Opening the black box and finding it – dark …
Epistemic opacity in computer simulation and machine learning
Conference series: Science and Art of Simulation IV
28.11. – 30.11.2018, HLRS, Stuttgart, Germany

Abstracts (Keynotes and all other participants)

Abstracts Keynote Speakers

Epistemic opacity and reasoning
Anouk Barberousse

I will present an analysis of epistemic opacity as related with the usual requirements of human reasoning, namely, intellectual control and accountability. From this perspective, I shall examine various sources of epistemic opacity when mathematical reasoning is performed with the help of machine computation.

Opacity and its Discontents
Till Grüne-Yanoff

When is it legitimate to employ epistemically opaque models, if ever? Epistemically opaque models have been the focus of a whole battery of criticisms, ranging from failing to provide understanding to disallowing responsibility attribution. The various criticisms assert that opacity hampers or makes impossible reaching certain goals. However, epistemically opaque models also offer advantages: only epistemically opaque models exploit the computational potential of today's computers to the full, and model users often do not want to be burdened with the details of the computational process. In this paper, I offer an *instrumental analysis* of epistemic opacity – investigating which goals opacity advances, for which it is an obstacle, and which kind of opacity is preventing which goal. My analysis helps identify a number of opacity-mitigation strategies, ranging from the identification of opacity-unaffected goals through the reconceptualization of scientific goals and various transparency measures to counteracting opacity's negative effects without eliminating opacity. Altogether, I conclude that although epistemic opacity occasionally poses obstacles for computational modeling, it does not amount to an overall indictment. Rather, there are plenty of ways to legitimately employ epistemically opaque models.

The Surprisingly Intricate Link Between Models and Simulations –
and How It Creates Opacity
Hans Hasse and Johannes Lenhard

The computer enlarges the realm of mathematical (or formal) models that are tractable. But what is tractable for a machine might be opaque for human beings. On aspect of this opacity which has gained considerable attention by philosophers is the number of computational steps involved when exercising a simulation. We want to change focus and scrutinize the *process* that leads from a mathematical model to a simulation result. We argue that this process

- is more intricate than many philosophical analyses suggest,
- transforms the mathematical model in substantial ways, and
- creates opacity.

We support our analysis by evidence from a recent round robin study in simulation-based engineering.

Epistemic Opacity: Sources,Varieties, and Epistemological Consequences
Paul Humphreys

The talk will identify different sources for epistemic opacity in computational science (rather then in simulations alone), contrast epistemic opacity with related concepts such as black boxing and unsurveyability, and provide a partial taxonomy of different types of opacity. A focus will be the special difficulties produced by deep neural nets and machine learning. The effects on the epistemology of science will be another focus and the prospects for using either some version of a reliabilist epistemology or an entitlement epistemology will be discussed.

Multiple modelling in tackling epistemic opacity
Andrea Loettgers & Tarja Knuuttila

Control is a central function of biological organisms, yet it is difficult to study both theoretically and empirically. Biological systems are notoriously complex dynamical systems providing an excessive supply of possible ways of instantiating control. One strategy to tackle this problem is multiple modelling in different materialities. We are examining multiple modelling in the study of genetic circuits. The research practice in question explores biological control by making use of an ensemble of different epistemic means: mathematical models and simulations, synthetic genetic circuits and intracellular measuring devices, and finally electronic circuits.

Reproducibility in Computational Science

Thomas Ludwig

Recently, there is intensive discussion about reproducibility of results in science. Particularly with computational science, the influence of IT infrastructures is tremendous. We are confronted with complex computer programs, massive amounts of data, and powerful machines beyond imagination. How will boundary conditions defined by these components allow us to attain the goal of getting reproducible results with numerical simulations on current high performance computer systems? Will the machine represent a deterministic system? Do we have control over the feature of being deterministic? Can we reproduce simulation results at a later point of time? The talk will discuss the characteristics of IT-based research infrastructures and their influence on reproducibility of results.

Contingent Opacity and the Distribution of Epistemic Responsibility

Julian Newman[1,2]

In this talk I shall argue for seven main points:-

1. Our knowledge of computer simulation models, as of all software artefacts, is empirical and not a priori. Hence I reject the arguments for their *essential* epistemic opacity.
2. Epistemic opacity is widespread in computer simulation models, arising not from their essential nature, but contingently from the neglect of good software engineering practice in model design and development.
3. Such contingent epistemic opacity is a serious problem that can undermine the adequacy of computer simulation for many practical and epistemic purposes.
4. The rise of "Post-Normal Science" has blunted the internal critical practice of scientific communities while enhancing scepticism in citizen communities.
5. Developers of computer simulations who intend them to be adequate for policy-related purposes need to recognise that the audience for Post-Normal Science has particularly demanding requirements for surveyability (what outside the field of computing would be called "transparency").
6. Provision of adequate support for the exercise of epistemic judgement on the part of multiple audiences should be recognised as a central responsibility of developers and users of computer simulation models.
7. In this context it would be irresponsible, as well as methodologically flawed, to accept that nothing can be done to avoid epistemic opacity in computer models.

The Essential Epistemic Opacity thesis (EEO) introduced by Humphreys (2004, 2009) and the Confirmation Holism of complex simulation models identified by Lenhard & Winsberg (2010) (CH) are two sides of the same coin. Revisiting Lenhard & Winsberg (2010), Winsberg (2018, p 145) comments: "Analytic Impenetrability makes epistemically inscrutable the effects on the success and failure of a global model of past methodological assumptions that are generatively entrenched." Therefore my critique is directed equally at both EEO and CH, and I refer to my target as EEO-CH.

