AI as statistical methods for imperfect theories

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Science has progressed by reasoning on what models could not predict because they were missing important ingredients. And yet without correct models, standard statistical methods for scientific evidence are not sound. Here, I argue that machine-learning methodology provides solutions to ground reasoning about empirically evidence more on models' predictions, and less on their ingredients.

Science uses false models as means for truer theory [Wimsatt, 1987]. How can statistical tools ground valid reasoning on empirical evidence without true models? Generalization is the key. Here I develop the argument that, unlike popular belief, reasoning from black-box models is good for science, because it builds the validity of inferences on prediction of observables.

1 Science has progressed by refining relevant constructs from wrong models

1.1 Observing motions of bodies, working out laws of physics

Early scientists, such as Aristotle, did not conceive mechanics in terms of acceleration and forces. Rather, they thought in terms of natural motion of objects, proportional to their weight. The notion of force made its way, as discussed by Ibn Sīnā, but motion was seen as proportional to external forces. The Copernican revolution motivated the importance of acceleration. Increasingly precise astronomical observations led to formulate planetary motion as elliptical trajectories. Scientists such as Kepler were seeking simple phenomenological rules, "harmonies" in his words, to explain the observations, *eg* that across the different planets the square of the period is proportional to the cube of the major diameter of the orbit. By introducing acceleration via differential calculus, Newton could propose laws of mechanics that explained observations of both celestial and earthly motion.

The birth of Newtonian mechanics illustrates how better *observations* and *statistical models* lead to better theories, even when starting *without the right theoretical framework*. It shows how *new ingredients* may be needed, such as introducing the construct of acceleration. It shows that progress is driven by seeking theories that *generalize* across many settings. The importance of acceleration was revealed by uniting motion of bodies on Earth and in astronomy. Indeed, as friction is ubiquitous on Earth, applying a force to an object often leads to a velocity roughly proportional to this force.

Later, better observations called for new frameworks, quantum or relativistic. Irregularities in the orbit of Mercury were first explained by adding a planet to the solar system, Vulcan. But observations of this planet turned out to be flawed, and the irregularities in Mercury's orbit are now understood as relativistic corrections. The Vulcan hypothesis illustrates how theoretical frameworks shape the interpretations of empirical results: observations are "theory laden" [Boyd and Bogen, 2021].

Today, the fundamental laws of physics are incredible precise. Are phenomenological models still important for their empirical validation? From a statistical perspective, the Neyman-Pearson lemma

Submitted to NeurIPS 2021 AI for Science Workshop.

tells us that the optimal way to compare models is to use their likelihood [van Dyk, 2014]. Indeed, particle physics has long polished probabilistic models, minute stochastic description of observations built from first principles [Sjöstrand et al., 2001, Aaltonen et al., 2008]. And yet, recent statistical analysis of Higgs bosons is powered by black-box machine learning models –such as boosted decision trees– as they capture best background sensor noise [Aaltonen et al., 2009, Radovic et al., 2018].

1.2 Cognitive neuroscience: uncovering the functional units of human vision

Cognitive neuroscience strives to explain cognitive functions from neural activity. Which ingredients to include in such a model is a more open-ended question in than in physics. Breaking down high-level functions into units of investigation is particularly challenging. This endeavor has made much progress for the specific problem of vision. Studying early visual cortex response to specially-crafted stimuli, Hubel and Wiesel [1959] revealed neurons that form localized edge detectors. Slightly more complex shapes isolated other brain units [Logothetis et al., 1995]. These findings are tied to the stimuli presented, themselves motivated by cognitive theories used to decompose mental processes. Theories of visual processing break down it into successive operations tuned to specific aspects of the stimuli [Marr, 1982]. As any cognitive theory, their empirical neuroscience validation is then bound to this decomposition. Even with modern neural measurements, a decomposition into invalid ingredients, such as “alimentiveness” or “philoprogenitiveness” of 19$^{\text{th}}$ century phrenology, would lead to a brain mapping valid from the statistical standpoint [Poldrack, 2010].

