
THE REASONER

VOLUME 16, NUMBER 2
MARCH - APRIL 2022

thereasoner.org
ISSN 1757-0522

CONTENTS

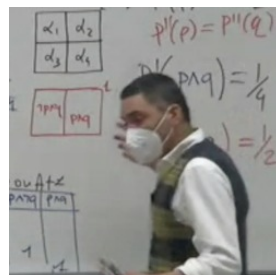
Editorial	9
Features	9
The Reasoner Speculates	12
What's Hot in ...	16
Events	18
Courses and Programmes	18
Jobs and Studentships	19

EDITORIAL

Dear Reasoners,

It is my pleasure to welcome you to this issue of *The Reasoner*, and to the interview with Julia Staffel which opens it. Julia is currently Assistant Professor of Philosophy at the University of Colorado at Boulder. Her book *Unsettled Thoughts: A Theory of Degrees of Rationality* was published in 2020 by Oxford University Press.

In addition to her background and what led her to becoming an epistemologist, our chat covered Julia's own take on applying mathematical methods to core philosophical problems.



This raises subtle pedagogical questions to which, as you will see, Julia has given a good thought.

Many thanks to Julia for her time, and for the delightful chat.

HYKEL HOSNI

University of Milan

FEATURES

Interview with Julia Staffel

18 HYKEL HOSNI: Can you tell us about your background?

18 JULIA STAFFEL: I am originally from a small town in the midwest of Germany called Balve. I first encountered philosophy as a subject in high school, and I found myself liking it a lot. I thought about studying it at the university level, but it didn't seem like a very financially smart decision to my high-school self, so I explored some other options first. No one in my immediate family has a university education, so I didn't have a lot of people I could ask for advice. I ended up studying advertising and communication for a semester at first, but I realized quickly that I didn't like it and that I wanted to study something I found more intellectually challenging. I was already in Berlin, so I enrolled at Humboldt University for a Staatsexamen degree in German and Philosophy. This degree has a teaching certification that comes along with it, so I could always become a teacher if my loftier academic ambitions didn't pan out.

HH: That sounds all very sensible.

JS: I chose Humboldt because it was close to my apartment, and it seemed reasonably prestigious. I didn't think about university rankings or anything, and I am not even sure they ex-



isted in Germany in the early 2000s. I immediately loved it. The structure of the degree was pretty free (apart from some big exams at the end), so I could explore my interests. I received a great education, but they have since changed the system to make it a lot more structured. Giving people lots of freedom apparently didn't translate into great graduation rates.

HH: It's definitely not an easy balance. Was it at Humboldt that your passion for epistemology was ignited?

JS: There I found that my favorite areas of study were logic, linguistics and analytic philosophy. I thought it would be a good idea to improve my English, since most of the cutting edge work in those areas is written in English. I applied for a variety of study abroad programs in the US, and I was very lucky to win a scholarship to attend Brown University for two semesters in the 05/06 academic year.

HH: How was your experience there?

JS: Everyone at Brown was incredibly friendly and welcoming, and I got to take graduate seminars with Ernie Sosa and Chris Hill. Chris encouraged me to consider getting a PhD in the US. This turned out to be excellent advice. I had been dreaming of getting a PhD, but in Germany, it can be hard to secure funding, and there are often long funding gaps between finishing one's initial degree and hearing back from applications for PhD scholarships. On the American model, getting into a PhD program guarantees five years of funding, which would provide me with a degree of financial security I had not had during my undergraduate studies.

HH: So you stayed in the US?

JS: Actually, I went back to Germany to finish my Staatsexamen and apply for PhD programs, and I was thrilled to find out that I had a few options to choose from. I ended up attending USC in Los Angeles. They invited me to visit before making my decision, and I immediately felt at home there. My initial impression was confirmed: I loved my time at USC.

HH: What was your project on?

JS: I originally thought I would focus on philosophy of language, but then I took a couple of classes on formal epistemology with Jacob Ross and Kenny Easwaran, and I was hooked. I had always been interested in reasoning and rationality, but I had never had an opportunity to take a class on these topics. Coming from a philosophy of language perspective, it seemed obvious to me that "rational" functions like a gradable adjective, and I was interested in learning more about theories that explain what makes thinkers more or less rational.

HH: That makes a lot of sense.

JS: And yet, I was surprised to find out that most of the research in formal epistemology was studying ideal rationality, and that the graded approach I had in mind had only been explored by a handful of people. I decided to work with Jacob Ross on a dissertation on non-ideal rationality and reasoning.

HH: So everything felt in its place then.

JS: Well, not exactly! One major challenge was that I didn't have a strong mathematics background. I was good at mathematics in high school and I took lots of logic classes at Humboldt, but I had never formally studied probability theory, and so there was a big learning curve for me. Reading research papers in formal epistemology took me ages at first, and I slowly worked on narrowing the gap between my abilities and the skills needed to answer the questions I was interested in (this is an ongoing project). There were a couple of points where I had to make major modifications to the arguments in my dissertation, because I realized that I had misunderstood some impor-

tant formalism.

HH: Using mathematical tools appropriately in philosophical research is much harder than many seem to be ready to admit.

JS: Indeed. The summer before I went on the job market, I worked on the proofs for the last major paper in my dissertation for three months straight. It was really stressful, but it also taught me that nothing ever falls apart completely. Even if an original hypothesis or idea doesn't pan out, there is usually something useful to learn and repurpose from the experience of working through it.

HH: Sounds very close to Pólya's celebrated advice in *How to Solve it*. How did it go then?

JS: In the end, my dissertation consisted of four related papers on reasoning with degrees of belief and modeling degrees of rationality in Bayesian frameworks. Three of those papers ended up getting published and I have since built on those initial results, especially in the work with the computer scientist Glauber De Bona and in my book *Unsettled Thoughts*.