[1] Department of Philosophy, Birkbeck College, University of London; jnewma09@mail.bbk.ac.uk
[2] Orcid: orcid.org/0000-0002-8291-5778

EEO-CH has a number of adverse consequences. Either we accept a simulation model as a superior epistemic authority which is not accountable to human judgement, or we fall back on un-analysed expert intuitions about the adequacy of the model. The latter alternative is particularly unpalatable when we realise the extent to which model construction practice in fields such as Climate Science is dependent on expert opinion rather than being validated by reference to independent empirical data. The widespread acceptance of EEO-CH is influenced by a constellation of perspectives on practice, norms and validity:-

a) The 'Practice Turn' in Philosophy of Science influences philosophers towards assuming the validity of current practices; related conceptual trends include pluralism and pragmatism. As a corrective to this, I draw attention to Russell's characterization of pragmatist instrumentalism as 'cosmic impiety' which he contrasted unfavourably with the ancient Greek avoidance of hubris (Russell, 1946, p 737). Philosophers should, of course, pay attention to scientific practices, but not accept them uncritically.
b) The increasing pursuit of policy-relevant science has led to a re-evaluation of scientific norms in the name of "Post-Normal Science". The substitution of policy goals for traditional scientific norms exemplifies the hubris of which Russell warned; moreover there is a moral hazard associated with the monopsonistic nature of the market for expertise, which (to employ Elgin's adaptation of Kantian terms) renders the members of the scientific community heteronomous (cf Elgin, 2017). As a consequence, dissent is restricted to those who have nothing to lose in career terms; meanwhile the 'consensus' is maintained by the depiction of dissenters as uniquely venial.
c) Both in scientific and in philosophical communities, discourse about the relationships between models and empirical data is beset by severe ambiguities about Con-firmation, Corroboration, Reliability and Validity. (The confusions of terminology and practice are too great and too many to permit a full treatment here).

A simulation model is an instance of software, and software is an immaterial artefact. The EEO thesis is based on the assumption that understanding software requires us to be able to follow every step of a computation in real time. In a previous paper (Newman, 2016) I have pointed out that this is at variance with the actual processes and disciplines whereby software is produced and maintained, and have suggested that Empirical Software Engineering (ESE) provides the key to an alternative account of Software Epistemology. Good engineering of software is not cost-free: hence the temptation to kludges and the reluctance to reconsider and re-engineer legacy code (or as the software engineering community would express it, to 'refactor'). From this perspective, epistemic opacity appears a contingent consequence of inadequate attention to good Software Engineering practice.

I further suggested that this problem may be understood in terms of the concept of 'Technical Debt' which has assumed increasing importance within software engineering in recent years. Technical Debt originated as a metaphor, according to which taking various shortcuts during the early stages of software development incurs obligations which will have to be repaid by additional work sometime later in the software lifecycle. Kruchten et al (2012) summarise current views of Technical Debt in terms of two dimensions: visibility/invisibility and maintainability + evolvability. Visible elements include new functionalities that need to be added and known defects that need to be fixed, but in their view "what is really a debt" is the invisible result of past decisions that negatively affect the future value of the software artefact. Ways in which this invisible debt can burden the developers and stakeholders include architectural problems giving rise to hard-to-correct multi-component defects, associated shortcomings in documentation, and factors making existing program code difficult to understand and modify, such as code complexity and violations of coding

style. They find that visible negative features depend to an important degree on less visible architectural aspects.

Studies of the distribution and persistence of errors in long-lived software have shown that persistent and hard to eliminate errors occur particularly around architectural boundaries (Li et al., 2012). In the field of computer simulation this is likely to be a particular problem with coupled models, where generative entrenchment (Lenhard & Winsberg, 2010) can act as a barrier to a clear, clean architecture for a model built out of previous models. For example, the failure of many global climate models to respect conservation of energy is thought to result from previous ocean and atmospheric models having different grid scales and different coastal representations. Ad-hoc fitting together of pre-existing models that have not been designed to be components of a global model creates problems that have to be fixed at the stage of model tuning, and the model tuning itself introduces further opacity into the behaviour of the overall model (Frisch, 2015; Mauritsen et al, 2012). From such examples I infer it is highly probable that observed symptoms of epistemic opacity in computer simulation arise from the acceptance of unmanageable degrees of Technical Debt, and not from essential characteristics of Computational Science with respect to the human agent.

On the basis that certain software engineering practices involve information hiding, Durán & Formanek (2018) reject this analysis (despite "find[ing] Newman's concerns reasonable, for they are based on the assumption that knowing how a method works gives insight into its outcome"). In particular they cite Colburn & Shute's description of information hiding as obscuring "details that are essential in a lower-level processing context but inessential in a software design and programming context" (Colburn and Shute, 2007, 176). Durán & Formanek interpret this as identifying "unavoidable degrees of opacity in standard software engineering practice that come with an agent being unable to relate a given computer program with its physical instantiation on the computer machine". But this begs the question, because the fundamental issue between us is whether one needs to follow every step of a calculation as it unfolds at the level of machine operations in order to justify a claim that one understands the functioning program. If our knowledge of software artefacts is empirical, it does not depend upon tracing the ***whole*** calculation step by step.

We know from a long tradition in Cognitive Psychology that the ability of people to assess an argument is highly dependent upon presentational forms (see also Jebeile, 2018). Mäki (2009) makes the useful suggestion that an account of a model ought to include its audience design and a commentary. In the context of computer simulation, one might then interpret 'commentary' as including any or all devices that render the model epistemically accessible by supporting surveyability for a particular audience of a particular model-based argument. (Compare Beisbart, 2012, on simulations as arguments). Surveyability of the model as argument is of particular importance with respect to the so-called "extended peer communities" of Post-Normal Science.

The concept of 'Post-Normal Science' is used both to justify and to criticize departures from traditional scientific norms. The term was originated by Funtowitz & Ravetz (1993, 1997) to designate a mode of scientific enquiry oriented to solving policy problems rather than basic epistemic issues. They distinguish Post-Normal Science from Applied Science which is also driven by practical problems but where traditional scientific norms remain applicable. Fields where they regard Post-Normal Science as relevant are those where "facts are uncertain, values are in dispute, stakes are high and decisions are urgent". Whereas in Kuhnian Normal Science problems are defined by the paradigm accepted by the scientist's peer community, and in Kuhnian Revolutionary Science a new

paradigm emerges in response to the breakdown and contradictions of the existing paradigm, Post-Normal Science is characterized by an "extended peer community" which participates in evaluating the quality of the scientific contributions to the decision process and resolving issues through public debate.

