intelligence technologies appear to "extend" beyond human cognitive capacities. Are these forms of outsourcing cognitive aspects of scientific practice similar to instrument-mediated perception? If not, how do they differ and should we be worried about their increasing role in science? I will take up this challenge by investigating AI from an instrumental stance. Does AI provide just another instrumental perspective for humans to use in gaining scientific knowledge, like microscopes and NMR spectroscopy? Are the means by which we trust the results produced by these "sensory" technologies transferrable to the results produced by AI "cognitive" technologies?

Routes to scientific knowledge

Science aims to accurately characterize features of nature that permit explanation, prediction and intervention in order to further our human goals. Support and justification for scientific knowledge comes from experience (observation or experimentation) and reason (concepts, logic and inference). I argue that the theoretically and experimentally based models that result from well-executed scientific practices always encode a limited perspective. Much has been written about the perspectival character of scientific instruments and models of natural phenomena (e.g. Giere 2006, Van Fraassen 2008, Massimi 2012, Price 2007). Some appeal to the location of the "observer", i.e. the vantage point of a distance or scale from which structures can be detected. I argue that perspectivism follows from the partiality of representation itself. My argument rests on the claim that no scientific model, whether it is derived from more general theories or from the results of an experiment, can provide a complete account of a natural phenomenon.

What could be meant by model completeness? Completeness in formal systems, like the propositional calculus, is tied to notions of proof and deductive inference. A set of axioms is complete if every theorem can be derived from it by the specified rules of inference. But what could it mean for a scientific model of natural phenomena to be representationally complete? Weisberg (2013) suggests that model completeness is a representational ideal referencing the inclusiveness of the model ("each property of the target phenomenon must be included in the model." P. 106) and fidelity (models aim to represent "every aspect of the target system and its exogenous causes with an arbitrarily high degree of precision and accuracy" p. 106). As

Weisberg acknowledges, this type of completeness is impossible to achieve.[2] Rather than reject completeness as a virtue of a model, Weisberg claims that we should treat it as a regulative ideal against which we judge the success of any given scientific model. Since no scientific model can satisfy the standard, we instead focus on how close or far from it a model comes. I disagree with this approach. As I have argued in the case of ideal, universal, exceptionless laws, (see Mitchell 2000, Mitchell 2009), we should develop normative standards that track the character of what can be accomplished, that is, what scientists in fact do. More or less complete often will be unmeasurable when we are considering models the use different variables that do not stand in inclusive hierarchies. Since all scientific models are partial, and since many will represent differing features, how would we determine which one of them was "more" complete? Counting the number of variables clearly will not be adequate.

Even if we came up with a way of measuring more or less complete models of natural phenomena, satisfying Weisberg's completeness ideal is neither necessary nor desirable for successful science. Not every describable feature of a system in every possible degree of precision is required for identifying features and relations that permit prediction, explanation, and intervention on that system. Suppose we could meet a strong completeness standard whereby our model represents each property of the target phenomena (at all spatial and temporal scales) with the highest degrees of precision and accuracy. That representation would fail to constitute usable knowledge of the phenomenon; it would be a duplicate. For the

[2] See also Madden 1967 "The incompleteness of science arises from the impossibility of describing every detail of nature, whether the universe be conceived as infinite or finite in space and time, and from the fact that any explanatory deductive system depends upon assumptions which are themselves not explained" Madden review of Schelgel POS 1967

purposes of facilitating explanation, prediction, and intervention, it would be no better than engaging directly with the very system we are trying to understand. Model "goodness" should be judged by its *accuracy* with respect to existing empirical data, and its *adequacy* with respect to specific goals, not how close it comes to an unachievable and non-useful "ideal." The assumption that if we could represent everything then we can achieve any and all of our goals is undoubtedly the intuition supporting completeness as an ideal. However, given we cannot represent everything, including more details in a model can be detrimental to both for its accuracy (e.g. by over parameterizing in ways that compound uncertainty) and its adequacy (e.g. by obscuring main factors whose manipulation might be sufficient to meet the goal).