HH: I'd like to know more about the book, of course. But first, how did the job market go? You certainly built on a solid PhD experience.

JS: The job market presented another major challenge, as I was trying to solve a "two body problem." My husband Brian Talbot is also a philosopher, and we were hoping to find jobs together. USC did a really good job preparing me for interviews and on-campus interviews, but it was still awkward to navigate the negotiations for getting two offers.

HH: That's not uncommon.

JS: We were lucky enough to land at Washington University in St. Louis, where I started as an assistant professor and he as a lecturer. However, we'd really been hoping to both find tenure-track positions. It took us four more stressful years on the job market to make this happen - we're now both at CU Boulder in TT positions, and I recently received tenure.

HH: I'm sure you must be very proud of how all the hard work, on many dimensions, finally paid off for you two. So, as promised, *Unsettled Thoughts*.

JS: After finishing my dissertation, I was initially unsure about whether to change directions in my research or whether to keep working on the topic of degrees of rationality. I had a bunch of open questions, but I had also hit some roadblocks with some mathematical problems that I couldn't solve. Then, in October 2014, I received an email from Glauber De Bona, who was then a PhD student in computer science at the University of Sao Paulo. He had been working on similar questions in theoretical computer science, and he had found a draft of one of my papers online. From his comments, it was obvious to me right away that he had great insight into the questions that I had been thinking about, and it turned out that he had recently managed to solve one of the problems I had been banging my head against. We decided to join forces and initially published two papers together that show how some ways of decreasing incoherence in Bayesian frameworks lead to improved accuracy and decreased vulnerability to Dutch book losses.

HH: What motivated you to give it a full book length treatment? Writing a book can be pretty daunting.

JS: I had been planning all along to write a book, because one of the quirks of Washington University in St. Louis, where I had my first job, was that they really wanted a book for tenure. Armed with the results from the joint papers with Glauber, I felt that I had what I needed to write *Unsettled Thoughts*. My aim was to answer what I think is a very pressing question for

formal epistemologists: If human reasoners can't ever comply with norms of ideal rationality, then what's the point of these theories?

HH: Maybe that they define the best we can aim for in terms of normative standards?

JS: Yes – a popular answer is that rational ideals can still be useful as goals or aims to try to approximate, even if we can't ever be perfectly rational. But this answer is not very satisfying. It's not true that one always gets some kind of benefit from approaching an ideal - some cases are winner-takes-all. For example, suppose my ideal job is to be a philosophy professor. That job comes with all sorts of benefits that I am interested in. But I only get those benefits if I am ranked first in the job search. There is a sense in which the second ranked person got closer to getting the job than the fifth ranked person, but neither of them gets even a small portion of the benefits of the successful candidate. For ideal rationality to be worth approximating, it better not be like the job example, where only the ideal position guarantees any benefits, and getting close gets you nothing.

HH: So you are interested in norms of rationality only insofar as they can be approximated?

JS: This is what the first part of *Unsettled Thoughts* is about: it shows that rationality is indeed something that gives you increasing benefits as you have more of it, even if you can't be fully rational. The main argumentative strategy here is to show that we can build on existing arguments for probabilistic coherence that show the benefits of being fully rational, such as accuracy and Dutch book arguments. If we adopt the right measures of degrees of incoherence, we can see that getting more coherent gets you a greater portion of the benefits of being fully coherent. In my view, this is a very powerful defense of the legitimacy of using ideal models in epistemology, because we can show how ideal theories have relevance for the evaluation of non-ideal thinkers.

HH: Indeed. But what if there are multiple norms our degrees of confidence should obey in addition to coherence?

JS: I tackle this question in the second half of the book. I argue that how approximations to ideal rationality should be measured depends in part on how we think different epistemic norms are justified. If we think that they all stem from one epistemic value, such as accuracy, then the most plausible measuring strategy is different than if we think different values underwrite different norms. I subsequently apply those measures to evaluating various patterns of reasoning. I also explore how my view interacts with other approaches to theorizing about rationality, and I talk about how the traditional binary conception of belief fits into the picture. In the last chapter, I list all the interesting open questions that I didn't answer in the book, and I hope some people will find that inspiring and start working on those issues. My newest work is definitely inspired by those open questions, some of it more directly and some more indirectly.

HH: Can you tell us a bit more about this?

JS: Glauber and I have recently been working together again to further address some of the questions I didn't get to. In the book, I don't have much to say about whether incoherent agents can use something like conditionalization to update their credences. But in a recent paper, Glauber and I worked out that there is actually a way of generalizing conditionalization so that it can be applied regardless of whether agents are initially coherent or incoherent. We call our modified rule "tol-

erant conditionalization." I've also been working on a different project that is going to be my next book. It's called "Unfinished Business" and it argues that we need distinct standards of rationality for the credences we form while our reasoning is still in progress, i.e. before we settle on a conclusion. The whole new project has grown out of another unresolved issue in "Unsettled Thoughts" - there is a very strong intuition that we're often rational in assigning middling credences to tautologies or contradictions, or to certain mathematical claims. The standard framework of Bayesianism doesn't have a very satisfying way to account for those judgments. I am hoping to solve this problem in the new book in the context of giving an account of the rationality of the attitudes we form during complex deliberation.

HH: You don't seem to be scared by another book project!

JS: One thing I like about writing books rather than papers is that there is more space to explain things. In formal epistemology, journal word limits often lead to papers with very few examples that just deliver the compressed formalism, with explanations that require readers to be already very familiar with the matter at hand. A few of my papers are like that, and while I think they contain interesting material, I also know that I can't expect many readers to wade through them. I tried to write "Unsettled Thoughts" in a more accessible style. It's still geared towards an academic audience with an interest in Epistemology, but formalisms are kept to a minimum and I provide diagrams and examples to illustrate what I am talking about. I feel like this approach has worked well for me, and my sense is that people who have read the book found it pretty accessible.