If non-specialist citizens are to function as an extended peer community, they will have particularly demanding requirements for surveyability in the evidence and results presented to them by specialists. Peterson (2006) presents a case study of the problems of communicating with the Dutch public the uncertainty issues relating to a simulation study supporting environmental policy. In the aftermath of a scandal in which an environment agency statistician turned whistleblower, a working party, in which Peterson participated, drew up recommendations for communication about uncertainty in any future research by the environment agency. It is not clear to what extent this contributed to public understanding or agreement. More generally, it has been shown (Curry, 2011) that the presentation of simulation-based results and policy-relevant constructed measures in Climate Science is often deficient in its treatment of uncertainty – yet uncertainty is crucial to public assessment of arguments for policy alternatives. Combining the insights drawn from Mäki with the variable experience of Post-Normal Science to date, we argue that it is incumbent upon simulationists in fields such as Climate Science to ensure that the reasoning and process underlying their projections, together with a clear account of the types and degrees of uncertainty attaching to these findings, is intrinsic to the communication of their investigations and results. Jebeile (2018) makes a powerful case for the role of visual presentations in making simulation results understandable; however her paper assumes that the simulation program itself will necessarily remain epistemically opaque and that the visualisations will continue, as now is commonly the case, to be produced by post-processing the simulation results. I would argue, rather, that what is called for is a radical refactoring of complex simulation models to ensure they are surveyable at all stages of their production and use.

Why is this not happening? Despite the idealistic vision propounded by Funtowicz and Ravetz, Post-Normal Science has had counterintuitive negative results. These, I would argue, arise from a virtual monopsony in the market for policy-relevant knowledge, which creates what the insurance industry would call "moral hazard". By "monopsony" I refer to a situation in which there is effectively only one buyer for a service, which prevents the emergence of a healthy competitive market (Koppl, 2018, pp 214-215). In a web post early in 2017, a leading climate scientist issued a statement explaining her decision to resign her position as professor and head of the Department of Earth and Atmospheric Sciences at Georgia Tech – a department which she had devoted 15 years to turning around from "struggling" to ranking "in the top echelon of earth and atmospheric schools globally" (Webster, 2017). She describes her growing disenchantment with universities, the academic field of Climate Science and scientists thus:

> I no longer know what to say to students and postdocs regarding how to navigate the craziness in the field of Climate Science. ***Research and other professional activities are professionally rewarded only if they are channeled in certain directions approved by a politicized academic establishment*** — funding, ease of getting your papers published, getting hired in prestigious positions, appointments to prestigious committees and boards, professional recognition, etc.[…] How young scientists are to navigate all this is beyond me, and ***it often becomes a battle of scientific integrity versus career suicide.*** (Curry, 2017; emphasis added)

According to the narrative of Post-Normal Science, *marginal groups* who have potential insights to offer *are welcomed* into the extended peer community; in practice, however, *members of the core scientific community are marginalized* if they challenge the official consensus. Consequently their views find expression only in the blogosphere or via think tanks, so that they are readily dismissed as 'merchants of doubt'.

Parker has noted that model evaluation often begins informally in the early stages of model development, and that more formal activity comes at a later stage. In this respect there is a parallel with experimentation, which often moves from an exploratory stage to more formal hypothesis-testing. Despite initiatives such as the Software Sustainability Initiative (Goble, 2014) the culture of Computational Science appears particularly resistant to adopting rigorous Software Engineering practices, as shown for example in findings from Kelly's group and in Easterbrook's argument that formal independent validation of climate models is unnecessary because they are validated by the actual practice of the developers (Easterbrook, 2010b, 2010c). But if development practice produces code which is epistemically opaque, the burden of formal validation at the later stage will be unmanageable and the move from exploration to validation will never take place.

Does this matter? I have said that it would be irresponsible to accept that nothing can be done to avoid epistemic opacity in computer models. I take some consolation from the fact that, despite their adherence to Essential Epistemic Opacity thesis, the practices which Durán & Formanek (2018) identify as a basis for 'Computational Reliabilism' are close to what I identify as the virtues of an empirically-based software epistemology.

Validation is relative to goals and the context of intended use. The consequences of resistance to software engineering practices supporting surveyability may be more serious in (e.g.) Climate Modelling than in Astrophysics. Where goals of computer simulation are purely epistemic, the context may be set purely by and within a specialist group or community, since the issues are not of any wider interest. This may lead to de-facto tolerance of model opacity on a pragmatic basis, particularly where model results are seen as consilient with the group's general understanding. Where goals relate to public policy, the modeller cannot rely upon a unified perception of goals, nor upon an agreed general understanding of the field. Paradoxically, then, adequacy for purpose in a public policy context is likely to make higher demands upon strength and clarity of argument and upon precision and accuracy in model results than adequacy in a context of pure research. Responsibility for purely scientific *epistemic decisions* is distributed, but only within a specialist community. Major *policy decisions* such as climate policy are of such broad significance that responsibility must be shared with whole communities of citizens. The 'expert' claim that 'the science is settled' tends to weaken rather than strengthen the publicly-perceived rigour of the argument. This reinforces the importance of 'audience and commentary' in the philosophy of simulation science.

References

Beisbart, C (2012) How can computer simulations produce new knowledge? *Euro Jnl Phil Sci* **2**, 395-434.

Butos, W & McQuade, T (2015) Causes and consequences of the Climate Science boom. *The Independent Review,* **20(2),** 165-196.

Colburn, T & Shute, G (2007) Abstraction in computer science. *Minds and Machines,* **17(2),** 169-184.

Curry, J (2011) Reasoning about climate uncertainty. *Climatic Change*, **108,** 723-732.

Curry, J (2017) Posting on *Climate etc* about the author's reasons for resigning from her post as Professor and Head of Department at Georgia Tech; reposted by Anthony Watts at

https://wattsupwiththat.com/2017/01/04/dr-judith-curry-chooses-integrity-over-the-state-of-climate-science/ Downloaded 02/02/2018

Durán, JM & Formanek, N (2018, in press) Grounds for trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines.*

Easterbrook, S M (2010a) Climate change: a grand software challenge. In *Proceedings of the FSE/SDP workshop on Future of software engineering research* (pp 99-104). ACM

Easterbrook, SM (2010b) Do climate models. need Independent Verification and Validation? *Serendipity: Applying systems thinking to computing, climate and sustainability* 27 November 2010. http://www.easterbrook.ca/steve/2010/11/do-climate-models.-need-independent-verification-and-validation/

Easterbrook, SM (2010c) Validating climate models. *Serendipity: Applying systems thinking to computing, climate and sustainability* 30 November 2010. http:// www.easterbrook.ca/steve/2010/11/validating-climate-models/

Elgin, CZ (2017) *True Enough.* MIT Press.