A model represents relations that provide explanations and predictions. Since it cannot be complete in the sense of including all that could be described, it must be partial. Scientific models expose what is deemed causally relevant, what is salient, what is expressible in a particular framework, etc. What is represented and what is left out sometimes is guided explicitly by explanatory or pragmatic goals. Scientific representations also reflect limitations of the representational medium. Models are abstract and, even when accurate, are not complete. There are two consequences of this fact. First, the partiality of scientific models requires us to embrace model pluralism as essential to science achieving its goals. Second, the partiality of scientific models entails perspectivism.

> *Premise 1*: A useful representative model can include only some aspects of its target phenomena.
>
> *Premise 2:* Therefore other aspects that might have been represented are omitted.

Premise 3: The omitted aspects could be (and typically are) included in other scientific models.

Premise 4: Some sets of representational models are not incompatible, nor intertranslatable/reducible, nor additive/mutually exclusive.

Conclusion: To explain, predict, and intervene on a given phenomenon, science may require a plurality of models to represent the features that are relevant in different contexts and for different purposes.

This raises new questions about the relationships among the multiple models that are developed to represent the same phenomenon. While reduction, unification, and elimination are ways in which models of the same phenomenon may be related, I have argued that when there are substantial instances of compatible pluralism explanatory integration better accounts for model-model relationships (Mitchell 2009, and forthcoming).

The flip-side of partiality of representation is perspectivism. One representational model, by leaving out some features or details explicitly and implicitly "selects" features to include. What is left in constitutes a perspective. Given there is no unique, complete representational model, there can be and frequently are multiple models that all satisfy the empirical demands science places on acceptance. The accuracy of a model is judged by its ability to account for accepted empirical data. Multiple models can provide equally accurate descriptions of a phenomenon, though they do so by referencing different features of it, perhaps at different scales, with different degrees of precision, etc. The adequacy of any model or set of models is judged by how well it or they serve some epistemic or non-epistemic purposes. Human purposes, as well

as human capacities are reflected in both these judgments. How should we evaluate AI deep learning algorithms? Surely, AI provides new tools for developing models, which describe and predict features of nature. These models are partial and perspectival. As such, they cannot provide a complete model, though they may provide new perspectives that no human could produce.

The New Technologies

Machine learning includes a variety of computational algorithms that detect patterns, or make inferences from data, without applying explicit, human programmed rules specifically designed to solve the problem at hand, rather they implement generic "learning" algorithms. From minimal information (a training set of input-output patterns bearing varying degrees of human-assigned labels) the machine builds its own models and new algorithms for making predictions from data in a specific domain. Artificial Neural Networks (ANNs), massively parallel systems with distributed computation were modeled on biological neural architecture in the brain. Research on ANNs has been a growth industry since the 1980s, due to, in part, the development of back-propagation learning algorithms for multilayer feed-forward networks that had widespread impact through Rumelhart and McClellan's 1986 book. While feed-forward networks are static, producing one set of output values for a given input, recurrent, or feedback network architectures are dynamic, where computed outputs modify the inputs. ANN's learn from examples rather than explicit rules. This means that, among other things, connection weights between the neurons are adjusted by means of a learning rule. For example a back-

propagation algorithm can implement error-correction to train the network to minimize output error (Jain, Mau and Mohiuddin 1996).

A human signature is ineliminable from all forms of machine learning. In so-called supervised learning, humans determine not just what the prediction problem is, but specify what counts as a "correct answer", the target, that is used as part of the training process. The machine learns how to reach the target by calculating an error signal between the target and actual outputs, and using that error to make changes in the weights in the algorithm. Unsupervised machine learning, infers a function from unlabeled input to output that relies on hidden structure. In both cases, what data is presented, and what problem is to be solved is set by humans. The feature that makes artificial neural net machine learning (ANNs) a challenge is that the functions it "infers" to map input data into output patterns, and how the functions are acquired may not always be cognitively available or meaningful for humans. If AI machine learning algorithms do not learn the way humans learn, do not make inferences or patterns the way humans do, and we cannot see what it is doing, then how can we trust it is doing the right thing?

This seems to me to be parallel to the situation of causal detection instruments. They detect features of phenomena that we cannot detect, and they do it in ways that are different from how we do it. Yet we trust the results of such instruments. How is the perspective of AI machine learning different from causal experimental perspectives?

