THE REASONER

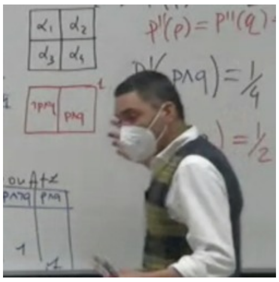
VOLUME 17, NUMBER 2
MARCH 2023

thereasoner.org
ISSN 1757-0522

CONTENTS

EDITORIAL

Dear Reasoners, it a pleasure to introduce this issue, featuring my interview with Erica Thompson. Erica is Senior Policy Fellow at the LSE Data Science Institute in addition to being fellow of the London Mathematical Laboratory and Honarary Senior Research Fellow at UCL Department for Science, Technology, Engineering and Public Policy. She recently authored *Escape from Model Land: How Mathematical Models Can Lead Us Astray and What We Can Do About It*, a book which brings to the wider audience a set of thorny methodological issues related to mathematical models. It has received great reviews in, among others, The Guardian, The Wall Street Journal and The Economist. In the interview Erica covers the path which led her to writing the book, and more generally to develop her highly interesting views on how uncertainty sometimes can be way trickier than we expect. I'm grateful to Erica for her time and for sharing her thoughts with the readers of The Reasoner.

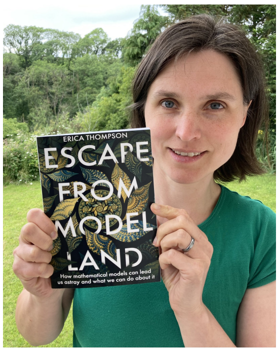


HYKEL HOSNI
University of Milan

FEATURES

Interview with Erica Thompson

HYKEL HOSNI: Can you tell us about your background?



ERICA THOMPSON: I studied Natural Sciences for my undergraduate degree, because I couldn't decide which subject I liked best. After specialising in theoretical physics, I realised that I had a choice between studying extremely big things (astrophysics) or extremely small things (quantum physics). But for me the interest of science has always been in the human scale, things that you can observe directly, and so I was more interested in statistical physics, electromagnetism and particularly fluid dynamics. To study those, it turned out you needed to be in the Maths department, so I switched subjects and pursued an MMath.

HH: Was it, in retrospect, a good choice?

ET: Well, I have no access to my counterfactual life and wouldn't trust a model of it, so I can't make a relative comparison but I am happy with how things have turned out! I took mostly fluid dynamics options and didn't distinguish myself in the exams, but I particularly enjoyed writing an essay about the mathematics of carbon sequestration in underground reservoirs, and doing a summer project on granular column collapses, both supervised by Prof Herbert Huppert of the Institute for Theoretical Geophysics in DAMTP at Cambridge.

HH: What was your dissertation on?

ET: There wasn't a dissertation as such, just an optional "essay" component, essentially a literature review, which was about fluid flow in porous media applied to carbon sequestration; I wrote about the maths but also about the practicalities of fluid injection, the problems of instability, the potential for leakage, and data from pilot projects.

HH: And then you realised it would be great to do a PhD

ET: Having done the essay and also a summer project, I realised that you could be a mathematician and tackle real-world problems. But I didn't do well enough in the exams to be considered for a PhD in DAMTP, so I applied for a studentship at Imperial College's department of Earth Science and Engineering, on the same topic of carbon sequestration, finding numerical solutions to equations for fluid flows through porous media. Although most work on this was (is) funded by oil companies, my post was part of Imperial College's new Grantham Institute for Climate Change.

HH: Please tell us a bit more about it

ET: This was 2007-8 and a period of heightened interest in climate change in the UK prior to the Copenhagen climate conference. The Grantham Institute at Imperial and its sister Institute at LSE had just been set up by the influential investor Jeremy Grantham with a large philanthropic donation. So I started this PhD and began work, but increasingly I felt unsure whether carbon sequestration was the answer - since it was primarily justified on the basis of enhancing oil recovery from depleting reservoirs - and sure that if it was the answer, the social and political issues were much more pressing questions than the mathematics and technology.

HH: The human scale.

ET: Yes, the human scale and the human context for our science. So I took an interruption of studies to spend a few months working as a researcher for a UKERC report on global oil depletion. A deep dive on this topic further convinced me of the need for rapid action on climate change and its systemic interlinkage with resource challenges in the 21st century. I was keen to put my efforts more towards the problem of climate rather than going back to the original PhD topic, and the Grantham Institute very kindly facilitated that move by offering me a new PhD position with Prof Brian Hoskins, a dynamical meteorologist and climate scientist.

HH: Awareness on climate change wasn't so widespread then as it is now, so it was brave of you to walk away from a scientifically consolidated area halfway through your PhD. Weren't you scared about your career prospects?

ET: Actually I think it was brave of them to take a risk on a flaky student! I think I would have left academia otherwise. Fortunately my new topic, North Atlantic storms, immediately felt more comfortable than Matlab models of oil reservoirs, and I enjoyed generating, analysing and comparing the outputs of different models, both general circulation models (GCMs) and very basic statistical models of storm occurrences. The Grantham Institute was a wonderful place for a PhD student with broad interests, since the seminars were on such a diverse range of scientific, economic and policy topics.

HH: Can you tell us a bit more about your PhD research?

ET: I was looking at what different models said about the prospect for changes to North Atlantic storm tracks in the 21st century due to greenhouse gas emissions. Once I started working with the models and constructing a literature review of previous results, I soon felt that the question of model uncertainty wasn't adequately addressed. Effectively, the error bars of the available models didn't overlap, and my own models and analysis were not going to be decisive. So I found it difficult to make any inferences about North Atlantic storms, but it really sparked my interest in models and understanding their relationship with the real world.

HH: Was that your first encounter with model uncertainty, or had you been exposed to it in your degree?

ET: In my physics degree we did a lot of "measurement error" and combining random uncertainties, but little or nothing about model inadequacy. But I guess the general approach of physics is that if you have a systematic error you can remove it and if you have a random error you can estimate it. In the theoretical geophysics group at DAMTP of course there were lots of idealised models and scaling laws, but discrepancies with observation were due to the messiness of experimental set up. Certainly I experienced more acknowledgment of the difficulties of imperfect models of nonlinearity (eg fluid instabilities, or equations without analytic solutions) in mathematics than in the physics department. But that's also partly just the topics that I studied. We always start with the nice linear things because they are easier - the trouble is when you develop intuition for the linear situations and apply it widely without realising they are a very special case.