Frisch, M (2015) Predictivism and old evidence: a critical look at climate model tuning. *Euro Jnl Phil Sci* **5,** 171-190.

Goble, C (2014) Better Software, Better Research. *IEEE Internet Computing,* Sept/Oct 2014, 4-8.

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

Humphreys, P (2009) The Philosophical Novelty of Computer Simulation Methods. *Synthèse* **169**: 615-626.

Humphreys, P (2011) Computational Science and its Effects. In Carrier, M & Nordmann, A (eds) *Science in the Context of Application.* Boston Studies in the Philosophy of Science, 274. Springer

Jebeile, J (2018) Explaining with Simulations: Why Visual Representations Matter. *Perspectives on Science* **26(2),** 213-238.

Kelly, D (2007) A Software Chasm: Software Engineering and Scientific Computing. *IEEE Software,* November/December 2007, pp 120, 118-119.

Kelly, D (2009) Determining factors that affect long-term evolution in scientific application software. *Journal of Systems and Software*, **82(5),** 851-861.

Kelly, D (2013) Industrial scientific software: a set of interviews on software development. In *Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research*, IBM Corp. 299-310.

Kelly, D. (2015). Scientific software development viewed as knowledge acquisition: Towards understanding the development of risk-averse scientific software. *Journal of Systems and Software*, **109,** 50-61.

Kelly, D, Hook, D & Sanders, R (2009) Five Recommended Practices for Computational Scientists Who Write Software. *Computing in Science and Engineering,* September/October 2009, 48-52.

Kelly, D, Smith, S & Meng, N (2011). Software engineering for scientists. *Computing in Science & Engineering*, **13(5),** 7-11.

Koppl, R (2018) *Expert Failure.* Cambridge University Press.

Kruchten, P., Nord, R.L., Ozkaya, I.: Technical debt: from metaphor to theory and practice. *IEEE Software,* **29(6),** 18-21 (2012).

Lenhard, J & Winsberg, E (2010) Holism, entrenchment and the future of climate model pluralism. *Studies in the History and Philosophy of Modern Physics,* **41**, 253-263.

Li, Z, Madhavji, NH, Murtaza, SS, Gittens, M, Miranskyy, AV, Godwin, D & Cialini, E (2011) Characteristics of Multiple-component Defects and Architectural Hotspots: A large system case study. *Empirical Software Engineering.* **16**: 667-702.

Mäki, U. (2009). MISSing the world. Models as isolations and credible surrogate systems. *Erkenntnis*, **70(1),** 29-43.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M.,Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H., Tomassini, L.(2012) Tuning the climate of a global model. *J. Adv. Model. Earth Syst.*, **4**, M00A01, doi:10.1029/2012MS000154.

Newman, J (2016) Epistemic Opacity, Confirmation Holism and Technical Debt: Computer Simulation in the Light of Empirical Software Engineering. In Gadducci, F & Tavosanis, M (eds) *History and Philosophy of Computing – Third International Conference, HaPoC 2015,* Pisa, Italy, October 8-11, 2015, Revised Selected Papers. Dordrecht: Springer. ISBN 978-3-319-47285-0. Pp 256-272.

Parker, W (2010) Scientific Models and Adequacy for Purpose. *The Modern Schoolman,* **LXXXVII,** *March and May 2010,* 281-293.

Petersen, AC (2006) Simulation Uncertainty and the Challenge of Postnormal Science. In Lenhard, J, Kűppers, G & Shinn, T (eds) *Simulation: Pragmatic Constructions of Reality.* Springer, 173-185.

Russell, B (1946) *History of Western Philosophy*. Routledge.

Webster, P (2017) A Personal Note. Posted on *Climate etc.* Reposted by Anthony Watts at https://wattsupwiththat.com/2017/01/04/dr-judith-curry-chooses-integrity-over-the-state-of-climate-science/ Downloaded 02/02/2018

Winsberg, E (2018) *Philosophy and Climate Science.* Cambridge University Press.

Epistemic Opacity: Outlines of a Research Program

Michael Resch, Andreas Kaminski

It is obvious that the sciences are characterized by an increasing use of computers. However, the extent to which this changes science is a controversial question. Paul Humphreys has put forward a thesis on this subject: The use of computer simulations goes hand in hand with the emergence of epistemic opacity. Humphrey's thesis is no less controversial than the question mentioned at the beginning.

The assumption that epistemic opacity is novel meets with objections that argue that science has always been opaque through the social division of labor and its technical organization. This objection is true, but it misses the point that there is indeed *one* novel form of opacity that occurs through a specific combination of mathematics and information technology in computer simulation and machine learning. To develop our argument, we seek clarification by distinguishing different forms of opacity associated with different sources. This shows that certain forms have always been part of science, but that there is a novel form of opacity that arises in computer simulation and machine learning.

For this purpose, our lecture will deal with the following questions:

1. **Nature:** Is it a property of something (e.g. model) to be opaque or is it a concept of reflection?

2. **Forms and sources:** Is there one opacity or are there different forms and sources?
3. **Novelty:** Is epistemic opacity a novelty or was science always partially opaque?
4. **Justification strategies:** How can computer simulations and machine learning systems be justified if they are opaque?

Abstracts of all other participants

01 - Agent-Independent opacity

Ramón Alvarado

The term epistemic opacity was used by Paul Humphreys (2004) to characterize the epistemic inaccessibility of the underlying processes and properties of some systems, particularly computer simulations. Epistemic opacity has important implications for debates concerning the reliability, trust, and novelty of computational methods, such as simulations in science. Computational methods are now ubiquitous in contemporary scientific inquiry. Alongside traditional non-software dependent scientific practices, software now plays an integral role in the process of empirical inquiry. This makes epistemic opacity an unavoidable topic in contemporary philosophy of science.