How is AI machine learning different from experimental instruments?

The Instrumental Stance

In commendation of y^e Microscope

Of all the Inventions none there is Surpasses
The Noble Florentine's Dioptrick Glasses
For what a better, fitter guift Could bee
In the World's Aged Luciosity.
To help our Blindnesse so as to devize
A paire of new & Artificial eyes
By whose augmenting power wee now see more
Than all the world Has ever doun Before.

Henry Power 1664 (Cowles 1934)

Power, one of the first scientists to be made a fellow of the Royal Society wrote the first book about microscopes (predating Hooke's *Micrographia* by two years). Power's message – that artificial eyes let us see more than anyone could have seen before continues to be true of the modern forms of spectroscopy. These, like X-ray crystallography and Nuclear Magnetic Resonance Spectroscopy, are less clearly extensions of human visual perception, than they are alternative detecting devices. Why do we trust the results of such instruments to provide data from which scientific inferences can be drawn? In what follows I will consider the use of these experimental instruments in the determination/prediction of protein structure. There are two components to trusting instrumental detection; one is reliability of the data output that result from the causal interaction with the target phenomenon; the other is the inferential or epistemic warrant that the measurements or models of data provide in supporting hypotheses about the phenomena.[3]

[3] See Bogen and Woodward 1988 for an important distinction between data and phenomena.

Briefly, proteins are the most common molecules found in living cells. They consist of one or more polypeptide chains which themselves are composed of amino acids. Proteins are coded for by DNA and produced through a process of transcription of RNA from DNA, then translation of the RNA on a ribosome to generate a chain of amino acids (the primary structure of a protein) that then folds into a functional conformation, of secondary, tertiary and sometime quaternary structure. Predicting the structure aims to identify the position of the atoms constituting the amino acids in their folded, functional form. Knowing the structure provides information about binding sites on a protein that are a clue to its function, since most proteins perform their biological function in response to or in conjunction with other molecules. This information can also aid in drug design to intervene on proteins for medical purposes. NMR and X-ray Crystallography are the two primary methods for experimentally ascertaining protein structure.

Obviously, humans cannot directly detect what is going on at the scale of atoms. NMR spectroscopy is an experimental alternative causal detection method that measures the way magnetic influences affect the behavior of the nuclei of atoms (e.g. hydrogen). Basically, a concentrated protein solution is placed in a strong magnetic field. Atomic nuclei, like hydrogen, have an intrinsic magnetization resonance or spin that is changed by the strong magnetic field. In the experiment, the initial alignment with the strong magnetic field is disrupted by a radio frequency (RF) electromagnetic pulse. As the hydrogen nuclei return to their aligned states, they emit RF radiation that is measured. The radiation emitted depends on the local

environment so that excited hydrogen nuclei in other amino acids induce small shifts in the signals of close-by hydrogen nuclei (magnetization transfer). Given information about the protein's constituent amino acid sequence, the measurements provide information about where each atom of each amino acid is located in the 3-D structure of the protein.

As Hans Radder puts it "An experiment tries to realize an interaction between some part of nature and an apparatus in such a way that a stable correlation between a feature of that part of nature and a feature of the apparatus is produced." (Radder 2003). The first level causal output in NMR protein spectroscopy is the RF radiation associated with hydrogen nuclei in the various amino acids composing the protein sample. From measuring the decay curves of the hydrogen atoms, the experimenter can recover information about the relative distance and rotational angles between atoms. From that, plus measurements of other types of atoms in the protein, an atomic structure of the protein can be inferred. What are our grounds for trusting the results of the instrument? We don't have any experience of "what is like to be a nuclear magnetic resonance spectrometer". Though we can decompose the causal process into more fine-grained steps, at some point we will not have any more direct perception. Thus, reliability of the causal process will be an inference we make – not a causal "experience" we have. That inference is made by appeal to theories, the "theory of the experiment" and the stability of the results of instrumental replication and multi-instrument convergence.

I will consider how this is achieved in the case of NMR experiments and question whether or not the same types of inference are available in the case of AI machine learning.