HH: Back to your PhD. How did it go with the model analyses?

ET: I finished them competently though without great enthusiasm, but I dived into various rabbit-holes in statistical inference and philosophy of science to try to work out how to do a more defensible uncertainty analysis. I really wanted to know what I ought to be doing but everything I found just opened more awkward questions.

HH: That's not unheard of! And then you met Leonard Smith.

ET: Yes, at some point he came to give a seminar at the Grantham Institute and his account of the dynamical characteristics of model error made more sense to me than anything else I had read.

HH: Did you tell him?

ET: Well I probably asked some very dumb questions in Lenny's seminar, and then I think I emailed him about forty pages of my working literature review document which was my exploration of all these rabbit-holes, and asked for comments. I don't think he ever got back to me with any comments on the document, but I guess he must have read and liked some of it because he suggested I apply for a postdoc position in his group at LSE, the Centre for the Analysis of Time Series (CATS).

HH: So after finishing the PhD you moved to CATS.

ET: I worked there mostly on short term research projects in a variety of areas, from climate to energy, insurance, weather forecasting and anticipation of humanitarian crises. It was super interesting to see the real-world implications of model imperfection in these different decision-making contexts, and the ways that different kinds of stakeholders find to deal with the limitations of scientific information.

HH: Is there an example you'd like to tell us about?

ET: Sure. The humanitarian one is interesting. In principle some disasters are predictable, particularly weather-related disasters such as hurricanes. And in principle if you could release money to take action before the event you could actually reduce the loss and damage suffered and save lives rather than just helping to clean up afterwards. But the forecasts aren't perfect, and if you want to predict further in advance then you trade off very strongly with accuracy. Acting on the basis of an imperfect forecast might have other downsides, like the dangers associated with evacuating a town, plus reputational downsides if you are perceived to overreact or waste resources. So I worked with the agencies to think about how to use forecasts of heatwave in Pakistan, and cyclones in Madagascar, to support real operational decision-making. That's partly a scientific question - what information is available? - but also partly a question of ensuring that the relevant information can interface well with whatever the decision-making procedures are. So to some extent I was working on tailoring the information to the decision, and also partly helping to advise on tailoring the decision and operational procedures to the information. There's no point even trying to make a decision ten days in advance if you don't have a reliable forecast then. But there's also no point in having a decision procedure which takes 48 hours to activate if the event will have happened by then. There are all sorts of really interesting questions about science in practice raised by this kind of work at the interface - I've enjoyed it and learned a lot.

HH: You have recently published a very successful book. Did you decide to write the sort of book you had hoped to find down the rabbit-holes?

ET: Maybe. I mean yes, I would have loved to find this book, but it's also very much informed by the work I have done in the ten years after my PhD. What I was always missing in my academic work was the time to integrate and synthesise these insights, so it was incredible to have the opportunity of a part-time fellowship at the London Mathematical Laboratory to work on a longer but accessible introduction to my thoughts. And then it was published by Basic Books (UK and US) in 2022 as a "popular science" book called Escape From Model Land: how mathematical models lead us astray and what we can do about it.

HH: I have seen it reviewed by a number of prestigious outlets, including The Economist. It looks as if the systematisation of your insights is rapidly becoming a big hit!

ET: I don't think it's destined to be a bestseller but I've been really pleased with the response to the book and it has already led to lots of interesting conversations which would never have happened without it.

HH: A great opportunity indeed. What is the key message you wanted people to take home?

ET: I suppose one key message of the book is that the map is not the territory, which of course is not a new insight at all. But I hope that I have done a reasonable job of explaining why it matters that the map is not the territory, and what sort of maps we might need to draw on to inform better decision-making. I cover the limitations of inference from models and explain how models embed all sorts of social and political value judgements. Climate change is one of my examples of course, and while writing my book about models I have watched an incredible case study of model-supported decision-making play out in real time with the spread of Covid-19, so that's another key example. And I also look at economic and financial models which are so important to the functioning of the global economy. I hope that these concrete examples help to make it more accessible, or at least help to illustrate why the points are so important for us all and not just to academics.

HH: Your writing is both engaging and accessible. I suspect juggling several research projects at CATS has forced you to develop cross-area communication skills. Still I think writing for such a broad audience is far from easy. Do you have any suggestions or good references for those who are embarking on similar projects?

ET: I think the best advice is probably just to give it a go. Do plenty of reading and note which styles are appealing and interesting. Do plenty of talks to varied audiences and see which ways of explaining your ideas work well. Get on social media and do some listening, even if you don't feel like talking there. And I guess also consider what the right way to reach your target audience is. For me a book felt like the right answer but for others it might be social media or YouTube videos or policy briefings or something else entirely. Also don't be too much of a perfectionist - I've done plenty of other writing that never reached the light of day but in this case I had an editor chasing me for a book manuscript so I was forced to hand it over. Deadlines are definitely a good thing.

HH: You have recently moved to the LSE Data Science Institute. What are you working on there?

ET: I have a longer-term funded research project of my own, a UKRI Future Leaders Fellowship which brings together the mathematical, philosophical and practical angles of using imperfect models to inform decisions.

HH: Looks like the kind of thing this community really enjoys! What kind of questions are you tackling?

ET: At the moment I am exploring the position that in order to understand what information models can contain, we need to move away from consensus-seeking (all models say X) and towards exploration of the limits (no plausible model could say not-X). If we reframe inference into the regime of "no plausible model could say not-X" then of course the word "plausible" is doing all of the work in the sentence. But that's critical - "plausibility" is primarily subjective, it depends on our assessment of the expertise of the individual/group creating the model.

HH: Not exactly the caricature of "evidence based policy" in which scientists tells us what is true.

ET: Indeed. There is a huge question there about confidence, trust, group dynamics, what we mean by expertise and how we do science. I'm also trying to translate these questions back into the mathematics of statistical inference using model outputs, and thinking about how real decision-makers actually use model information (spoiler: generally not the way scientists would like them to!) and what the implications might be for more automated decision-making systems such as those employing methods described as "artificial intelligence".

HH: Many of our readers work on AI-related topics. Can you say something about those implications?