In recent work, Kaminski, Resch and Küster (2018), and Humphreys himself (Alvarado and Humphreys, 2017), offer a more detailed analysis of epistemic opacity. Kaminski et al., in particular, suggest that more attention ought to be given to the specific sources of opacity. They have argued for a distinction between social, technical and mathematical opacity. When it comes to software, they argue, it is through the lens of mathematical opacity that successful responses to questions about the reliability, trust and novelty of software systems, such as simulations, in science can be given (Kaminski, 2017). Similarly, in her paper "How the machine 'thinks': Understanding opacity in machine learning algorithms" Jenna Burrell (2016) draws a distinction between three kinds/sources of opacity: intentional and institutional secrecy, technical illiteracy, and scale/dynamic complexity. Humphreys and Alvarado (2017) for their part suggest that current statistical software methods such as machine learning algorithms elicit instances of representational opacity. As an example of partially opaque representations they use the discovery, by automated methods, of statistical structures within large data sets that are hidden to humans.

Here, I offer a taxonomy of epistemic opacity related to software which includes Humphreys' general and essential formulations of opacity (2009), representational opacity (Alvarado and Humphreys, 2017), as well as social, technical and mathematical accounts of epistemic opacity (Lenhard and Hasse, 2016; Burrell, 2016; Kaminski, 2017), and which together I will call conventional accounts of opacity in software. I argue that these accounts ultimately fail to capture the full range of epistemic opacity instances related to software systems. This is because conventional accounts are mostly agent-based accounts of epistemic opacity. That is, although it is true that epistemic opacity is a relational feature between an agent and a system/process and not merely a property of a system (Kaminski et al., 2018), conventional accounts of epistemic opacity like the ones mentioned above only capture those instances that are due to limitations ascribed to the agent. In this way, they

neglect the fact that opacity can also emerge from features that are not responsive to agential differences.

02 - Black Boxes or Rube Goldberg Machines? Neural Networks as Ashby Regulators

Cameron Beebe

Analyzing an ANN is like analyzing a brain or foreign system we presume to be intelligent. Thus, the epistemic situation we have with respect to an ANN falls primarily under the jurisdiction of cybernetics and cognitive systems theory. A cybernetic regulator is a complex system which controls environmental inputs by appropriate actions, resulting in a state aligning with a regulatory goal. Ashby (1958, §11) outlines the idea with a very simple decision game between two players.

		R	
E	α	β	γ
	β	α	γ
	γ	α	δ
	δ	ϵ	γ
	γ	δ	ϵ

S

Both players ('Environment', 'Regulator') have access to actions (rows and columns), and can see all possible outcomes. E goes first, and R plays for some outcome. If R plays for γ, any row that E chooses can be responded to with a move such that R wins every time. In fact, if the variation of outcomes in any given row were always and only 3 (i.e. only α,β,γ) then R could win no matter what was played for as long as it was on the board. As the game stands, there are goals (such as playing for δ) which can never be satisfied under certain choices from E. The variation of regulatory responses by R is, in this sense, insufficient to control the variation in E. Broadly, Ashby's Law of Requisite Variety says that "only variety in R can force down the variety due to [E]". A good regulator R has a sufficiently variable strategy profile for responding to moves by E, enabling it to accurately and effectively control the outcomes of such a game. The notion of a regulator can be generalized to an ANN, and can be seen as the same principle underlying the VC dimension and shattering a set of points. We could end the analysis here, if one is satisfied with the above insights. However, I anticipate a particular objection stemming from one of Ashby's theorems that good regulators must be models of the environment:

> In this regard, the theorem can be interpreted as saying that although not all optimal regulators are models of their regulands, the ones which are not are all unneccessarily complex. Conant and Ashby (1970)

When doing a machine learning task, we might scoff at the fact that there are millions (or more) trainable parameters. This may give us the impression of a black box, but these parameters are not hidden. Epistemic clarity about ANNs can still be enhanced further. If we consider groupings and sub-networks (and their specialized functions), the black box impression fades alongside the idea that parameters in the model are "unnecessarily complex". Convolutional networks, for example,

pass filter kernels over an image, detecting features such as edges. Capsule networks provide more structure which helps preserve spatial relationships between features. These examples continue to deflate the epistemic opacity worries of ANNs, providing structure to a sea of regulatory parameters. If we are still justified in treating these objects as opaque, it is not because they are like black boxes. Rather, they are opaque like biologically inspired Rube Goldberg machines.

References

Ashby WR (1958) An Introduction to Cybernetics. Chapman and Hall Conant RC, Ashby WR (1970) Every good regulator of a system must be a model of that system. International Journal of Systems Science 1(2):89–97.

03 - Two Levels of Opacity in High Energy Physics and Their Contingent Nature

Florian Boge / Paul Grünke

High Energy Physics (HEP) is a field of research in which the use of Computer Simulation (CS) abounds. Searches for particles, as well as the determination of their properties, depends on CSs from experimental design to evaluation of data.
In recent years, Machine Learning (ML) techniques have become more and more important in HEP, e.g. for the task of separating signal from background data (cf. Baldi et al. 2014 or the 2014 Kaggle Higgs Boson challenge). Even more recently, the HEP community has begun to explicitly acknowledge the “black box”-nature of ML algorithms and associated, induced features of opacity (e.g. Chang et al. 2018).
In this paper, we are going to argue that, although the involvement of ML in HEP introduces a new kind of opacity, the epistemic situation of the individual researcher at experiments like CERN’s ATLAS has not changed at all.
In particular, we first distinguish between (i) opacity induced by complexity, which is a feature of all traditional experimental procedures in modern HEP; and (ii) opacity induced by method, which is the specific opacity that accompanies searches using ML techniques.
We then argue that both kinds of opacity are contingent on the epistemic situation of the experimenting individual: It is in principle always possible to bypass the opacity induced by complexity by learning more about the functioning of, and interrelations between software packages at ATLAS or similar experiments; and it is in principle always possible to bypass opacity induced by the method by tracking the procedures that the computer undergoes during its training.
In a final step we argue that (a) it is unclear whether overcoming any of the two kinds of opacity is 'quantitatively more involved', meaning that both may require equally many working steps to be overcome, but that (b) overcoming opacity induced by method is 'qualitatively more involved', meaning that it may require the actual experimenter to develop new techniques for getting insight into the machine's workings. We use this observation to explain the recent interest in opacity in ML, as cited above.

References
Baldi, P. et al. (2014) “Searching for Exotic Particles in High-Energy Physics with Deep Learning”, Nature Commun., 5: 4308 (9pp).