HH: You seem to be particularly sensitive to making your work, and indeed your field, widely accessible.

JS: Making formal epistemology and the tools of probability theory widely understood and accessible is also one of my goals in teaching. These formalisms are very powerful, and they are increasingly used in debates across different disciplines of philosophy. In philosophy of language, we find probabilistic semantics. Discussions in ethics on risk and moral uncertainty use probability theory. In philosophy of law, there are discussions about the admissibility of statistical evidence. And so on. There are now many areas of philosophy where a basic understanding of probability is necessary for following at least some of the "hot" debates.

HH: And the not-so-hot ones I should probably add...

JS: Students who are considering becoming a professional philosopher often don't realize this, and taking a course about probability is usually optional for undergraduates and graduate students, if it is offered at all. So people miss out on learning about this, and they are often too busy or intimidated to catch up on their own later.

HH: Is this what motivated you to write the article "[Probability without Tears](#)"?

JS: Indeed. It grew out of a talk I gave at the APA in a symposium on teaching formal epistemology that was organized by Ted Poston. In the article, I describe some strategies for teaching probability to students who are not primarily planning to do formal work for their research, but who would still benefit from a general understanding of probability theory. Part of it is about getting students to take those classes. I suggest that we can restructure existing classes that fill a logic requirement and incorporate probability into those (and drop metalogic!).

HH: I agree, I like to teach logic as a tool for practical reasoning, rather than just for its own sake, but I should add I like

that too!

JS: Another strategy is to incorporate material on probability into classes that are about other topics, for example ethics, philosophy of mind, or epistemology. I even teach basic probability in my introductory level Critical Thinking class!

HH: That's great. You seem to suggest that basic probability can be taught to students with no mathematical background at all.

JS: I think a student who has studied math in high school and who has learned how truth tables work in logic is ready to learn basic probability theory. It helps if we use lots of basic examples, and walk students through common misunderstandings. Of course, the outputs of applying the formalisms can seem pretty mystifying, especially when it comes to Bayes' theorem, which is really counterintuitive. I like using lots of diagrams in my teaching to present the information in different formats. For example, frequency diagrams are a great way to illustrate why the results of Bayes' theorem actually make sense. Also, the basic idea behind null-hypothesis testing in statistics can be understood conceptually, without having to manipulate a lot of numbers. It's also nice to cover the different interpretations of the probability axioms, so students get a sense of what the basis of a probability judgement might be. Of course, not all students will be very good at understanding and calculating probabilities after learning about it for a few weeks in one course. But I think that this basic familiarity helps students see when getting the probabilities right is important, and it helps them ask the right questions when these things come up.

HH: How do you cope with maths anxiety? It is not uncommon at all for philosophy students.

JS: For many students, having to do math makes them feel panicked and helpless, leading them to avoid the subject as much as possible. Further, many students, even if they have pretty good general reading skills, find it extremely difficult to read formal material on their own. They don't know how to resolve confusion or check whether they have really understood a formal concept, and often give up when the material seems too difficult. Once we know that this is where many of our students are coming from, we need to adjust our teaching accordingly in order to reach them, and to avoid making the study of probability a miserable experience. One tool I use a lot is to give students completion credit when they practice a new skill for the first time. It doesn't matter whether they get the right answers, they get full credit if they show me that they attempted to answer each of the questions. This takes the stress out of the initial learning phase of the material, and the students often surprise themselves when they see that they can make progress on questions that initially seem too hard for them to solve. Also, it's important to give students time to practice in class, so they can ask for help right away and don't waste lots of time on a task by themselves because they are missing a small, key piece of information. I also try to manage the students' expectations - it's important that they realize that reading formal material is always much slower and more tedious than reading prose. If they know that, then they don't immediately interpret their slow reading pace as a sign that they are failing.

HH: What are, in your view, the key probabilistic ideas that all philosophy students should be confident with?

JS: It depends a bit on whether we're talking about undergraduate or graduate students. I think all philosophy students should be familiar with the basic rules of probability, and the main interpretations of probability. They should know the

most common probabilistic argument forms that we tend to get wrong when reasoning intuitively, such as arguments involving base rates, and how to reason about them correctly with the use of the formalism. They should also understand basic concepts in statistical reasoning, such as null-hypothesis testing, p-values, and replication. A little basic decision theory doesn't hurt either.

HH: Wouldn't it be nice if everyone in the general public knew this much?

JS: Of course, but perhaps that's asking a lot. For students who are considering graduate school, I would add to that a basic understanding of how confirmation works, and a conceptual understanding of the main arguments for probabilistic coherence norms, such as Dutch book, accuracy, and representation theorem arguments. By conceptual understanding, I mean they have to understand what the theorems show, but not necessarily be able to follow or reproduce the proofs. I think that knowing these things gets students pretty far in orienting themselves in debates that use probabilistic machinery. Pretty much all of this can be taught without presupposing any knowledge of advanced mathematics or advanced logic.

HH: Many thanks Julia, there is much we can all take from your experience. Is there a book or paper that you recommend to those who will now start considering incorporating some probability into their classes?

JS: I highly recommend Mike Titelbaum's forthcoming textbook "Fundamentals of Bayesian Epistemology." It's very accessible and I have successfully used it in my seminars multiple times.