ET: In my book, I distinguish between two "escape routes from Model Land". One is the quantitative route, whereby we take our model and compare it with real data. If we have enough data and the underlying conditions are not changing, then that data-driven escape is fantastic. The modern world is built on models in which we have high confidence because we can test them extensively, like the laws of physics and electromagnetism, or the prediction of ballistic motion. But if we are trying to make predictions of social systems like financial markets, then "past performance is no guarantee of future success". If we are trying to predict high impact low likelihood events like severe weather risks, then we simply don't have very much relevant data to go on. If we are trying to model a system which is intrinsically changing, like the climate system, then we don't necessarily know whether the parameterisations developed on past climate will be appropriate in future. So in all those cases we have to make a qualitative exit from Model Land: a strong expert judgement about model quality (for the pedants, yes, we made a strong expert judgement in the previous case as well, it's just that most people agree it to be reasonable). So for the data-driven approaches where we can agree that past performance is a good indicator of future performance, AI and machine learning methods can be incredibly powerful tools. But I think there are strong constraints to the (constructive) use of AI in the more extrapolatory domains, because of the reliance on expert judgements about model quality. Which is not to say that you can't do it, only that you are going to have to be super careful about the assumptions and accept that people may disagree with them. Any kind of autonomous decision-making implies value judgements, and also differential impacts which need to be examined, as Cathy O'Neil showed in her book on algorithms (Weapons of Math Destruction).

HH: In conclusion I'd like to ask you a question on post-pandemic academic life, if you don't mind.

ET: Sure, go ahead!

HH: As the restrictions were being eased pretty much everywhere, many in the profession realised that we were flying too much, and sometimes for really negligible benefit – think of the day-return flights which were so common for departmental seminars etc.. Several (strong) opinions on reducing dramatically in-person conferences and workshops have been put forward over the last couple of years. What's your view on this?

ET: Well, I haven't been on a plane for fifteen years now - since starting my PhD in climate science - so I certainly support calls to reduce the frankly ridiculous and wasteful attitude towards frequent flying that has been commonplace in academic circles.

HH: Wow! Many of us have become really aware of this very recently (if at all). How did you play the academic rituals when so few people understood the importance of the question?

ET: It used to be awkward and weird to decline invitations on climate grounds but it is becoming less unusual, and of course we now have much more effective online options that didn't exist fifteen years ago. I thought it would be hard to continue a career in academia without flying but I have put a lot of effort into developing local and virtual connections, and somehow I do seem to still be here! I hope it's also become clearer that there are other good reasons than carbon emissions to make online and hybrid collaboration options available. For those with little/no travel budget, caring responsibilities, disabilities, neurodivergence and even those who don't drink alcohol, traditional conferencing options can be exclusionary and inaccessible.

HH: Indeed.

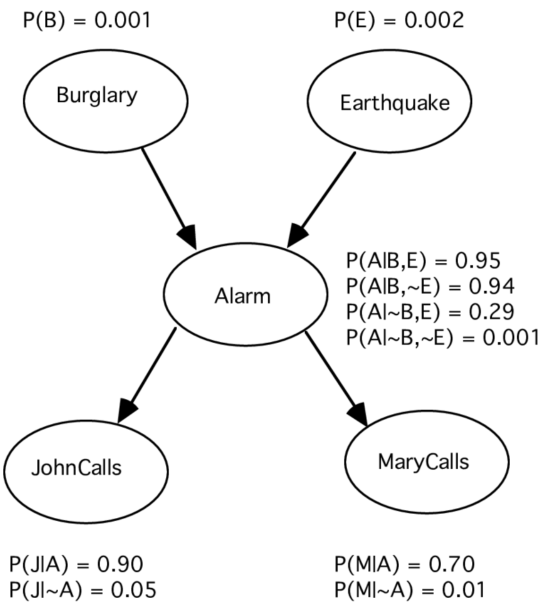
ET: So while there is a place for in-person networking, we should be capitalising on the benefits of our involuntary trial of non-flying academia. How about putting travel budgets towards accessibility for junior researchers from less advantaged countries and institutions? Of course there will be plenty of edge cases and exceptions, and everyone has to make their own decisions, but I think a majority of people agree that the right amount of flying is a lot less than was the pre-Covid norm. So, as I wrote in 2011, "*If we are going to continue to do good science in a low-carbon future, then we need to find more efficient ways of working. As well as providing moral leadership, those researchers, universities and institutions who explore new, low-carbon paradigms now will become the architects of new collaboration methods, and the leaders of future science. Don't get left out!*"

Causal Attribution Tool

We introduce a new tool, Causal Attribution Tool (CAT), using Causal Bayesian Networks (CBNs), which render causal inferences graphically and explicitly. CAT is a new kind of tool which allows you to interrogate a causal Bayesian Network to see: updated probabilities given observation evidence (as does every other BN tool, of course); updated probabilities given causal interventions plus optional observations (which only a very few tools have allowed before); determine the causal attributions for the effect given those interventions and observations, according to various criteria, which is unique. The attribution criteria implemented so far are: a variation of Glymour and Cheng (1998: Causal mechanism and probability: A normative approach, in Oaksford and Chater (eds.) *Rational models of cognition*) causal power; the Fraction of Attributable Risk (FAR), widely used in epidemiology and climate science; our own Causal Information theoretic criterion. CAT is open source, so this list can be extended, either by us or by you, to include, for example Halpern and Pearl (2005: Causes and Explanations: A Structural-Model Approach. Part I: Causes, *British Journal for the Philosophy of Science,* 56(4)) criteria, or Hitchcock's (2001: The intransitivity of causation revealed in equations and graphs, *Journal of Philosophy, 98(6), 273-299* criterion or others. Indeed, suggestions are welcome.