Theoretical support

Instruments have perspectives, i.e. they selectively interact with some features of the target phenomena, not all, and in this sense, are causally "biased"(see Giere 2006). Measurements or some other meaningful representation of the causal effects of an instrument/phenomenon interaction encode theoretical assumptions, and are not perfectly objective or "given" by experience. Numbers, graphs, and natural language represent the causal process in terms that express units, scales, and other content. As Tal (2017) clearly and correctly claims "To attain the status of an outcome, a set of values must be abstracted away from its concrete method of production and pertain to some quantity objectively, namely be attributable to the measured *object* rather than the idiosyncrasies of the measuring instrument, environment, and human operators." (p. 35). With respect to NMR experiments, we want to assign measurements (relative distances) to the atoms of the protein itself. How much, which and where theories are involved in detection matter to our judging the process and the outcomes as a reliable means to do that. In NMR determination of protein structure experiments, theories of how electromagnetic fields and pulses affect the behavior of atomic nuclei are required both for designing the experiments and interpreting their results. Additional theories are also required pertaining to the materials used in building the instrument, how the preparation of the sample might affect the target properties, the confounding influences of the environment in which the experiment is conducted and more. Theories are essential for both performing an experiment – the causal theory of the experiment – and for producing epistemically relevant information. We might require different causal operations to elicit measurements relevant to determining the

structure of particular proteins, which are biologically functional only in conjunction with other molecules. We might need different sorts of information, and hence experimental access, to design drugs that will bind with a protein to negate a detrimental function it would otherwise perform.

The role of theoretical assumptions in experimental hypothesis testing long has been a subject of philosophical scrutiny. Concerns about theory-ladenness challenge the "objectivity" of experimental results for testing theories. But theories are required. Duhem points out that "The same theoretical fact may correspond to an infinity of distinct practical facts….The same practical fact may correspond to an infinity of logically incompatible theoretical facts…." Duhem (*Aim and Structure*, p 152). Thus, for an experiment to be a test of a specific hypothesis or prediction, there needs to be some way to translate the causal outcome (practical fact for Duhem) into a claim relevant to the prediction of the hypothesis being tested (theoretical fact). There is no escaping *some* semantic infection from the theory being tested in describing the outcomes of an experiment. [4] There is no way to isolate an observation or measurement from the theoretical assumptions. But which theories are involved?

In its most extreme version, what is called Duhemian holism, it is claims that every experiment implicates "a whole theoretical group" (Duhem 1906, 1862, p. 183) and thus no isolated hypothesis can either be confirmed or falsified by an experimental result. There is always the

[4] There are different forms of theory ladenness. See Bogen's SEP article distinction of perception loading, semantic theory loading, and salience. On Bogen's classification, Duhem's claim is about semantic theory loading.

possibility that a negative test result is due to one of the auxiliary assumptions, and not the theory under test. This has challenged the "objectivity" of experimental results to both accurately reflect the features of the phenomena and provide test for accepting or rejecting individual hypotheses.

Hasok Chang (Chang 2004) argues, in his examination of the history of the thermometer, that scientists can avoid the worst forms of holism by adopting a "principle of minimalist overdetermination". For Chang, overdetermination is the agreement by different methods on the measurement value ascribed to some phenomenon, e.g. a temperature determined by both calculation and measured by a mercury thermometer, or measured by two different types of thermometer. The necessity of invoking auxiliary hypotheses in order to make predictions, build apparatus and interpret the results of an experiment is unavoidable, but by minimizing assumpitons, Chang argues, the damage can be contained. "The heart of minimalism is the removal of all possible extraneous (or auxiliary) non-observational hypotheses." (Chang 2004, p. 94). Overdetermination by multiple experiments which make similar ontological assumptions about the nature of the phenomena, uses fewer assumptions than are required by predictions from high-level theories. His approach rests on the appeal for the most direct, or least theory-mediated correlation between what is in the world and the measurement. I understand this principle to require that the theories of the experiment should rely on *as few assumptions* about the function that associates the target feature with the measured feature as possible. Chang's principle acknowledges that there is not way to eliminate theoretical assumptions in generating experimental results, in contrast to others who have required something stronger,

namely the independence of the theory of the instrument and theory to be tested by the experimental results using the instrument. Complete independence is certainly too strong, since there needs to be some form of translation between experimental measurements and theoretical predictions in order for the former to be a test of the latter at all (Darling 2002). By minimizing, removing all possible auxiliary assumptions, Chang suggests we can provide the strongest grounds for taking experimental results to be confirmations or refutations of the hypothesis being tested.