Complete models of cortical visual processing assemble brain functional units, each implementing specific operations [Riesenhuber and Poggio, 1999]. They derive from many studies of neural responses to elementary manipulations of visual stimuli. But their neuroscience validity faced a chicken-and-egg problem as long as each functional unit had been studied in isolation: each study had investigated only one aspect of otherwise very complex stimuli, natural images. Models of vision can be derived without invoking neuroscience arguments, as in computer vision where computational models are optimized directly on natural images, *eg* for object recognition [Pinto et al., 2009, Sermanet et al., 2014]. In fact, *encoding* studies showed that pure computational models explain better neural activity than models based on hand-crafted reductions of natural images [Yamins et al., 2014]. These computer-vision models, based on artificial neural networks, extract intermediate representations of natural images, which can be mapped to brain responses, confirming functional units obtained in more hypothesis-laden neuroscience experiments [Eickenberg et al., 2017].

The large computational models do not answer some cognitive-neuroscience debates, such as the specific semantic tuning of functional areas. For instance, a brain area crucial to recognizing human faces is known as the *fusiform face area* [Kanwisher et al., 1997]. Yet, some researchers claim that its role is best explained by implementing visual expertise, rather than face recognition [Tarr and Gauthier, 2000]. As the corresponding brain area responds to both types of stimuli, the debate became trapped in a ontological disagreement: which of visual expertise or face recognition is a more central mental function? One side argues visual expertise leads to face recognition, and the other that face recognition is innate to the social human.

Encoding studies use as ingredients to map brain responses the internals of large computational models of vision. As such, they circumvent questions related to finding valid ontologies of cognitive processes: on the one hand, they cannot bring evidence in favor of ontological choices, but on the other hand they enable empirical evidence without buying into one framework. There are two ingredients to this robustness. First, encoding studies can work on more ecological and richer stimuli. Hence they capture all facets of cognition, but must rely on computational models of the stimuli, typically borrowing from artificial intelligence [Varoquaux and Poldrack, 2019]. Second, they model brain responses using high-dimensional statistical models focused on prediction. These can fit more ingredients jointly, avoiding difficult modeling choices. As a result, they can generalize findings across stimuli probing different parts of a cognitive ontology: natural images, simplified faces, or wedges traditionally used for retinotopic mappings [Eickenberg et al., 2017]. This is in sharp contrast with conventional brain mapping methodology: based on oppositions between stimuli, it does not lead to formal models bridging results from different experimental paradigms.

2 How do statistical tools fit in scientific progress

2.1 From scientific evidence to scientific knowledge: more than data

Internal versus external validity The validity of a study's findings is more than a statistical question. Internal validity controls inferences about the relations across the quantities in the study, for instance that measurements have no unmodeled errors. External validity, more important but less discussed, asserts that those relations are maintained beyond the study's settings [Cook and Campbell, 1979]. It may for instance fail when running a study on a sample non representative of the population.

Validity of constructs Scientific theories and models are constructed from abstract ingredients such as "intelligence" or phrenology's "alimentiveness" in psychology. These *constructs* are central to reasoning about empirical evidence, to position it in a broader context. A good construct is one that is useful to explain many different observations, beyond a single study [Cronbach and Meehl, 1955]. Interpreting an empirical study in a theoretical framework requires *construct validity* of its measures and manipulations: that these indeed to relate well to the construct of interest. For instance, to be interpretable as intelligence, IQ tests should not be counfounded by cultural knowledge.

Stances on theories Models, and thus theory, are needed to interpret empirical finding. The acceptance of these theories often builds upon implicit stances on their ingredients. In psychology, Fried [2020] argues that statistical models should build on "strong theories" and provide "explanation of a phenomenon" relating valid psychological constructs, beyond mere data fit. Yarkoni [2020] points out that such a view carries implicit preferences on choices of construct that may be difficult to defend. In particular, such model esthetic assumes realism about psychological constructs: that these have an absolute existence beyond the minds of the scientists. A scientific discourse must position its claims on unobservable constructs, for instance centers of gravity in mechanics. *Realism* accepts to build scientific knowledge on unobservable entities only if they are objective and mind-independent. *Instrumentalism*, rather, accepts that some ingredients of theories are mere instruments needed to tie together observable outcomes, and that the success of a theory is asserted solely on these observables.

Questions on the validity of basic modeling ingredients are less discussed in a well-established science such as physics, as there is a consensus on the ingredients: forces, acceleration, temperature –which has a non-trivial definition–... And yet, this consensus was achieved through iterations. Planetary observations in the times of Kepler were analyzed with phenomenological models lacking the ingredients of dynamics, but were fundamental to nourishing Newtonian mechanics.