Questions about attribution — What caused this cancer? Is this flooding event due to anthropogenic global warming? — are critical for society. Our tool is unlikely to resolve such questions, but it can provide a platform where such questions can be rendered explicitly and the different answers implied by different causal criteria compared one with another. We hope this will focus debate about attribution, which has largely proceeded in a haphazard fashion to date, with proponents and detractors of particular approaches going after each other at random. With CAT, they can potentially form a circle and collectively shoot at each other at the same time, as in The Good, the Bad and the Ugly:)

Causal Bayesian Networks (CBNs) A CBN is a graph whose nodes represent the important factors, causes and effects, in a network system where arrows represent the direct causal influences of the parent (cause) on the child (effect). Thus, in the CBN above, Judea Pearl's home alarm network (Pearl 1988: Probabilistic reasoning in intelligent systems: Networks of plausible inference, Morgan Kaufmann), the Alarm sounding is directly connected to everything: burglaries and earthquakes directly cause the Alarm to sound and in turn that will directly cause Pearl's neighbors to call and warn him at work when it does sound. The connections cause changes in the effect node's probability distribution. Thus, for example, when there is an earthquake alone Pearl's Alarm will sound with probability 0.29 (as indicated by the line "$P(A|\neg B, E) = 0.29$", where: ¬ means negation; | means conditioning on).



It doesn't take a sophisticated attribution theory to see that earthquakes can cause Pearl's Alarm to go off. Where theories differ lies in such questions as **how much** responsibility to put on one cause versus another, especially when causal influences are mediated by a complex subnetwork. Ideally, we would have percentage attributions for distinct causes given a context, much like the explained variation of Sewell Wright's path models (Wright 1934: The method of path coefficients, *Annals of Mathematical Statistics, 5(3), 161-215*).

Causal Criteria Currently supported causal criteria are:

1) Glymour and Cheng (1998)'s Causal Power begins with "positive causal contrast":

$$\Delta P_c = P(e|c) - P(e|\neg c)$$

Assuming this is positive, it reports to what extent intervening and setting C to value c raises the probability of an effect value of interest, $E = e$. Cheng proposes restricting this measure to cases where, without c, e would not have occurred; that is, there are cases where e would have occurred anyway, and we shouldn't count them. Hence:

$$p_c = (P(e|c) - P(e|\neg c))/(1 - P(e|\neg c))$$

which is their **causal power**. An additional constraint is using this measure only with Naive Bayes models, where the Cause is the only ancestor to the Effect; multiple causes are disallowed. We enforce a similar restriction by cutting off any parents of the Cause when measuring causal power, although the relevant subnet may not have a Naive Bayes structure.

2) Fraction of Attributable Risk (FAR) is:

$$FAR = 1 - P(e|\neg c)/P(e|c)$$

This is in widespread use in epidemiology (as a minor variation on "risk difference" , biology and climate science (Stone and Allen 2005: The end-to-end attribution problem: From emissions to impacts, *Climate Change, 71, 303-318*; Stott et al. 2016: Attribution of extreme weather and climate-related events,*Wiley Interdisciplinary Reviews: Climate Change,7(1), 23–41*). The idea behind it is that, since c and $\neg c$ are binary, e can be attributed to one or the other. Since the maximum attribution percent is 100%, we can divide this in two parts, one for c and one for $\neg c$q. If the ratio $P(e|\neg c)/P(e|c)$ is, say, 1/3, then c is tripling the probability of e compared with $\neg c$, which we can recognize by attributing an actual occurrence of e 67% to c.

3) Causal Information (CI) Criterion. Simply put, our concept of CI is identical to Mutual Information (MI), except that it is measured in a causal Bayesian network which has undergone a causal intervention; also it is measured in comparison with a maxentropy distribution in the Effect, unlike MI proper. There are potentially many varieties of intervention possible; currently CAT only supports a uniform distribution on the Cause and cutting all other parent arcs, see (Korb et al. 2004: Varieties of causal intervention, in *Pacific Rim International Conference on Artificial Intelligence*, 322-331) for alternatives. These three criteria can be applied to any CBN given to CAT, compared with each other and compared with your own judgment about causation in any given case.

Other Criteria There are many other criteria people have considered. Since CAT is open source code (`https://github.com/voracity/CAT`), readers are invited to code up whatever criteria they may like to see in action. Readers are also invited to contribute causal Bayesian networks of interest to the CAT tool itself (there is an upload and publish option for CBNs). That webpage also includes an expanded version of this article.

P.S. We are editing a special issue of *Algorithms* on Bayesian Networks and Causal Reasoning. If anyone is interested, please look at the CFP.

Kevin B. Korb
The University of Melbourne
Steven Mascaro
Bayesian Intelligence Pty Ltd
Erik P. Nyberg
Monash University
Yang Li (Kelvin)
Deakin University

Aristotle's Logicism

Bertrand Russell credits Gottlob Frege with being the first in "logicising" mathematics, '*i.e.* in reducing to logic the arithmetical notions which his predecessors had shown to be sufficient for mathematics.' (1919: *Introduction to Mathematical Philosophy*, London: Allen and Unwin, p. 7) It therefore seems an abundant anachronism to speak of Aristotle and logicism in the same breath, let alone the fact that he was not a mathematician either in practice or profession by any stretch of imagination. The purpose of this short paper is thus to argue otherwise by setting the project of logicism within a larger context than mathematics and to offer some grounds for the logicist credentials of Aristotle.

To set the scene before locating the logicist landmarks in the Aristotelian text, there are a couple of cardinal characteristics of the logicist project that need to be highlighted. First, for the logicism project to get off the ground, the initial necessary step is to set up a formal deductive system of logic adequate for formalising the reasoning of one domain into another one.

Specifically, in the case of Fregean logicism and its recent descendants in the form of neo-logicism, the formal deductive system must possess the ability to formalise mathematical reasoning. This indeed constitutes the principal *prerequisite* or *precondition* of any attempt in the implementation of logicism. Secondly, the claim of logicist reduction can be understood in one of two senses: either in the *strong* sense of claiming that all *truths* of the reduced domain comprise a subset of logical truth or in the *weak* sense of claiming that all *theorems* of the reduced domain comprise a subset of logical truth.

Now, with this perfunctory prelude to a few features of logicism, the question is whether the Aristotelian corpus does afford any textual evidence in support of his allusion and allegiance to the doctrine of logicism. In this regard, one of the most promising sources is Aristotle's epistemological and ontological ruminations and pronouncements in his later work of *Metaphysics*. There is a notable consensus among scholars that Aristotle's *Metaphysics* is intentionally concerned with the problem of scepticism as an integral part of a universal or special science of being. Indeed, his discussion of the Protagorean doctrine, arising out of the problem of conflicting appearances, is purposefully tied to the denial of the *law of non-contradiction* which in turn epitomises itself in the Aristotelian corpus as radical scepticism.