The theory of the instrument does not need to be simple, and certainly is not in the case of NMR spectroscopy, but is should, if Chang is correct, rely on no more theory overlapping the hypotheses to be tested than is necessary. NMR spectroscopy relies on theories of nuclear spin and the effects on spin from magnetic influences (chemical shift). While the atoms in each functional protein are described by these theories, which particular arrangement of molecules occurs in a given protein is not. So although NMR spectroscopy experiments may succeed or fail to correctly detect the atomic locations, the assumptions involved in addition to the theory of the instrument seem to be minimally overlapping with a hypothesis that protein P has conformational structure C. At least on Chang's account, it could be argued that NMR experiments are minimally overdetermined.

Chang's approach has the virtue of recognizing the ineliminabilty of influence of theoretical assumptions related to the hypothesis being tested in acquiring reliable measurements from any experimental set up. However, like Weisberg, his approach seems to take the impossible –

complete independence – as a regulative ideal. Minimizing means using fewer assumptions, thus getting closer to independence. However, as in the case of Weisberg's ideal completeness ideal, reliance on the impossible-to-realize ideal norm of independence strikes me as an inappropriate goal. Instead, what matters is *what* is assumed in an experimental set up, not *how much*.[5]

In the case of NMR for protein structure prediction assumptions invoke general theories about atomic nuclei and chemical steric constraints (what angles of rotation are possible, and that no two atoms occupy the same space at the same time) and a host of other theories. However, the specific location of atoms in the conformation of a functional protein in solution is not directly implicated in those theories. NMR can be trusted in testing rival structure predictions since NMR is not tuned to a specific protein structure but only to nucleic atomic behavior more generally. A more detailed explication of the theories involved in all the steps of generating measurements in NMR spectroscopy, akin to what Tal (2017) calls "white-box calibration" in contrast to "black-box calibration" would be required to convincingly make the case (see also Humphries 2004). At some point, however, there will be no more sub-boxes between the phenomenon and detector that can be theoretically unpacked. There will just be input (some sample of the phenomenon) and an output. This is the rawest sense of data. With no theories left to support the veridicality of the correlation between phenomenal feature and instrument detection, we appeal to stability of the results to warrant claims of reliability.

[5] See Glymour 1980 for other ways to manage the theory-ladenness of experimental observations.

Replicability and Convergence

A theory or model of the experiment itself describes and predicts the causal relations between the target phenomenon and the output of the experiment. It involves theoretical constraints as well as sources of random and systematic error. Even if we have good grounds for accepting that the theory of the experiment is correct, there remains a question about if our experimental apparatus instantiates what the theory describes.

NMR spectroscopy, based on theories of nuclear magnetization had been attempted in 1936 to testing Lithium compounds. However, the experiments were unable to produce any measurable signals. In 1941 it was reported that NMR signals of hydrogen in water had been detected, but there was insufficient replicability of the results. It wasn't until 1946 that the first successful results of NMR were reported by two labs, by Bloch at Stanford and Purcell at Harvard for which they were jointly awarded the Nobel Prize in 1952. There were many developments that refined the instrument; importantly Richard Ernst's 1964 introduction of the use of short, intense RF pulses to simultaneously excite all magnetic resonances and using Fourier transforms to computationally analyze the response (H. Pfeifer 1999). The RF technique increased the sensitivity of the instrument thus improving the signal to noise ratio. The first NMR spectrum of the protein ribonuclease was detected in 1957. In 1958 Kendrew and Perutz published the first high-resolution protein structure, using by X-ray crystallography. In 1970 Wuthrich published a lecture showing that NMR spectroscopy and X-ray studies of proteins could yield similar spatial structures of protein under extremely different experimental

conditions (NMR of proteins in solution and X-ray of proteins in a crystalline form). This is taken to be the pivotal moment for NMR in the history of protein experimentation. (Schwalbe 2003).