Kaggle Inc. (2014) “Higgs Boson Machine Learning Challenge: Use the ATLAS experiment to identify the Higgs boson”, url: https://www.kaggle.com/c/higgs-boson (checked 05/18).
Spencer C., Timothy C., and Ostdiek B. (2018) “What is the machine learning?” Phys. Rev. D, 97(5): 056009 (6 pp).

04 - Switching on the lights inside the black box: how computer simulations observe the unobservable

Arianna Borrelli & Martin Warnke

The dream of watching the hidden is dreamt by humankind for a long time. At least I catched myself exclaiming when watching the wax models in the Medizinhistorisches Museum in Vienna or the Glass Flowers by Blaschka in Harvard: „In the end I now can see what it actually looks!“
Simulations, in wax, glass, or the computer, give these insights.
But the doubt remains: is this mere illusion or can these simulations tell some truth?
A recent event based computer simulation on the EPR experiment will demonstrate the claim that they actually can produce new knowledge: by illuminating the black box from the inside.

05 - An Agnostic Strategy of Solution and the Topology of Errors

Matthias Brandl, Johannes Lenhard

Computer methods that utilize large amounts of data recently have become able to tackle a class of problems that had been unsolvable before. Statistics based translation is one prominent example, pattern recognition by so-called deep learning presents another. We claim that these methods work with an agnostic strategy of solution. Given that agnostic strategies can be successful, they raise an important question about the topology of errors. There might occur errors that are neither explainable nor eliminable.

Artificial intelligence researchers have for a long time tackled the problem of machine translation. Knowledge-based accounts centered on grammar and parsing, but progress was much slower than expected because common language proved to entail far more surprises than anticipated. Statistics based machine translation, like it is implemented in the Google translator, brought a breakthrough in terms of functionality. It presents an agnostic solution strategy because it ignores the intricacies of grammar and its rules to a large part. Instead, relatively simple algorithms statistically evaluate small pieces of existing translations, like words and short word combinations. Notably, the solution calls for a very large base of given data.
Our thesis is that such agnostic strategies undermine a fundamental assumption about the topology of errors. The usual concept of numerical solution assumes that a numerical solution approximates the correct solution. The better the approximation, the smaller the error and the properties of the numerical solution approach the properties of the correct one. We call this the continuity assumption about errors – and we argue that it is misleading in the case of agnostic strategies.
It is misleading because it builds on the way how human beings improve by learning more vocabulary, or more grammar, i.e. improving their knowledge. But strategies like statistics based machine

translation or deep learning improve in a completely different way. A striking class of cases in support of my claim is the classification of patterns, like sorting pictures of birds according to the bird's species. Recently, deep learning approaches have become capable of correctly sorting such pictures to a high percentage. Our analysis focuses on the set of pictures that are misclassified. This set looks surprisingly strange to human evaluators because it contains patterns that do not appear similar but completely dissimilar to the correct solution – in contrast to the continuity assumption. For a human classifier, this type of errors appears not explainable. Furthermore, this set might contain a rather small fraction of all pictures (patterns) that resists elimination by further learning attempts of agnostic strategies.
We conclude by pointing out two open problems. Does the missing understanding of errors mark a technical limit for agnostic solution strategies? Second, to what extent is the misleading continuity assumption ethically relevant, e.g. when classification is implemented in automated cars?

06 - Three Forms of Transparency in Scientific Machine Learning

Kathleen Creel

Philosophers' existing analyses of transparency target complex climate models because their complexity and opacity contribute to public doubt about the predictions of the models. But computational opacity extends beyond climate models. Whether using machine learning at the LHC, artificial neural networks for cancer detection, or visual recognition software for fossil pollen identification, scientific researchers have capitalized on advances in machine learning algorithms without addressing the epistemic problems that can accompany such advances (Castelvecchi 2015; Tcheng et al. 2016; Esteva et al. 2017). They are now searching for ways to make machine learning more transparent in order to better detect errors and provide scientific explanations.

When computational opacity is ineliminable, philosophers such as Paul Humphreys recommend treating complex computational systems instrumentally (Humphreys 2004, 150). However, researchers are not and ought not to be satisfied with instrumental solutions, which fail both to deliver the epistemic goods which transparency can provide and to take into account computer scientists' recent successes in increasing transparency by novel means. It is both premature and unnecessary to "abandon the insistence on epistemic transparency for computational science" (ibid.). Instead, we need an analysis of transparency that captures more ways that systems can and should be transparent.

This paper argues for the need for transparency and claims that transparency comes in three forms. The first is functional transparency, or knowledge of the algorithmic functioning of the whole. Having functional transparency is being able to know which algorithm the program instantiates. The second is structural transparency, or localized knowledge of how mid-level components function. Being able to know sub-components that realize an algorithm and their relations is having structural transparency. The third is system transparency, or knowledge of the program as it was actually run in a particular instance, including the hardware and input data used. Having such transparency allows the detection of artifacts caused by interaction effects between the program and hardware, unexpected input data, and its implementation in a particular programming language. Identifying all three forms of transparency, as compared with recent papers that each identifies one, is necessary to explain recent successes in reducing opacity. These successes include LIME, which queries the decision space of an existing program from the outside and creates a sparse linear model to explain the original system's decisions, and visualization strategies used at Google DeepMind (Ribeiro, Singh, and Guestrin 2016). The tripartite analysis also gives those attempting to increase

transparency in future systems an analytical tool with which to identify the type of opacity to eliminate given their epistemic goals.