Prima facie, one may suspect a dissonance here as any discussion of the law of non-contradiction seems to be more ensconced in the domain of logic and its foundation in contrast with a study of the content and details of a universal or special discipline dedicated to the overarching subject of being and existence. However, Aristotle in his pioneering role as the first *metalogician* (Jonathan Lear: 1980, *Aristotle and Logical Theory*, Cambridge: Cambridge University Press) attempts to shed light on the nature of proof and consequence and, in particular, the status of the law of non-contradiction in his *Metaphysics* with the ultimate aim of demonstrating the *intelligibility* of the broad structure of *reality* in the same breath.

In Aristotle's own articulation, this metaphysical and metalogical interplay and interaction takes place in the following manner: 'Obviously then it is the work of one science to examine being *qua* being, and the attributes which belong to it *qua* being, and the same science will examine not only substances but also their attributes'. (Richard McKeon, ed.: 1941, *The Basic Works of Aristotle*, New York: Random House, 1005^{a} 13-16, p. 735) And, lest there is a minimalist or broad understanding of substances and their attributes in this context, Aristotle adds that: 'We must state whether it belongs to one or different sciences to inquire into the truths which are in *mathematics* called axioms, and into *substance*. Evidently, the inquiry into these also belongs to one science ... for these truths hold good for everything that is, and not for some special genus apart from others.' (*Ibid.*, 1005^{a} 18-24, pp. 735-36; emphasis added)

To reinforce his point, he continues by cautioning against two sets of false contenders here. For the first set, he targets mathematicians and, specifically, geometers and arithmeticians: 'since these truths clearly hold good for all things *qua* being (for this is what is common to them), to him who studies being *qua* being belongs the inquiry into these as well. And for this reason no one who is conducting a special inquiry tries to say anything about their truth or falsity – neither the geometer nor the arithmetician.' (*Ibid.*, 1005^{a} 27-31, p. 736) That is, not only the mathematical axioms are not the fundamental principles of what Aristotle's special science is going to ascertain but also they are not in themselves sufficiently *sui generis* to form an independent set of their own.

For the second set of contenders, Aristotle rebukes natural philosophers for harbouring such ontological ambitions. This is quite interesting in view of Aristotle himself being a naturalist *par excellence* evidenced by his iconoclastic revolt against his master's suprasensible and supernatural entities of the platonic forms. In his dismissal of natural philosophy as the home of being *qua* being, he writes: 'Some natural philosophers indeed have done so, and their procedure was intelligible enough; for they thought that they alone were inquiring about the whole of nature and about being. But since there is one kind of thinker who is above even the natural philosopher (for nature is only one particular genus of being), the discussion of these truths also will belong to him whose inquiry is universal'. (*Ibid.*, 1005^{a} 31-35, p. 736) In particular, he goes after those who offer the discipline of physics as furnishing the foundational principles of existence. Although Aristotle readily acknowledges the status of physics as 'a kind of Wisdom', he chides the advocacy of physics as the special science of being 'due to a want of training in *logic*'. (*Ibid.*, 1005^{b} 1-3, p. 736; emphasis added)

So, the question is which discipline or branch of knowledge has the necessary wherewithal and the *logical* capability to deliver the objectives and goals of the universal or special science of being. Aristotle's answer is unhesitatingly categorical with a tantalising twist: 'Evidently then it belongs to the philosopher, i.e. to him who is studying the nature of all substance, to inquire also into *the principles of syllogism*. (*Ibid.*, 1005^{b} 6-8, p. 736; emphasis added) The significance of the twist, *viz.* the reference to the theory of syllogism, can be best appreciated against the backdrop of the forgoing first observation about the project of logicism: namely, the prerequisite or precondition of the availability of a formal deductive system of logic adequate for formalising the reasoning of one domain into another one.

It should be noted that for Aristotle this appeal to the syllogistic formal system in the context of studying being *qua* being is neither accidental nor incidental. The idea of a reduction process in the discovery, classification, and ordering of the principles of *each* genus of being is a fundamental feature of his formal methodology. Indeed, the burden of his *Prior Analytics* is primarily to provide a formal apparatus through which such determinations and reductions can take place with apodeictic necessity. aAristotle reiterates the same commitment here in the context of *Metaphysics* again: 'he who knows best about each genus must be able to state the most certain principles of his subject, so that he whose subject is existing things *qua* existing must be able to state the most certain principles of all things.' (*Ibid.*, 1005^{b} 8-10, p. 736)

But, what is the outcome of the study of being as being by inquiring into 'the principles of syllogism'? It is a principle, remarks Aristotle, that 'is the most certain of all': '*It is, that the same attribute cannot at the same time belong and not belong to the same subject and in the same respect*': that is, the law of non-contradiction. (*Ibid.*, 1005^{b} 17-20, p. 736; emphasis added) Yet, to leave no room for doubt as to the core fundamentality and centrality of this principle *vis-à-vis* any other principles including *mathematical* ones, Aristotle sharpens his 'logicist' stance by the following observation: 'This, then, is the most certain of all principles ... *that all who are carrying out a demonstration reduce it to this as an ultimate belief; for this is naturally the starting-point even for all the other axioms.* (*Ibid.*, 1005^{b} 22 and 31-34, pp. 736-7; emphasis added)

Thus, in Aristotle's ontology, what ultimately underwrites being and existence is logic, or, more specifically, the law of non-contradiction, and thereby metaphysics and metalogic seem to be intrinsically coextensive in the Aristotelian architecture. On this basis, it may not be an anachronism after all to think of Aristotle as an early proponent or a precursor of *logicism*, except on a grander scale than its circumscribed mathematical variety as presented in the works of Frege, Russell and later neo-logicists when it comes to the overall ontological structure of reality.