2.2 Reasoning with statistical tools

Statistics gives the scientist tools to reason from noisy observations. The prevailing approach is **model reasoning**: a probabilistic model describing data generation is built, encompassing ingredients of the application domain. Parameters estimated using this model are interpreted within its logic [Cox, 2006, chap 9]. Cox [2001] goes as far as saying that statistical models are "efforts to establish data descriptions that are potentially causal". Another form of reasoning –design-based [Cox, 2006, chap 9] or **warranted reasoning** [Cook, 1991, Baiocchi and Rodu, 2021]– relies on specific experimental design, as randomization, for causal inferences without a model of the data-generating mechanism. Finally, Breiman [2001] famously noted that increasingly many statistical tools forgo data modeling, to focus on algorithmic capacity to approximate relations. Their success is established by **outcome reasoning**: gauging predictions on observables [Baiocchi and Rodu, 2021], key to machine learning.

3 Grounding more statistical reasoning on output rather than models

With a historical emphasis on data modeling, statistics has an implicit realism stance. Yet, as we have seen in physics or vision neuroscience, scientific progress is achieved despite analyzing observations without the right conceptual framework. Outcome reasoning, with tools of machine learning, gives a robust statistical framework for science: given imperfect premises, it fails less.

3.1 Robustness to model mis-specification

With model reasoning, parameters can be interpreted only conditional to the choice of model, which is outside of statistical control. Statisticians often assume that many hard modeling questions can be resolved by domain experts. Yet science is performed by limited beings [Wimsatt, 2007] and even experts have finite resources to dedicate to a given problem [Simon, 1955]. Model imperfections can have vast consequences on statistical conclusions. Botvinik-Nezer et al. [2020] asked 70 different teams of experts to analyze the same brain imaging data. Variations in modeling choices –all based on linear models– lead to vastly different parameters, and qualitatively different neuroscience findings.

Controlling predictions instead of model parameters leads to a different statistical regime. Even the simple case of the linear model changes drastically: with learning theory, analysis is possible even in the miss-specified setting, showing that multi-colinearity in the design is not an issue [Hsu et al., 2014], unlike when performing inference on model parameters. Higher-dimensional settings are possible, which means that the analyst no longer has to cherry-pick a small number of descriptors. In neuroscience, it has enabled studying richer descriptions of the stimuli, generated by artificial intelligence techniques rather than set in a specific reductionist theoretical framework. Switching to output reasoning requires reinventing analytical paradigms: in brain imaging switching to *decoding* models that gauge the ability to *predict* neural responses.

3.2 Putting explicit generalization at the center of the inference

Judging a model by its predictions is good science. It shifts the burden on validity on observables. These may suffer biases, such as censoring, which must be accounted for even in machine-learning settings [Ishwaran et al., 2008]. But in the long run, the validity of scientific theories is established by their ability to generalize across many settings.

Cross-validation on a study sample is however not a test of a strong ability to generalize; it gives no evidence of external validity. Machine-learning models may easily create local approximations which do not generalize to new settings, bad scientific models. Yet, their ability to generalize can be explicitly tested. This is unlike model-based tests of qualitative theories, as in psychology or sociology. Indeed, a methodology based on machine learning can be applied to rich descriptions of the objects under study –the raw images presented–, while model-based reasoning is applied to a small number of features, specially crafted to represented the constructs of interest –a face-place opposition. In the former, the generalization is readily tested on data from different settings via a quantitative prediction error. In the latter, the finding is more conceptual and given a new setting it must be instantiated with a new modeling effort.

Beyond broad generalization, an oft-requested feature of an analytical model is to provide "understanding". For domain reasoning, it is helpful to try to tease out the contribution of various ingredients. An emerging non-parametric statistical toolbox is catering to this purpose: black-box explanation techniques [Molnar, 2020], such as partial dependency plot [Friedman, 2001] or the knock-off [Barber and Candès, 2019]. These tools ground their inferences on model outputs, the quantities amenable to strong empirical validation. Demanding more from an analytical model, for instance opposing phenomenological data explanations with valid theoretical understanding, forces buying into a given theoretical framework Yarkoni [2020], with the risk of circular reasoning on the evidence.