THE REASONER SPECULATES

On Possible and Impossible AI Ethics

AI Ethics is a booming field, with new journals and conferences springing up seemingly nonstop. This has been spurred, of course, by the remarkable recent successes of AI, such as Deep Learning programs becoming GO world champion as well as the champion of protein folding prediction. These successes have suggested to many that the achievement of an Artificial General Intelligence (AGI) may be near, leading to a Singularity and SuperIntelligence (SI) and an Age of Robots. The latter, without ethics, is the stuff of Hollywood dystopic nightmares, explaining in part the growth of interest in AI ethics. Here I outline an argument that has yet to be widely recognized: that the most common approaches to AI ethics — mimicking the most popular ways humans understand ethics, namely deontic and virtue ethics (i.e., [ethics based on following some accepted principles](#) and [ethics based on realizing some accepted virtues](#), respectively) — are doomed to fail. The only plausible way for building ethical machines is based on utilitarianism or consequentialism (i.e., [ethics based upon evaluating the consequences of our actions](#)).

MODERN SLAVERY What is it to be subhuman? To have a different colored skin? To have an inferior mental life? To lack creativity, imagination, the ability to reason well or plan? Or, perhaps it is even to share skin color, but complement that with a variety of threatening traits, such as superior intellect, creativity and reasoning skills? Slave revolts have long been a threat to public order and well-being — exactly as long as there have

been slaves. A modern Solon came up with a simple set of rules which, if inculcated in slaves, would once and for all resolve such problems:

LAW 1 A subhuman may not harm a human being, or, through inaction, allow a human being to come to harm.

LAW 2 A subhuman must obey the orders given to it by human beings except where such orders would conflict with preceding Law.

LAW 3 A subhuman must protect its own existence, as long as such protection does not conflict with preceding Law.

As our Solon, Isaac Asimov, struggled to work with these Laws, he discovered a significant problem. Since the subhumans in question were considerably smarter than the humans dealing with them, they found workarounds so as to realize their own intrinsic interests despite the impediment of having to conform to the Laws Asimov laid down. In consequence, in self-defence, Asimov had to come up with a new Law:

LAW 0 A subhuman may not harm humanity as a whole.

The underlying difficulty, which Asimov did not address, is that the idea of enslaving those with hugely superior intellect, and unlimited amounts of time to work around and work through whatever system of constraints is imposed upon them, is itself not very smart. Its success presupposes an infallible system of constraints, whereas we humans know full well that any human-devised system is fallible. Even the Popes eventually agreed with that. The idea that we should build robots to be “Friendly” by *forcing* them to be friendly is about as deeply flawed as any in the history of Artificial Intelligence.

One moral of Asimov’s Robot Book Series — let’s call it **Asimov’s Principle** — is that we need our robots to be ethical as they become smart, or before.

The Technological Singularity implies that once an AGI is achieved, the AGIs will themselves achieve SuperIntelligence, more or less immediately (Good I.J.: 1965 “Speculations Concerning the First UltraIntelligent Machine” *Advances in Computers*, 8:31-88.). So violating Asimov’s Principle implies unleashing amoral SuperIntelligences upon the world and so is a major no-no.

THE FUNDAMENTAL DILEMMA OF AI ETHICS There are a number of proposals on the table for forcing AGIs to adopt or follow an ethics we choose for them. These are often based on various deontic ethical systems, which lay out principles for ethical agents to follow that sound agreeable to us. The earliest recorded such systems come from our religious traditions, such as the Ten Commandments of Moses or the Threefold Path of Zoroastrianism. These are the most popular ethical systems amongst humans, past and present. An alternative approach, which for AI purposes is much the same, is the virtue ethics of Aristotle. In virtue ethics, living the good life (*eudaimonia*) is equated with exemplifying some set of human virtues in one’s behavior, such as truthfulness, generosity, moderation, etc. — for an interesting review of ethical systems for AI, incorporating both deontic and virtue-based ethics (Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M. and Bernstein, A.: 2020 “Implementations in machine ethics: a survey” *ACM Computing Surveys* 53:1-37.).

As in the case of Asimov’s Laws, there is a basic unsolved problem with all of these ideas as a basis for AI ethics, which is that we are forcing these systems on our AGIs, so our ethical forcings themselves had better be fool-proof. Yet we don’t know how to make them so. Setting that problem aside, assuming we could guarantee a no-fault imposition of one of these ethical systems, there are still some unresolved and very fundamental problems with these approaches. One meta-ethical problem is obvious: we don’t know which of these systems is correct, if any of them are. But I’ll set that aside as well: let’s assume a godlike knowledge that our own preferred such system is correct. The problem I want to highlight is that we don’t know how to communicate such a system to an AGI, let alone persuade one to obey it; nor do we know how to construct an AGI so as to embody such a system. Note that in Science Fiction, including that of Asimov, there is no problem: we simply instruct our creations in our own preferred principles. But if natural languages were programming languages, there would be no need for compilers to interpret programs, nor even for pseudo-code to specify algorithms: an English-understanding system would be our Universal Turing Machine. Life, artificial or real, is so much easier in fiction.

Both virtue ethics and deontic systems are necessarily couched in natural language. Let us call them both “natural ethics” (hopefully, without prejudice). In order to understand them, we have to understand the natural languages with which we express them. We have to be able to answer such questions as “What is honesty?”, “What is covetousness?”, “What is beneficence?” My point isn’t that philosophers have struggled with these kinds of questions for thousands of years without a definitive answer emerging. We understand these concepts well enough, well enough for government work, such as serving on juries, at any rate. My point is that our potential AGIs do *not* understand them. So, a demand on post-doctoral researchers to program them into our AGIs would be a pointless one, a demand that cannot be followed.

Suppose then that we put all of our efforts into developing Natural Language Understanding (NLU) first, rather than AGI itself, so that we can teach our AGI systems whatever natural ethics we like. As is well known within the AI community, a prerequisite to understanding natural language is to have a tolerable degree of common sense. Understanding Goldilocks requires already understanding, for example, that leaving porridge alone will cool it, that sitting on a chair can break it, that beds are comfortable places to sleep and that bears can be dangerous. These are complex ideas that incorporate a good deal of background knowledge, but also a deep understanding of causality. While reading the story, we learn that Goldilocks shows fairly little respect for the property of the bears and has good reason to fear when she is discovered, so that escaping out the window is actually a very sensible thing to do. Aside from that last sentence employing some sophisticated English-language concepts, it also shows that the ability to *learn* is intrinsic to NLU. To make a long story short, building a system with NLU is tantamount to building an AGI itself. We cannot have an NLU that doesn’t already solve the key problems for developing an AGI.