Majid Amini
Virginia State University

Dissemination Corner

BRIO

Opacity, bias and risk are pressing issues in the Artificial Intelligence (AI) community. This is mainly due to the development of modern AI systems based on Machine Learning (ML) algorithms, which have at least two critical aspects: 1) a training phase that uses a large amount of data; 2) a large number of parameters that are fixed only at the end of the training phase and that define the final behaviour of the system. These aspects make AI systems essentially black boxes, so that although they may exhibit very high performance, they may not be very reliable, safe and transparent. Our group, in collaboration with the other researchers in the BRIO (Bias, Risk and Opacity in AI) project - funded by the National Research Project (PRIN MIUR) - aims to contribute to the development of trustworthy AI systems by improving the transparency of such AI systems developed with ML techniques in general and deep learning in particular. To this end, we intend to develop algorithms in the field of eXplanable Artificial Intelligence (XAI) that are capable of providing explanations for the behavior of AI systems that are as comprehensible as possible for human beings. Notice that different senses of interpretability for AI systems have been distinguished and analyzed in the literature, e.g. to explain the behavior of ML classification systems for which only I/O relations are accessible. This type of approach is known as model agnostic. Instead, if the internal mechanisms of the model are available for building the explanations, the XAI method is said to be model specific. Several model agnostic approaches have been developed to provide global explanations, for example, consisting of a class prototype to which the input data can be associated. These explanations answer queries that are usually expressed as why-questions: ”Why were input data x assigned to class C?” On the other hand, specific why-questions of the type: ”Why was this credit application rejected?” and ”Why was this image classified as a fox? In this case, the XAI system responds by providing local explanations that highlight salient parts of the specific input. From another point of view, an important distinction in the XAI field is between model-based and post-hoc explainability, the former consisting in AI systems that are explainable by design (e.g. decision trees), since their internal mechanisms are easy to interpret, and the latter proposing explanations built for systems that are not easy to understand. In the XAI literature, much of the research tends to provide agnostic and post-hoc solutions, since they can be applied to a wide range of ML systems. In this context, successful XAI strategies consist of providing explanations in the form of visualizations and, more specifically, low-level input features such as relevance scores or heat maps of the input, such as sensitivity analysis or layer-wise relevance propagation methods. However, it is important to note that the main problem with such methods is that they impose a significant interpretive burden on the human user. Indeed, starting from the relevance of each low-level feature, the human user has to identify the overall input properties that are perceptually and cognitively salient to him. Thus, an XAI approach should alleviate this weakness of low-level approaches and overcome their limitations by providing the possibility to construct explanations in terms of input features that represent more salient and understandable input properties for a human, which we call here Middle-Level Input Features (MLFs). These types of approaches align very well with the aims of the BRIO project insofar as such approaches lead to significant advances in the implementation and verification of less opaque and more trustworthy systems. In the XAI literature, however, there are relatively few approaches that pursue this line of research focusing on explanations in terms of MLFs. In Ribeiro et al. (2016: ”why should i trust you?” explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135–1144), the authors proposed LIME, a successful XAI method which is based, in case of image classification problems, on explanations expressed as sets of regions, clusters of the image, said superpixels which are obtained by a clustering algorithm. These superpixels can be interpreted as MLFs. In one of our recent papers Apicella et al. (2019: Explaining classification systems using sparse dictionaries, ESANN 2019 - 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 495 – 500) we propose a solution by which the explanations are formed of elements selected from a dictionary of MLFs, obtained by sparse dictionary learning methods. However, these approaches propose specific solutions which cannot be generalized to different types of input properties. We want to investigate the possibility of obtaining explanations using an approach that can be applied to different types of MLFs, which we call General MLF Explanations (GMLF). We want to propose a general framework insofar as it can be applied to several different computational definitions of MLFs and a large class of ML models (model-agnostic and post-hoc explanations). Consequently, we can provide multiple and different explanations based on different MLFs. In particular, we are building upon the idea that the elements composing an explanation can be determined by encoders/decoders able to extract relevant input features for a human being, i.e., MLFs, and that one might change the type of MLFs changing the type of encoder/decoder or obtain multiple and different explanations based on different MLFs. For example, in Apicella et al. (2022: Exploiting auto-encoders and segmentation methods for middle-level explanations of image classification systems, Knowledge-Based Systems, 255:109725), a general framework to obtain humanly understandable explanations is proposed, considering three different ways (segmentation, hierarchical segmentation and variational auto-encoder) to obtain MLFs, which are based on encoder/decoder systems. The proposed approach enables one to obtain different types of explanations for a ML system in terms of different MLFs for the same pair input/output, paving the way to XAI solutions able to provide human-understandable explanations.

Roberto Prevete, Andrea Apicella
Department of Electrical Engineering and Information Technology, University of Naples Federico II

News

Call for Papers

Logic for the new AI spring: special issue of *International Journal of Approximate Reasoning*, extended deadline 15 March..
Formal and Cognitive Reasoning: special issue of *International Journal of Approximate Reasoning*, deadline 30 April.

What's Hot in . . .

Statistical Relational AI

A key distinction in statistical relational AI is between directed approaches, where dependencies between propositions are conceived as an influence that the validity of one proposition has on the likelihood of another, and undirected approaches, which see dependencies merely as statistical correlations. The origin of this dichotomy lies in probabilistic graphical models, where Bayesian networks on directed acyclic graphs exemplify directed approaches and Markov networks on undirected graphs exemplify undirected approaches.

To the logically inclined, one of the beauties of statistical relational AI is that they provide a bridge between logical specifications on the one hand and such probabilistic (graphical) models on the other. So how does this fundamental feature of directedness reflect in the logical basis of the frameworks?

Probabilistic logic programming, arguably the most mature directed paradigm in statistical relational AI, is founded on the premise of a complete separation between the probabilistic and the logical part of the program, as can be seen in our classic example:

Example The probabilistic logic program *Smokers and Friends* consists of the probabilistic facts

```
0.2 :: befriends(X,Y).
0.5 :: influences(X,Y).
0.3 :: stress(X).
```

and the rules

```
friends(X,Y) :- befriends(X,Y).
friends(X,Y) :- befriends(Y,X).
smokes(X)  :- stress(X).
smokes(X)  :- friends(X,Y), smokes(Y), influences(Y,X).
```

All the sophisticated logic is in the rules rather than in the probabilistic part, which merely encodes an independent set of Boolean random variables with specified success probabilities. From a logician's point of view, the rules in the deterministic part *define* the meaning of the intensional predicates `friends` and `smokes` in terms of the random predicates `befriends`, `influences` and `stressed`. These definitions are cast in the clausal Datalog language of logic programming, but they could equally be written as logical formulas. In fact, finite model theorists have long known that such Datalog programs can be expressed in *least fixed point logic*, a much-studied extension of first-order logic. So one could say that probabilistic logic programming is really just least fixed point logic over free random structures. This brings their analysis very close to classical finite model theory. For instance, Ron Fagin's celebrated zero-one law says that every first-order formula has asymptotic probability zero or one when evaluated on precisely such random structures, and the extension of this law to least fixed point logic is well-known among finite model-theorists.