07 - Epistemic opacity in agent-based models of social phenomena

Thomas Durlacher

This paper explores the role of epistemic opacity in agent-based models designed to simulate social phenomena. Social scientists often praise agent-based models explicitly for their ability to provide a deep understanding of the simulated social phenomena. (Epstein 2008; Lemos 2018) In these accounts, understanding is generally related to the ability of agent-based models to explain social phenomena via the representation of causal mechanisms at the level of individual actors. In contrast, philosophers of science commonly emphasize that agent-based models suffer from epistemic opacity. (Humphreys 2004, 2009; Nersessian 2017) I am going to consider a recent proposal by Nicole Saam to distinguish social simulations resembling the epistemology and methodology of thought experiments and social simulations resembling the epistemology and methodology of experiments. (Saam 2017) According to her view only the former are not affected by epistemic opacity and are therefore preferable to the latter. I will argue that agent-based models do not easily fit into both categories. Missing from Saam's work are also details why some social simulations are opaque. To answer this question I am going to introduce several case studies of agent-based models. With the help of theses case studies, it is possible to bring the philosophical debate closer to the practices of social scientists and explore the reasons why opacity occurs. These agent-based models can be divided into two kinds, highly idealized toy-models and data-driven representative models. The two model-types provide different kinds of explanations, how-possible and how-actually explanations and present different levels of opacity depending on the complexity of the model. I will distinguish between the complexity of the model behaviors and the complicatedness of the model structure. Finally, I will present several strategies to cope with this kind of opaque complexity within agent-based models, the modularity of models and the question of the replicability of agent-based models.

Literature

Epstein, Joshua M. “Why Model?” Journal of Artificial Societies and Social Simulation 11, no. 4 (2008): 12. http://jasss.soc.surrey.ac.uk/11/4/12.html.

Humphreys, Paul. “The Philosophical Novelty of Computer Simulation Methods.” Synthese 169, no. 3 (2009): 615–26.

Humphreys, Paul. Extending Ourselves: Computational Science, Empiricism, and Scientific Method. New York: Oxford University Press, 2004.

Lemos, Carlos M. Agent-Based Modeling of Social Conflict From Mechanisms to Complex Behavior. Cham: Springer, 2018.

Magnani, Lorenzo, and Tommaso Bertolotti, eds. Springer Handbook of Model-Based Science. Dordrecht, Heidelberg, London, New York: Springer, 2017.

Nersessian, Nancy, and Miles MacLeod. “Models and Simulations.” In Springer Handbook of Model-Based Science, Dordrecht Heidelberg London New York: Springer, 2017. p. 119–132.

Resch, Michael, Andreas Kaminski, and Petra Gehring, eds. The Science and Art of Simulation I: Exploring - Understanding - Knowing. Cham: Springer, 2017.

Saam, Nicole J. “Understanding Social Science Simulations. Distinguishing Two Categories of Simulations.” In The Science and Art of Simulation I. Exploring - Understanding - Knowing. Cham: Springer, 2017. p. 67-84.

08 - Grounds for trust: Essential Epistemic Opacity and Computational Reliabilism

Nico Formanek & Juan Durán

Several philosophical issues in connection with computer simulations rely on the assumption that their results are trustworthy. Examples of these include the debate on the experimental role of computer simulations [Parker, 2009, Morrison, 2009], the nature of computer data [Barberousse and
Marion, 2013, Humphreys, 2013], and the explanatory power of computer simulations [Krohs, 2008, Durán, 2017]. The aim of this talk is to show that these authors are right in assuming that results of computer simulations are to be trusted. We claim that this trust is warrented exactly because computer simulations are reliable processes. After a short reconstruction of the so-called epistemic opacity, we present computational reliabilism, an amended form of traditional process reliabilism. While process reliabilism is externalist with regard to justifications we argue that this kind of radical externalism is not possible for computer simulations. Rather a subdued form of externalism is necessary which allows for at least one instance of the J→JJ principle. That is one can trust the results of a possibly epistemically opaque simulation if one has a method at hand which ensures its reliabilty. We discuss four such methods establishing reliability namely, verification and validation, robustness analysis for computer simulations, a history of (un)successful implementations, and the role of expert knowledge in simulations. We conclude by arguing that the general sceptical challenge concerning the universal reliability of such methods is theoretically unsolvable but poses no threat to practicing science with computer simulations.

09 - Epistemic opacity of computer simulations: a black-boxing feature

Julie Jebeile

Epistemic opacity of computer simulations: a black-boxing feature

Against Frigg and Reiss (2009), it has been argued that epistemic opacity is a novel feature of computer simulations (Humphreys 2009). "[M]ost steps in the [simulation] process are not open to direct inspection and verification" (Humphreys 2004, p. 148), mainly because a simulation run too fast for the simulationist to follow the computational processes in detail and, even if it was possible to slow down the simulation, the simulation would be still too long to be cognitively grasped by a human mind. Epistemic opacity therefore produces a black-boxing effect: it is hard for the simulationist to explain how the simulation outputs have been obtained from the model components.

In this paper, I argue that epistemic opacity is not the only black-boxing feature, and may not even be a novel feature.
Beforehand, I explicate the different reasons for epistemic opacity. They are heterogeneous in nature in that they include:
mathematical properties of numerical calculations such as incompressibility and long iterations;
limited cognitive abilities of the simulationist;
division of scientific labor, which sometimes goes with opaque epistemic dependency (Wagenknecht 2014), non-disclosure of data, and industrial property.

In the second part, I differentiate epistemic opacity from other black-boxing features, including complicated interactions between model components in computer simulations (Frisch 2015), and entrenchment that is due to the lacking track of the design choices in the model development (Lenhard and Winsberg 2010). It follows that all these features invite the simulationist to present the computer program as a black box, interacting with it in an experimental manner.

Importantly, I show that each feature entails a specific account for how the user should overcome in practice the gap between model components and simulation outputs. In particular, epistemic opacity entails two possible accounts: either (a) the user should go all over the series of logical and mathematical operations in the simulation in extenso, and survey every step of the computation; or (b) she should identify at least the epistemically relevant elements of the process.

In the third part, I question whether epistemic opacity is specific to computer simulations. As I have shown elsewhere with co-author, analytical solutions make sometimes numerical applications difficult or impossible. This is the case of infinite series. More particularly, to get a sufficiently good approximation from slowly convergent infinite series would require summing a considerable number of terms, which is not possible in practice. Even though the computational steps here are transparent to the user in principle, infinite series may still black-box the numerical application in practice.