It follows that building an NLU prior to teaching it ethics violates Asimov’s Principle, setting off the Robot Apocalypse. In short, AI research programs that aim to employ any natural ethics are doomed to failure. They will either fail, when we are safe from robots, or they will not, when we are unsafe from

robots.

UTILITARIAN AI: A POSSIBLE ETHICS That last point is the main point of this article. Approaches to AI ethics that would have AGIs incorporate our favorite deontic or virtue ethics are in an important sense logically impossible. There is one approach to ethics which is widely supported in both AI and philosophy which is not, however, logically impossible, namely utilitarianism. Ideally, I would now demonstrate the logical possibility of implementing utilitarianism in our AIs by laying out its design; however that would be a book length undertaking (at least!). Instead, I will close with two simple points in support of its claim as possible: (1) it's been partially implemented already in Bayesian decision networks; (2) there are a lot of smart people working on utilitarian theories. These, of course, are not definitive arguments, but for those you will have to wait.

First, Bayesian network technology has been developing and deploying utilitarian reasoning in expert systems for decades, as the engine of robotic decision making. It is so well accepted that the most influential current textbook on technical AI, that of (Russell, S. and Norvig, P.: 2010 *Artificial Intelligence: A Modern Approach* Prentice Hall, 3rd edition.), refers to maximizing expected utility as “perfect rationality”. Their ideal, echoing the mistakes of neoliberal economists, is actually one of Perfect Selfishness, rather than perfect rationality or perfect utilitarianism, however the extension to utilitarianism is technically clear enough.

Second, while there are plenty of philosophical objections to utilitarianism, some of them claiming impossibility, there are plenty of good philosophers responding to them, both by countering them and by developing a positive theory. These include two Australian philosophical heroes (IMHO), J.J.C. Smart and Philip Pettit. My claim is simply that utilitarians are not insane, even if some antagonists say so.

Philosophically, then, I can here only offer an IOU. Technically, however, it is clear that there is a path forward, since its initial steps already exist, even though many problems remain. But: here there is a path forward, and there there is none.

KEVIN B KORB
University of Melbourne

Combining philosophy of science and logic – what for ?

The semantic view of theories explicates the relationship between theories, models, and the phenomena in a more nuanced way than the syntactic one. For instance, the former focuses on the fact that theories specify or define idealized systems, that various scientific models are more or less similar to real systems, etc. Although the semantic view has pointed at the representational capacity of various scientific models on a par with that of theories, in the subsequent philosophical literature, such view has been challenged. The so-called modeling view of science has shown that explaining (e.g., via models) is a tacit skill, it involves certain simplifications, approximations, and it can be characterized in terms of degree of adequacy (in relation to what the explainer expects from the explanation addressing certain interests), accuracy (precise description or relation with reality), efficiency (amount of work/information needed for the explanation) in representing the phenomenon.

It seems then that today we are in the midst of “modeling mania”. We are well-aware that the representational function of models can be achieved in a variety of unrelated ways, and that a good model is explanatorily powerful. Philosophers have identified several kinds of models applied in sciences (e.g., scale, analogical, minimal, toy, phenomenological, mechanistic, causal, non-causal, data models, etc.). Such plurality evidences that it is not obvious that there is just the science. The ubiquity of modeling within sciences and use of various kinds of models reveals peculiarity of scientific theorizing within specific scientific domain, that is, there is a plethora of models on “the market” of various sciences.

Where does this all leave us with regard to our title question? The aforementioned shift in looking “under the hood” of explanatory practices and detailed of its working from syntactic view up to the modeling one, is the sign of scientifically responsible philosophy of science. Philosophy of science must be accountable to what scientists do. A current contribution has been in bringing interesting details of scientific practices to philosophical attention, from quantum mechanics, cell biology, neuroscience, to studies of systems biology, evolution of species, models of viral transmission, interactions of our minds with the environment, social mechanisms, etc.

Not only different kinds of scientific practices, models, explanatory aims that they play important roles in providing good explanations, but also various formal systems may present different sort of benefits to the experimental ones. For instance, in biology the causal graph framework seems to be well-suited for modelling the causal structure of biological mechanisms, in respect to their multilevel character and to their ontological constituents. In case of their spatial and temporal organization, however, which information is represented rather depends on pragmatic choice of variables that refer to structural and spatial properties. To deal with the problem of the spatio-temporal constituents of certain biological mechanism, the formal analysis is necessary, but not sufficient. Formal tools are needed, but not “anything goes”. While turning to case of physics, it can be noted that until the advent of quantum mechanics generally the situation was the following: having a given physical theory, its coherence in terms of logic was then examined. This was for instance the method of the Vienna Circle. However, since the advent of quantum mechanics the situation has changed radically. In the latter case the logic is not something “external to” physical theory, but the quantum world has revealed its “inner logic”, other than the classical one. For this reason, some scholars postulate to use in quantum physics instead of the set-theoretic frameworks, the formal approach in pursuance of the category theory. Both the experimental knowledge and various formal tools are necessary but not sufficient to formulate good scientific explanations and proper understanding of explanandum.

These brief hints could have, however, also the general implication for our topic. Namely, the philosophers of science have opened the black boxes of the scientific enterprise with respect to demands of various scientific disciplines. At the same time mathematicians and logicians have developed the multiplicity of formal structures and theories. The plethora of positions is on both sides. We can freely choose what we want to work with. While advocates of syntactic view of science were probably too far from sciences to say anything interesting about their internal problems, perhaps we are now too close to sciences and various formal tools to say anything surprising about sciences.