Compare this to Markov logic, the logical formulation of Markoc Logic Networks. In Markov logic, a knowledge base is a finite set of (quantifier-free or) first-order formulas, either absolutely true or annotated with real-valued weights. Then the likelihood of any given structure satisfying the absolute constraints depends on the (weighted) number of formula instances in the knowledge base that it satisfies.

Example A Markov logic network for *Smokers and Friends* could be written thus:

```
-1 :: friends(X,Y).
      friends(X,Y) <=> friends(Y,X).
-1 :: smokes(X).
 1 :: smokes(X) <= friends(X,Y), smokes(Y), influences(Y,X).
```

Where probabilistic logic programming uses first-order logic as a language of definitions, Markov logic uses it as a language for *constraints*. The absolute formulas are considered to be universally quantified, as they must be true under every variable assignments. Weights represent penalties incurred for every violation of a rule (that is, every variable assignment under which the formula is false). Eventually, the ratio between the probabilities of two structure is computed as the exponential of the difference between their incurred penalties.

This gives us a new, logical perspective on the distinction between directed and undirected representations: While directed approaches view logical formulas as definitions of concepts, inducing a direction to the term defined from the terms used in the definition, undirected approaches view formulas as constraints to be respected, without any hierarchy among the prdicates in the language.

This distinction is mirrored in the tools used to deal with such representations: While finite model theory with its asymptotic analysis and the arsenal of descriptive complexity theory lends itself easily to the study of directed models, On the other hand, inference in Markov logic networks allow a more straightforward application of first-order weighted model counting. In this sense, the translation of inference in probabilistic logic programs to a problem in first-order weighted model counting, replacing definitions with constraints, implies a forgetting of the directionality inherent in the program structure.

Our observation ties in beautifully with Judaea Pearl's calculus of causal reasoning. There, causal Bayesian networks as directed models suffice to calculate the effect of external interventions, but to ascend to true counterfactual reasoning, we need structural causal models. Structural causal models are nothing but definitions of the child random variable in terms of the parent variables and error terms—the very idea behind probabilistic logic programming.

Felix Weitkämper
Computer Science, LMU Munich

Uncertain Reasoning

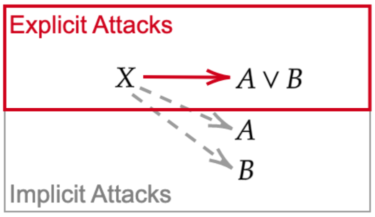
Let us consider the following statements:

X : Gas price rises;

A : Hybrid car sales fall;

B : SUVs sales grow.

Then, we could agree that whenever the gas price rises, it is not the case that hybrid cars sales fall or SUVs sales grow, i.e. X *attacks* $A \vee B$, in symbols $X \longrightarrow A \vee B$. X and $A \vee B$ should be understood as claims of two different arguments where arguments are entities made of three parts: the support, the claim (or conclusion) and the method of inference between the support and the claim. Suppose now that these two arguments, presumably, belong to a bigger argumentation framework and in principle, there might be not just one, but a class of them. Therefore, by only looking at the explicit attack $X \longrightarrow A \vee B$, we shall say that the implicit attacks $X \longrightarrow A$ (whenever the gas price rises, it is not the case that hybrid car sales fall) and $X \longrightarrow B$ (whenever the gas price rises it is not the case that SUVs sails grow) should also belong to the class of argumentative frameworks compatible with $X \longrightarrow A \vee B$.



These kinds of rules referred to as Attack Principles (APs), have been introduced in (Corsi and Fermüller, 2017), they have been defined for all main four connectives and they refine the existence, or not existence, of specific attack relations once the arguments involved share in their claims some propositional formula. APs are defined on an intermediary level of abstraction between Dung-style argumentation frameworks where both the arguments and the attack relation are abstract entities (Dung, 1995) and deductive or logical argumentation frameworks where the arguments are defined as above and also the attack relation is instantiated using some logical inferences that the argument involved might or might not satisfy (Besnard and Hunter, 2001). Even though the above attack principle seems very reasonable and easy to formally justify, this might not always be the case. For example, the strong attack principle for conjunction states that if an argument with claim X attacks an argument with claim $A \wedge B$, then the former argument either attacks also an argument with claim A, or an argument with claim B. The use of the term *strong* for the attack principle just introduced refers at the fact that this principle, in contrast with one concerning disjunction previously introduced, is hard to justify. Thus, if our explicit attack is $X \longrightarrow A \wedge B$, we might be hesitant about how to close off the argumentative framework and consider compatible both the argumentative frameworks in which $X \longrightarrow A$ and $X \not\longrightarrow B$ and those in which $X \longrightarrow B$ and $X \not\longrightarrow A$. The symbol $\not\longrightarrow$ stands for *not attacking*.