For an instrument to be reliable is for the output of its detection process to capture the targeted features of the phenomenon thus permit accurate measurements that represent those features. But since we have no direct access to the phenomenon outside of some detection device, there is no way to directly compare the phenomenon with the instrumental response. Instead, instrumental reliability is supported by the stability of instrumental results in different places and times using the same experimental protocols (replicability) and stability of results among very different kinds of experiments (reproducibility or convergence). For NMR to reliably indicate the atomic structure of a protein is for the signals detected (chemical shift) and measured (decay curves) to reflect the distance and rotational angles between atoms present in the molecules constituting the protein.

The comparison of different instrumental techniques, like X-ray and NMR spectroscopy, permits cross-validation. If there is a designated standard technique with which to compare a new instrument, this is called calibration, if not, then multiple experiments each with different sources of systematic errors can be taken to be validating each other. Notice that just as there is no independent-of-theory test of a single measurement, neither is that any independent-of-theory calibration. Eran Tal (2017) has recently argued that successful calibration depends on two conditions. First the measurement outcomes “are mutually consistent within their ascribed uncertainties” and second “the ascribed uncertainties are derived from adequate

models of each measurement process" (2017:43). Since the conditions are different, and the models are driven by different theories, or parts of a theory, Tal argues that the shared results are context invariant. This, Tal suggests, is what can be meant by the objectivity of the result. They are not theory independent, but are independent of any particular theory. Replication under the same experimental protocol can support claims of internal stability – the phenomenal feature-instrument interaction is generating the signal measured, not a signal from some fluctuating or random extraneous source. Convergence can support a broader warrant. Tal suggests it permits prediction of experimental outcomes for other types of measurements with other sources of bias and uncertainty. As Tal claims, convergence accounts for "how it is possible to assess the reliability of measuring instruments despite the inaccessibility of 'true' quantity values, and despite the fact that measurement standards do not provide absolutely exact, infallible quantity values." (Tal 2017: 45). Stability across replication and convergence plays the same role as Chang's principle of minimalist overdetermination. However instead of minimizing the number of assumptions required for an experiment, on Tal's account, reliability is gained by varying the sources of uncertainty, no matter how many assumptions each source requires.

To sum up: trust in the reliability of the models inferred from the causal processes of experimental instruments derives from our theories of the instrument (how is nuclear spin affected by magnetic influences) and from the stability of the results across multiple trials (replication) and multiple instruments (convergence).

AI instrumental perspectives

I have presented a view that the inferences supporting the reliability of the causal capture of phenomenal features through experimentation is supported by theories of the experiment and by the stability of results in replication and convergence. Are AI machine-learning practices subject to the same tests for reliability? How do they fare?

The instrumental stance focuses attention on the causal interaction between the target phenomena and the detecting device in the experiment that can output a measurement. Note that the measurement is a description that depends on interpretation and inference. In particular the measurement (or model of the data) involves distilling the signal from the noise in the initial data output. There are a variety of techniques to accomplish this in NMR protein experiments, from averaging out random error to 'correcting' for known systematic errors. In addition the measured data must be represented in a way that can be related to the hypotheses at issue, i.e. decay curves in NMR experiments are translated into relative distance and rotation angles between atoms.

Instrumentally, machine-learning algorithms enable identification of associations and patterns in observed data, permitting scientists to build explanatory models and make predictions. The algorithms do not have explicit programmed rules that are applied directly to accomplish the pattern recognition, but rather "learn" rules based on past "experience". There are many different protocols for the type of learning implemented (reinforcement, supervised) and the character of the algorithms (linear regression, logistic regression, etc.). The point at issue is

that how AI learns and then generates patterns and predictions seems to diverge from the way a human scientist learns and generates patterns and predictions. For some architecture of multiple layer (deep) neural networks, there may be no way in which a human can recover the rules that have been learned. Thus, even though the architecture of ANNs is inspired by analogy to the way humans reason, it is difficult to know if they actually reason in the ways humans reason. There is some evidence (Dodge and Karam 2017) of significant differences between deep neural nets and humans in recognition performance of distorted images in success rates and in types of errors.