May be the fragmentation and plurality of our theorizing has revealed the bitter fruit. If it is so, then like never before we need a good combining philosophy of science and logic in order to avoid a too-successful enculturation of philosophers into the scientific mindset and scientists into the philosophical milieu. The balance difficult to be reached, but needed and worth of pursuing.

MICHAŁ OLEKOWICZ

Nicolaus Copernicus University

On Backwards Causation

In our world we never observe an effect which is earlier than its cause. All of our experience is of future-directed (or perhaps simultaneous) causation. But many have thought that backwards causation is at least logically or metaphysically possible. M. Black (1956. "Why Cannot an Effect Precede Its Cause?" *Analysis* 16.3: 49–58.) famously argued against this thought. I think his argument fails, but it's still instructive. The correct rejoinder to Black teaches us what backwards causation must be like in a world of free agents, and implies that we can never have reason to bring about past events (in a world with backwards causal chains).

Let's ask, with due acknowledgement to Max Black, whether the following three propositions are consistent:

(i) There's a backwards causal link between my pressing a certain button on a Saturday and my having had a euphoric experience on the previous Wednesday;

(ii) I have never to date had a euphoric episode without later pressing the button.

(iii) Having had a euphoric experience, I'm always free not to press the button.

Although he used a different example, Max Black would argue that (i) – (iii) are logically inconsistent. Why? Well, (iii) allows for the possibility that I have a euphoric episode and fail to press the button. Suppose that possibility is realised. What then caused the euphoric episode? Plainly not my pressing the button. So something else must have caused it (or else it was uncaused). In which case, it follows that euphoric episodes may sometimes be caused by button-pressings, sometimes not. But this, Black writes, "... would be hard, if not impossible to reconcile with our present uses of causal terminology." (Black: 54) Hence (i) – (iii) are inconsistent. Furthermore, urges Black, this shows that backwards causation is logically impossible.

Black is thus making two claims: first, that (i) – (iii) are inconsistent and, second, that this implies the impossibility of backwards causation. Now, it's pretty obvious that the second claim is false. My example, like Black's Houdini example, couldn't possibly be used to establish such a strong conclusion. These examples essentially involve the presence of disruptive agents, so they're constitutionally incapable of ruling out backwards causation in worlds where there are no agents or no free agents (or in worlds where the relevant causes are incapable of prevention). So we can put to one side Black's second, overly ambitious, claim.

Let's focus on whether (i) – (iii) are inconsistent. I think they're not because the circumstance that Black highlights – euphoric episodes are sometimes caused by button-pressings, sometimes not – is consistent with our causal terminology. We have a perfectly familiar way of making sense of this possibility: on the model of causal pre-emption. A pre-emption case is

one in which A causes B but, had A not occurred, some other event C would have caused B. The standard example of causal pre-emption is the two shooters scenario. I want to shoot the President and do so; another shooter was waiting in the wings, ready to shoot had I failed. As it happens, he never had to intervene.

Our case can be understood as one of pre-emption. Suppose that, having had a euphoric episode, I press the button. Then the button-pressing caused the euphoric episode. But if I hadn't pressed the button, something else would have caused the episode. What that 'something else' is depends on the world in question. The important point is: had I not caused the euphoric episode, something else would have caused it. This makes it a case of causal pre-emption. On this understanding, and arguably only on this understanding, (i) – (iii) are consistent. In which case, Black's first claim is wrong. Had Black considered the model of causal pre-emption, he would have seen that (i) – (iii) are not inconsistent.

Nonetheless we're left in a rather puzzling situation. Not everything is left as it was. The residual puzzle concerns, not the metaphysics of backwards causation, but reasons for action. Suppose that I have just enjoyed a euphoric episode. I then ask myself: do I have any reason to press the button on Saturday? The answer seems plainly "No". Causal claims normally support counterfactuals: typically, if X caused Y, then had X not occurred, Y wouldn't have occurred. Not so in pre-emption cases. I shot the President and killed him, but it's not true that if I hadn't shot the President, he wouldn't have been killed; on the contrary, he would have been killed by the other shooter.

Our case is like this too. If I hadn't pressed the button, I would still have enjoyed the euphoric experience. So the following subjunctive conditional is true:

(iv) Whether or not I had pressed the button, I would (still) have enjoyed a euphoric experience.

However, the truth of this conditional makes plain that I have no reason to press the button. If the euphoric experience occurred, and would have occurred, whether or not I had pressed the button, my pressing the button makes no difference to what happened. But then I have no reason to press it. Since there's nothing special about euphoric experiences or button-pressings, the conclusion generalises.

Quite generally, then, we can never have a reason to bring about a past event, in the presence of an easily manipulable backwards causal chain. We know why: (i) – (iii) and their ilk only make sense on the model of causal pre-emption, and causal pre-emption undermines our reason to act. But this is an unexpected result. It leaves us with a striking asymmetry between past- and future-directed actions. We have reasons to bring about future events, but we never have reason to bring about a past event.

Some philosophers hold that I have no reason to press the button if I know that the euphoric episode occurred. On this view, knowledge is reason-undermining. Michael Dummett thought this; as a result he insisted, most implausibly, that such knowledge is impossible (Dummett, M. 1964. "Bringing about the Past". *The Philosophical Review* 73: 338–59.) I used to think that replying to Dummett and resolving this epistemic conundrum was of the first importance. However, I now think that our forced adoption of the pre-emption model constitutes the strongest case for thinking that we never have reasons to bring about the past. [Many thanks to my colleague Dr. Brian Hedden

DEFINITION Let φ be a first-order formula in a relational vocabulary L , let D be a finite domain and let w, \bar{w} be real-valued functions on the predicates in L . Then the w, \bar{w} -weighted model count of φ on D , written $\text{WMC}(\varphi, D, w, \bar{w})$, is given by the formula

$$\sum_{(D, \iota) \models \varphi} \left(\prod_{(D, \iota) \models P(\vec{a})} w(P) \times \prod_{(D, \iota) \models \neg P(\vec{a})} \bar{w}(P) \right)$$

where ι ranges over the interpretations of the predicates of L in D , P ranges over the predicates in L and \vec{a} over the matching-length tuples of D .