The general approach in abstract and deductive argumentation theory is that, given an argumentation framework, identify which argument or set of arguments is more acceptable than others. Thus, all the arguments are known and the attack relations are explicit. Through the attack principles, we can change the perspective and given some attack relation among arguments, that might be seen as *evidence*, we are able to identify the class of frameworks compatible with it. However, some attack principles are easier to justify than others and depending on the set of attack principles we consider acceptable the class of argumentation frameworks compatible with the evidence might change. The understanding of the compatible argumentation frameworks given a specific attack can be seen as an *explanation* of the existence of that attack. E.g., going back to the example, a possible explanation for the attack $X \longrightarrow A \vee B$ is that the argument with claim X also attacks both the argument whose claim is A and the argument whose claim is B. This inference process can be related to *abductive reasoning* in which, given some data (or evidence) we infer the best explanation. In our specific example, the attacks $X \longrightarrow A$ and $X \longrightarrow B$ are naively acceptable and once the arguments and the attack relations are instantiated in logical argumentation frameworks, we can find a counterexample whenever the attack relation is, for example, *defeat*. In addition, in abductive reasoning, the conclusion does not deductively follow from the premises. In the context of argumentation theory and attack principles, it is not even clear which should be the logic of the arguments and if classical logic seems to be a good candidate, many are the reasons (e.g. the expressive power or the computational complexity) to consider weaker logics. In a recent work by Arieli, Borg, Hesse and Straßer, (Arieli et al. *Abductive Reasoning with Sequent-Based Argumentation.* Proceedings of the 20th International Workshop on Non-Monotonic Reasoning, Part of FLoC 2022) where arguments are understood as sequents, the authors introduce *abductive sequents* which are expressions of the form $A \Dashv \Gamma[\varepsilon]$ with the intuitive meaning that "(the explanandum) A may be inferred from Γ, assuming that ε holds". Abductive arguments are a new type of hypothetical argument that is subjected to potential defeats. Given its high degree of modularity, these new enriched sequent-based frameworks might represent a good starting point to formalise the argumentative reasoning depicted above. The explanandum could be understood as the attack on the argument with claim A, Γ could be instantiated with the attack on $A \vee B$ and $[\varepsilon]$ could be seen as the satisfaction of the attack principle **(A.** $\vee$**)** If $X \longrightarrow A \vee B$, then $X \longrightarrow A$ and $X \longrightarrow B$. Thus, given an attack principle and an explicit attack, the corresponding abductive sequent can be defined and a new argumentative framework that works at meta-argumentative level is introduced. Then, the several argumentation frameworks compatible with the initial attack relation can be characterised by the different abductive arguments that can be generated considering the different ways of "closing off" the initial attack considering the various APs.

Esther Anna Corsi
University of Milan

Events

March

HPS: Integrated History and Philosophy of Science, University of South Carolina, 16–18 March.

May

UnRe: Uncertain Reasoning, Clearwater Beach, Florida, 14–17 May.

June

LC2023: Logic Colloquium 2023, University of Milan, 5–9 June.

Courses and Programmes

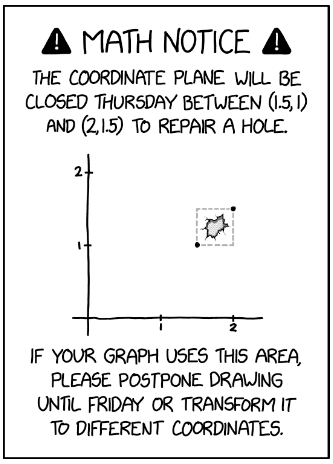
Programmes

MA in Reasoning, Analysis and Modelling: University of Milan, Italy.
APhil: MA/PhD in Analytic Philosophy, University of Barcelona.
Master Programme: MA in Pure and Applied Logic, University of Barcelona.
Doctoral Programme in Philosophy: Language, Mind and Practice, Department of Philosophy, University of Zurich, Switzerland.
Doctoral Programme in Philosophy: Department of Philosophy, University of Milan, Italy.
LogiCS: Joint doctoral program on Logical Methods in Computer Science, TU Wien, TU Graz, and JKU Linz, Austria.
HPSM: MA in the History and Philosophy of Science and Medicine, Durham University.
LoPhiSC: Master in Logic, Philosophy of Science and Epistemology, Pantheon-Sorbonne University (Paris 1) and Paris-Sorbonne University (Paris 4).
Master Programme: in Artificial Intelligence, Radboud University Nijmegen, the Netherlands.
Master Programme: Philosophy and Economics, Institute of Philosophy, University of Bayreuth.
MA in Cognitive Science: School of Politics, International Studies and Philosophy, Queen's University Belfast.
MA in Logic and the Philosophy of Mathematics: Department of Philosophy, University of Bristol.
MA Programmes: in Philosophy of Science, University of Leeds.
MA in Logic and Philosophy of Science: Faculty of Philosophy, Philosophy of Science and Study of Religion, LMU Munich.
MA in Logic and Theory of Science: Department of Logic of the Eotvos Lorand University, Budapest, Hungary.
MA in Metaphysics, Language, and Mind: Department of Philosophy, University of Liverpool.
MA in Mind, Brain and Learning: Westminster Institute of Education, Oxford Brookes University.
MA in Philosophy of Biological and Cognitive Sciences: Department of Philosophy, University of Bristol.
MA programmes: in Philosophy of Language and Linguistics, and Philosophy of Mind and Psychology, University of Birmingham.
MRes in Methods and Practices of Philosophical Research: Northern Institute of Philosophy, University of Aberdeen.
MSc in Applied Statistics: Department of Economics, Mathematics and Statistics, Birkbeck, University of London.
MSc in Applied Statistics and Datamining: School of Mathematics and Statistics, University of St Andrews.
MSc in Artificial Intelligence: Faculty of Engineering, University of Leeds.
MSc in Cognitive & Decision Sciences: Psychology, University College London.
MSc in Cognitive Systems: Language, Learning, and Reasoning, University of Potsdam.
MSc in Cognitive Science: University of Osnabrück, Germany.
MSc in Cognitive Psychology/Neuropsychology: School of Psychology, University of Kent.
MSc in Logic: Institute for Logic, Language and Computation, University of Amsterdam.
MSc in Mind, Language & Embodied Cognition: School of Philosophy, Psychology and Language Sciences, University of Edinburgh.
MSc in Philosophy of Science, Technology and Society: University of Twente, The Netherlands.
MRes in Cognitive Science and Humanities: Language, Communication and Organization: Institute for Logic, Cognition, Language, and Information, University of the Basque Country (Donostia San Sebastián).
Open Mind: International School of Advanced Studies in Cognitive Sciences, University of Bucharest.
Research Master in Philosophy and Economics: Erasmus University Rotterdam, The Netherlands.
Doctoral Programme in Philosophy: Language, Mind and Practice, Department of Philosophy, University of Zurich, Switzerland.
MA in Philosophy: Dept. of Philosophy, California State University Long Beach.

Jobs and Studentships

Jobs

Assistant Professor: in Philosophy of Science, Radboud University, deadline 12 March.
Tenure track position: in AI Ethics @TU Wien, deadline 16 March.


MATH NOTICE
THE COORDINATE PLANE WILL BE CLOSED THURSDAY BETWEEN (1.5,1) AND (2,1.5) TO REPAIR A HOLE.
2
1
1
2
IF YOUR GRAPH USES THIS AREA, PLEASE POSTPONE DRAWING UNTIL FRIDAY OR TRANSFORM IT TO DIFFERENT COORDINATES.