In the case of AI, the "phenomena" are the data sets fed into the computer and the output is a pattern or a prediction. The output is more like a measurement or model of the data than a physical causal effect, as is the case for the spectroscopy instruments. What the learning algorithm does is eliminate noise and minimize error. In the NMR experiment the model of the data is based on both theories of the instrument and on the stability of repeated outputs.

> Most modern NMR experiments rely heavily on the reproducibility and stability of the spectrometer, because in order to select those signals that carry useful chemical information, it is necessary to cancel those that do not…. In some experiments…the sensitivity of the technique is often determined not by the intrinsic signal-to-noise ratio of the instrument, but by its stability. The limiting factor is the ability to distinguish between the signals of interest and a background of unwanted signals derived from instrumental errors, rather than a background of random noise." Morris (1992)

I suggest that the warrant of the output of machine learning is similar to the warrant of the measurement output of a causal detection device. Both rely on a theory of the instrument and

the stability of the results. To assess reliability, one needs an analysis of the learning rules implemented in AI algorithm. What is the support for reinforcement learning vs. supervised learning relative to particular problem types? Just as in physical experiments, the theories "behind" the instrument in AI machine learning are subject to the norms of contrastive confirmation that apply across science. Which learning rules, representation protocols, etc. are theoretically warranted to solve specific kinds of problems cam provide support for the reliability of AI.

In addition, it is clear that the stability of results through replication and convergence are also sources of warrant in the case of AI. In a recent *Science* article (Hutson 2018), a "replication crisis" was announced for artificial intelligence. Scientists who wanted to test a new algorithm against a benchmark, a classic example of calibration, found they could not replicate the benchmark itself. Replication is a norm that AI scientists clearly adopt. Making explicit how a detector instrument is built or a program is designed will accomplish two things. First, a replication can be attempted, checking the stability of the results across different iterations in space and time. Second, the assumptions implicit in the instrument or program can be made visible and the theories underlying them can be endorsed or challenged.

Other sources of failure to replicate include not having access to the training set upon which the algorithm learns the functions that execute predictions on new data input. Learning from different data sets can generate major differences in output predictions. (Nishikawa et al 1994). Other factors that are not central to the code may also influence the stability of the outcomes

(see Islam et al 2017). Factors in NMR experiments that are not directly in the source-signal-detector path also can influence outcomes, like input electricity source. Although there is no way to eliminate factors that disrupt the fidelity of the source features to the detected features, there are strategies in both NMR and in AI to identify and manage what may bias the results. Replication and convergence of results in each case can warrant the reliability of the instruments.

That the failure of reproducibility in AI is deemed a 'crisis' indicates that the sources of reliability and trustworthiness typical for "extensions" of our perceptual means of acquiring knowledge of nature, namely instrumented detection devises, also apply to the new "extensions" of our cognitive components of acquiring knowledge of nature, namely machine learning AI programs. In the case of NMR spectroscopy, the way it "sees" the world is not the same as the way we do and yet we can trust its results based on robust theories of the experiment and evidence of the stability of the output. The way AI machine learning detects patterns in data is also not the way we do it. Like the NMR case, artificial intelligence deploys capacities that are beyond our unaided human perceptual and cognitive abilities. Are there differences in how we can interrogate the causal instrument and the cognitive instrument to apply the norms required for judgments of reliability? Certainly yes. However that the same general norms for reliability apply to both types of scientific practice supports taking an instrumental stance towards AI. It is a tool we can use, if reliable, in the same way we use NMR spectroscopy or other tested and trusted scientific instruments.

References

Bogen, James and Woodward, James. 1988. “Saving the Phenomena.” *The Philosophical Review* 97:303-352.

Bogen, James, (2017) "Theory and Observation in Science", *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2017/entries/science-theory-observation/

Chang, Hasok (2004) *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Cowles, Thomas (1934) “Dr. Henry Power’s Poem on the Microscope”, *Isis*, Vol 21, No 1, pp. 71-80.