Parametric models are appealing for intuitive counterfactual reasoning [Angrist and Pischke, 2008]: they appear as "data descriptions that are potentially causal" [Cox, 2001]. Yet, more than a parametric model, valid causal inference needs a structural characterization of variables, distinguishing confounders, colliders, mediators... [Greenland et al., 1999]. In such settings, machine learning models shine by their potential robustness to mismodeling [Rose and Rizopoulos, 2020].

Black-box models for thinking outside the box Empirical validation of a theory tied to its ingredients smells of self-fulfilling prophecies. This is the risk of model-based statistical reasoning. Science needs statistical reasoning based more on model predictions. Machine learning will provide the building blocks, for broad generalization and counterfactual reasoning.

References

T Aaltonen, J Adelman, T Akimoto, Michael G Albrow, B Álvarez González, S Amerio, D Amidei, A Anastassov, Alberto Annovi, J Antos, et al. Measurement of the single-top-quark production cross section at cdf. *Physical review letters*, 101(25):252001, 2008.

Terhi Aaltonen, J Adelman, Tb Akimoto, B Álvarez González, Sara Amerio, Da Amidei, Aa Anastassov, A Annovi, J Antos, G Apollinari, et al. Observation of electroweak single top-quark production. *Physical review letters*, 103(9):092002, 2009.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics*. Princeton university press, 2008.

Michael Baiocchi and Jordan Rodu. Reasoning using data: Two old ways and one new. *Observational Studies*, 7 (1):3–12, 2021.

Rina Foygel Barber and Emmanuel J Candès. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, 2020.

Nora Mills Boyd and James Bogen. Theory and Observation in Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

TD Cook and DT Campbell. *Quasi-experimentation: Design and analysis issues for field settings 1979 Boston*. MA Houghton Mifflin, 1979.

Thomas D Cook. Clarifying the warrant for generalized causal inferences in quasi-experimentation. In *Evaluation and education: At quarter century*. 1991.

David R Cox. [statistical modeling: The two cultures]: Comment. *Statistical science*, 16(3):216–218, 2001.

David Roxbee Cox. *Principles of statistical inference*. Cambridge university press, 2006.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52:281, 1955.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

Eiko I Fried. Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, 31(4):271–288, 2020.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Sander Greenland, Judea Pearl, and James M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

Daniel Hsu, Sham Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14, 2014.

D H Hubel and T N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.*, 148: 574–591, 1959.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

N Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, 17(11):4302–4311, 1997.

Nikos K Logothetis, Jon Pauls, and Tomaso Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current biology*, 5(5):552–563, 1995.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. The MIT press, Cambridge, 1982.

Christoph Molnar. *Interpretable machine learning*. Lulu.com, 2020.

Nicolas Pinto, David Doukhan, James J DiCarlo, and David D Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, 5(11): e1000579, 2009.

Russell A Poldrack. Mapping mental function to brain structure: how can cognitive neuroimaging succeed? *Perspectives on psychological science*, 5:753, 2010.

Alexander Radovic, Mike Williams, David Rousseau, Michael Kagan, Daniele Bonacorsi, Alexander Himmel, Adam Aurisano, Kazuhiro Terao, and Taritree Wongjirad. Machine learning at the energy and intensity frontiers of particle physics. *Nature*, 560(7716):41–48, 2018.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

Sherri Rose and Dimitris Rizopoulos. Machine learning for causal inference in biostatistics. *Biostatistics*, 21(2): 336–338, 2020.

Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.

Torbjörn Sjöstrand, Patrik Eden, Christer Friberg, Leif Lönnblad, Gabriela Miu, Stephen Mrenna, and Emanuel Norrbin. High-energy-physics event generation with pythia 6.1. *Computer physics communications*, 135(2): 238–259, 2001.

Michael J Tarr and Isabel Gauthier. Ffa: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature neuroscience*, 3:764, 2000.

David A van Dyk. The role of statistics in the discovery of a higgs boson. *Annual Review of Statistics and Its Application*, 1:41–59, 2014.

Gaël Varoquaux and Russell A Poldrack. Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current opinion in neurobiology*, 55:1–6, 2019.

William C Wimsatt. False models as means to truer theories. *Neutral models in biology*, pages 23–55, 1987.

William C Wimsatt. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press, 2007.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Tal Yarkoni. Implicit realism impedes progress in psychology: Comment on fried (2020). *Psychological Inquiry*, 31(4):326–333, 2020.