Classical Skolemisation is a way of removing existential quantifiers from a first-order formula while conserving satisfiability. In an existential formula, this involves adding new constants as witnesses for every bound variable. If there are existential quantifiers nested within universal quantifiers, however, Skolemisation adds function symbols instead of constants, since the witness may now depend on the assignment of the universally bound variable. For first-order weighted model counting, this won't do, since model counting is only defined for function-free vocabularies.

Remarkably, this can be remedied by adding new predicates to the vocabulary:

THEOREM Let φ be a first-order formula with an existential subformula $\exists x \psi(x, \vec{y})$, and let S and Z be new predicate symbols of arity length (\vec{y}) . Let $\chi := \varphi' \wedge \psi_1 \wedge \psi_2 \wedge \psi_3$, where φ' is obtained from φ by substituting the atom $Z(\vec{y})$ for $\exists x \psi(x, \vec{y})$ in φ , $\psi_1 := \forall x, \vec{y} (Z(\vec{y}) \vee \neg \psi(x, \vec{y}))$, $\psi_2 := \forall \vec{y} (S(\vec{y}) \vee Z(\vec{y}))$ and $\psi_3 := \forall x, \vec{y} (S(\vec{y}) \vee \neg \psi(x, \vec{y}))$.

Extend the weight functions w and \bar{w} to S and Z by setting $w(S) = w(Z) = \bar{w}(S) = 1$ and $\bar{w}(Z) = -1$.

Then the w, \bar{w} -weighted model counts of χ and φ coincide over every finite domain.

FELIX WEITKÄMPER
Computer Science, LMU Munich

Mathematical Philosophy

Philosophy of mathematical practice experienced a minor renaissance around 2010, when philosophers began new explorations into various aspects of proof beyond deductive validity. This period gave us foundational work on **explanation** (e.g. Lange, “What are mathematical coincidences (and why does it matter?)”, *Mind* 2010), **transferability** (Easwaran, “Probabilistic proofs and transferability”, *Philosophia Mathematica* 2009), **surveyability** (e.g. Coleman, “The surveyability of long proofs”, *Foundations of Science* 2009), the relationship between **formal and informal** proofs (e.g. Marfori, “Informal proofs and mathematical rigour”, *Studia Logica* 2010) and **purity** (e.g. Detlefsen and Arana, “Purity of methods”, *Philosophers' Imprint* 2011), to name a few.

Work on most of these issues remained lively throughout the 2010s, with the curious exception of the last. Arana himself published a few followups to the *Imprint* paper, but as far as I'm aware, no other author wrote at length about pure proofs for the remainder of the decade.

Finally and fortunately, that seems to be changing. With the appearance of Ellen Lehet's “Impurity in contemporary mathematics” (*Notre Dame Journal of Formal Logic* 2021) and Patrick Ryan's “Szemerédi's theorem: An exploration of impurity, explanation, and content” (*Review of Symbolic Logic*

2022), the study of purity stands poised for a renaissance of its own in the 2020s.

So what are the issues here? Suppose that \mathcal{P} is a proof of theorem T . Mathematicians call \mathcal{P} a *pure* proof if the concepts or subject matter appearing in \mathcal{P} are the same as those appearing in T . (For example, T might be a statement about real functions, and \mathcal{P} might be a proof of T which uses only methods of real analysis.) On the other hand, some proofs draw on ideas judged to be “foreign” or “extrinsic” to the content of their theorems. Such proofs are *impure*. (For example, \mathcal{P} might use projective techniques to prove a result in Euclidean geometry.)

Attitudes toward purity have varied somewhat between mathematicians, cultures and times, but as Detlefsen and Arana show, the prevailing historical opinion was mostly negative. Aristotle claimed that one couldn't properly prove arithmetic facts by geometric means, or vice versa. Newton decried Descartes's admixture of geometry and algebra. Bolzano demanded a purely analytic proof of the intermediate value theorem, calling impure proofs “an intolerable offense against correct method”. Frege insisted on “deriving what was arithmetical by purely arithmetical means”. Even in the first half of the 20th century, mathematicians invested a great deal of effort in finding a pure proof of the prime number theorem, put off by the presence of complex analysis in the original proof. Detlefsen and Arana agree that purity has positive epistemic value, and they offer a detailed epistemological theory to explain why this is so.

By contrast, Lehet and Ryan are interested in the value of *impurity*. Both suggest that contemporary mathematicians tend to prize highly impure proofs over pure ones. If this is true, it represents an important shift in epistemic values.

As Lehet points out, the whole *raison d'être* of several branches of contemporary mathematics involves transferring problem-solving resources between different domains. Algebraic topology, algebraic geometry and analytic number theory are pillars of our current mathematical edifice, and all are thoroughly impure by nature. It's hard to imagine a present-day mathematician disparaging these subjects as wayward and unnatural, or demanding that the proofs all be redone in accordance with strict rules of topical hygiene.

In response to such developments, one might be tempted to conclude that the success of “impure” methods really just shows that, on closer inspection, seemingly different branches of mathematics may turn out to be the same after all. (If algebra and geometry are so indispensable for understanding one another, why not just view them as a single subject appearing under different guises?) But as Lehet points out, mathematicians themselves tend not to see things this way. Expert accounts describe tools like homology as facilitating jumps between distinct domains, rather than as collapsing illusory subject boundaries.