Danks, David and Alex London (2017) “Algorithmic Bias in Autonomous Systems” *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Pages 4691-4697. https://doi.org/10.24963/ijcai.2017/654

Darling, K. M. (2002) “The complete Duhemian underdetermination argument: scientific language and practice” *Studies in History and Philosophy of Science* 33, pp. 511-533.

De Fauw, J, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O’Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, (2018) “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nature Medicine* vol. 24, pp. 1342—1350,

Dodge, Samuel and Lina Karam (2017) “A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions”, *2017 26th International Conference on Computer Communication and Networks (ICCCN)* pp. 1-7

Duhem, Pierre [1906] 1962. *The Aim and Structure of Physical Theory*. Translated by Philip P. Wiener, New York Atheneum.

Giere, R. (2006) *Scientific Perspectivism* (Chicago: University of Chicago Press).

Glymour, Clark N. (1980): *Theory and evidence*, Princeton University Press

Guest, Dan, Kyle Cranmer and Daniel Whiteson (2018) "Deep Learning and Its Application to LHC Physics", *Annu. Rev. Nucl. Part. Sci*. 2018. 68: pp. 1-22.

Humphreys, Paul (2004) *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press

Hutson, Matthew (2018) "Artificial intelligene faces reproducibility crisis", *Science* 15 Feb 2018, Vol 359, Issue 6377, pp 725-726.

Islam, R.; Henderson, P.; Gomrokchi, M.; and Precup, D. (2017)"Reproducibility of benchmarked deep reinforcement learning tasks for continuous control." *ICML Reproducibility in Machine Learning Workshop*

Jain, A. K., J. Mao and K. M Mohiuddin (1996) "Artificial Neural Networks: a Tutorial", *Computer*: pp. 31-44.

Madden, Edward H. (1967) "Book Review of Richard Schlegel. Completeness in science. *Philosophy of Science*, pp. 386-388.

Massimi, M. (2012) "Scientific Perspectivism and its Foes", *Philosophica* 84(2012) 25-52

Mitchell, Sandra D. (forthcoming) "Perspectives, Representation and Integration" in in M. Massimi and C.D. McCoy (eds.): *Understanding Perspectivism: Scientific Challenges and Methodological Prospects*, Taylor & Francis

Mitchell. Sandra D. (2009) *Unsimple Truths: Science, Complexity and Policy*, Chicago: University of Chicago Press

Morris, G. A. (1992) "Systematic Sources of Signal Irreproducibility and t_1 Noise in High-Field NMR Spectrometers", *Journal of Magnetic Resonance* 100, pp 316-328.

Nishikawa, R. M., Giger, M. L., Doi, K. , Metz, C. E., Yin, F. , Vyborny, C. J. and Schmidt, R. A. (1994), Effect of case selection on the performance of computer-aided detection schemes. Med. Phys., 21: 265-269.

Pfeifer, H. (1999) "A short history of nuclear magnetic resonance spectroscopy and of its early years in Germany" *Magnetic Resonance in Chemistry*, 37, pp. S154-S159.
Pittsburgh: University of Pittsburgh Press, pp. 1-18.

Price, H. (2007) "Causal perspectivalism", in R. Corry and H. Price (eds.) *Causation, Physics, and the Constitution of Reality* (Oxford: OUP).

Radder, Hans 2003 "Toward a more developed philosophy of experimentation" in Hans Radder (ed.) The Philosophy of Scientific Experimentation.:

Rumelhart, David, James L. McClelland, and the PDP Research Group (eds., 1986): *Parallel Distributed Processing: explorations in the microstructure of cognition*. MIT Press, Cambridge.

Schwalbe, H. (2003): Kurt Wüthrich, the ETH Zürich, and the Development of NMR Spectroscopy for the Investigation of Structure, Dynamics, and Folding of Proteins" *ChemBioChem* 4, pp. 135-142.

Tal, E. (2017) "Calibration: Modeling the Measurement Process", *Studies in History and Philosophy of Science* 65-66, pp 33-45.

van Fraassen, B. (2008) *Scientific Representation. Paradoxes of Perspective* (New York: OUP).

Weisberg, Michael (2013) *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.