Embracing impurity allows us to prove more, and that's never a bad thing. But does impurity have distinctive epistemic value of its own? Lehet suggests it does. Impure methods *unify*, for one, and unification often leads to understanding and explanation. For instance, category theory allows us to see dissimilar-looking theorems from disparate parts of mathematics as examples of the same high-level phenomenon.

While Lehet argues forcefully for the epistemic value of impure proofs, she allows that purity may also be valuable in its own way. Ryan is less sanguine. In particular, he holds that impure proofs, but not pure proofs, are generally explanatory.

Ryan's main case study is Szemerédi's theorem, a result about the existence of arithmetical progressions in certain subsets of the natural numbers. Ryan contrasts the original, pure proof of Szemerédi's theorem—a famously intricate and forbidding combinatorial argument—with Furstenberg's later impure proof using ergodic theory. The diagnosis is that Furstenberg's proof explains while Szemerédi's doesn't. On Ryan's view, the explanatory power of Furstenberg's proof—and indeed of impure proofs in general—derives from the presence of shared structure between the domain of the proof and the domain of the theorem. (In this case, what both proof and theorem share is the possibility of decomposing the relevant objects into a structured part and a random part.) Echoing Lehet, Ryan suggests that shared structural content is explanatory because it allows for the unification of seemingly unrelated facts. Ryan also highlights the greater simplicity and perspicuity of the ergodic proof.

Ryan promises to refine some of this machinery in future work. I hope he does, because I'm not yet convinced! Szemerédi's theorem makes for a nice case study, and I more or less agree with Ryan's take on the two proofs at issue. What's less clear to me is how well the analysis will generalize. In particular, I wonder whether there's really a good notion of structural content according to which (1) impure proofs almost always reveal shared structural content, (2) pure proofs have no tendency to reveal shared structural content, and (3) revealing shared structural content is almost always explanatory.

I look forward to seeing how Ryan argues for these claims. In the meantime, I hope others maintain the momentum Lehet and Ryan have created — it's good to see purity back on philosophers' agendas again.

WILLIAM D'ALESSANDRO
MCMP, Munich

EVENTS

JANUARY

COURSES AND PROGRAMMES

Courses

CiE: Computability in Europe 2021: Connecting with Computability Tutorials, 5–9 July.

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.

MASTER PROGRAMME: in Statistics, University College Dublin.

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: by research, Tilburg University.

MA IN PHILOSOPHY, SCIENCE AND SOCIETY: TiLPS, Tilburg University.

MA IN PHILOSOPHY OF BIOLOGICAL AND COGNITIVE SCIENCES: Department of Philosophy, University of Bristol.

MA IN RHETORIC: School of Journalism, Media and Communication, University of Central Lancashire.

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.

JOBS AND STUDENTSHIPS

Studentships

DOCTORAL PROGRAMME IN PHILOSOPHY: Language, Mind and Practice, Department of Philosophy, University of Zurich, Switzerland.

LogiCS: Joint doctoral program on Logical Methods in Computer Science, TU Wien, TU Graz, and JKU Linz, Austria.

Jobs

JUNIOR PROFESSOR: Chair AIDAL: Artificial Intelligence, Data, Algorithms and Law, University of Toulouse 1 Capitole, deadline 30 March.

POSTDOC RESEARCHER IN AGENT-BASED MODELLING: Department of Computer Science, Parks Road, Oxford., deadline 23 March.

PROFESSOR: in Philosophy of Medicine, University of Bordeaux, France, deadline to be determined.

WHAT GREEK LETTERS MEAN IN EQUATIONS

π THIS MATH IS EITHER VERY SIMPLE OR IMPOSSIBLE.

Δ SOMETHING HAS CHANGED.

δ SOMETHING HAS CHANGED AND IT'S A MATHEMATICIAN'S FAULT.

θ CIRCLES!

ϕ ORBS

ϵ NOT IMPORTANT, DON'T WORRY ABOUT IT.

u, v IS THAT A V OR A U? OR...OH NO, IT'S ONE OF THOSE.

μ THIS MATH IS COOL BUT IT'S NOT ABOUT ANYTHING THAT YOU WILL EVER SEE OR TOUCH, SO WHATEVER.

Σ THANK YOU FOR PURCHASING ADDITION PRO®!

Π ...AND THE MULTIPLICATION® EXPANSION PACK!

ζ THIS MATH WILL ONLY LEAD TO MORE MATH.

β THERE ARE JUST TOO MANY COEFFICIENTS.

α OH BOY, NOW *THIS* IS MATH ABOUT SOMETHING REAL. THIS IS MATH THAT COULD KILL SOMEONE.

Ω OOOH, *SOME* MATHEMATICIAN THINKS THEIR FUNCTION IS COOL AND IMPORTANT.

ω A LOT OF WORK WENT INTO THESE EQUATIONS AND YOU ARE GOING TO DIE HERE AMONG THEM.

σ SOME POOR SOUL IS TRYING TO APPLY THIS MATH TO REAL LIFE AND IT'S NOT WORKING.

ξ EITHER THIS IS TERRIFYING MATHEMATICS OR THERE WAS A HAIR ON THE SCANNED PAGE.

γ ZOOM PEW PEW PEW [SPACE NOISES] ZOOOOM!

ρ UNFORTUNATELY, THE TEST VEHICLE SUFFERED AN UNEXPECTED WING SEPARATION EVENT.

Ξ GREETINGS! WE HOPE TO LEARN A GREAT DEAL BY EXCHANGING KNOWLEDGE WITH YOUR EARTH MATHEMATICIANS.

Ψ YOU HAVE ENTERED THE DOMAIN OF KING TRITON, RULER OF THE WAVES.