References

Frigg, R., Reiss, J. 2009. The philosophy of simulation: hot new issues or same old stew? Synthese 169:593–613.
Frisch, M. 2015. Predictivism and Old Evidence: a Critical Look at Climate Model Tuning. European Journal for Philosophy of Science 5 (2):171–190.
Humphreys, P. 2004. Extending Ourselves. Computational Science, Empiricism, and Scientific Method. OUP.
Humphreys, P. 2009. The Philosophical Novelty of Computer Simulation Methods. Synthese, 169(3):615–626.
Lenhard, J., and Winsberg, E. 2010. “Holism, Entrenchment, and the Future of Climate Model Pluralism”. Studies in History and Philosophy of Science Part B, 41(3):253–262.
Wagenknecht, S. 2014. Opaque and translucent epistemic dependence in collaborative scientific practice. Episteme 11 (4):475-492

10 – Distinguishing between algorithmic opacity and epistemic opacity in the context of machine learning applications

Koray Karaca

Machine learning (ML) consists of an optimization procedure that is carried out through the following modelling tools: learning algorithm, hypothesis set, and training data. Due to the speed and complexity of the optimization process, it is beyond human cognitive capacity to keep track of all the steps of the optimization process. Therefore, the optimization process is a black box process in the sense that a human agent can only access the inputs and the corresponding outputs. This in turn

brings about algorithmic opacity, namely that the implementation of the learning algorithm is opaque, i.e., not transparent to direct inspection.

Since the training data cannot uniquely determine the modelling tools, in order to construct a unique ML model consistent with the available training data, ML modelers need to make judgments based on both epistemic values (e.g., simplicity, accuracy, robustness) and societal values (e.g., ethical, political, financial values) values in order to justify their specific choices of the modelling tools. Similarly, they need to make both epistemic and societal value judgments in justifying their specific choices of methods of validation and the data against which ML models are validated. Therefore, as I shall argue, unlike the optimization process, the aforementioned modelling aspects are transparent, in that they can be analyzed and interpreted by human agents in terms of their underlying value judgments.

An important implication of the above discussion concerns the notion of epistemic opacity. One prominent definition of this notion has been offered by Paul Humphreys as follows: "[a] process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process" (Humphreys 2009, p. 618). Humphreys also suggests that:

distinguishing between the weaker and stronger senses [of epistemic opacity] is useful. It is obviously possible to construct definitions of `partially epistemically opaque' and 'fully epistemically opaque' which the reader can do himself or herself if so inclined. (Ibid.)

In the case of ML, as I shall point out, what Humphreys calls "epistemically relevant elements" concern the aforementioned modelling aspects of ML. I shall argue that despite algorithmic opacity, the modelling aspects of ML are amenable to epistemic analysis and thus justification to the degree that their underlying value judgments are justifiable. This in turn means that algorithmic opacity in ML does not amount to full epistemic opacity, but rather only partial epistemic opacity, which is caused by the fact that the process of optimization cannot be accessed by human agents. I shall thus conclude that in the context of ML, epistemic opacity comes in degrees, thus showing the usefulness of the concept of partial epistemic opacity as suggested by Humphreys.

11- Shedding Light on Climate Change with Black Boxes

Benedikt Knüsel

Understanding is an important epistemic aim of science (de Regt 2009). In climate science, process-based computer models are one of the essential tools to advance understanding. Using climate models for this purpose rests on two assumptions, namely that (a) the relationships are adequately represented in the model, and that (b) no important causal factor is missing from the model (Parker 2014). These assumptions are made based on the coherence of the models with background knowledge and specifically their rooting in scientific theory, which is also one of the key reasons for confidence in climate model projections (Baumberger et al. 2017). Hence, although climate models suffer from epistemic opacity and confirmation holism (Lenhard and Winsberg 2010), they can be useful for increasing the understanding of the climate system.
In recent years, increasing volumes of data have opened up pathways for new, data-driven methods (Pietsch 2016). Such data-driven models, e.g., machine learning, can produce accurate predictions

of complex phenomena (Mayer-Schönberger and Cukier 2013; Pietsch 2015). It has been argued that causality is the reason for data-driven models' predictive success (Pietsch 2016), which indicates that they might be useful for advancing understanding, too. Since data-driven models can be trained when understanding of the target system is insufficient for constructing process-based models, machine learning could be an interesting tool to advance understanding of ill-understood phenomena.

In my paper, I discuss the example of attribution of climate change in temperature data. I compare a process-based climate model (see Huber and Knutti 2011) with a set of data-driven models (see Pasini et al. 2017) to assess how the two types of models can be used to increase understanding. Based on these results, I argue that machine learning can be used to better understand a given target system based on assumptions similar to (a) and (b). However, in the case of data-driven models, the two assumptions are conflated because assumption (a) generally holds only if assumption (b) holds, too. Furthermore, the lack of transparency of many machine learning algorithms makes plausibility checks difficult that could help to affirmatively argue for assumptions (a) and (b). Hence, while both types of models are epistemically opaque, this poses more serious problems for data-driven models in terms of understanding. I discuss why the two model types do not exhibit the same kind of epistemic opacity, and how these differences lead to different kinds of understanding that they can advance. Finally, I make suggestions for a strategy that makes data-driven models more useful for increasing scientific understanding despite the difficulties mentioned. The strategy is based on hierarchies of data-driven models with respect to their opacity, whose outputs need to be interpreted in light of the relevant background knowledge.

References

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (2017). Building confidence in climate model projections: an analysis of inferences from fit. Wiley Interdisciplinary Reviews: Climate Change, 8(3), e454. doi:10.1002/wcc.454

de Regt, H. W. (2009). The Epistemic Value of Understanding. Philosophy of Science, 76(5), 585–597. doi:10.1086/605795

Huber, M., & Knutti, R. (2011). Anthropogenic and natural warming inferred from changes in Earth's energy balance. Nature Geoscience, 5(1), 31–36. doi:10.1038/ngeo1327

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 41(3), 253–262. doi:10.1016/j.shpsb.2010.07.001

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution that Will Transform how We Live, Work and Think. John Murray.

Parker, W. S. (2014). Simulation and Understanding in the Study of Weather and Climate. Perspectives on Science, 22(3). doi:doi:10.1162/POSC_a_00137

Pasini, A., Racca, P., Amendola, S., Cartocci, G., & Cassardo, C. (2017). Attribution of recent temperature behaviour reassessed by a neural-network method. Scientific Reports, 7(1). doi:10.1038/s41598-017-18011-8

Pietsch, W. (2015). Aspects of Theory-Ladenness in Data-Intensive Science. Philosophy of Science, 82(5), 905–916. doi:10.1086/683328

Pietsch, W. (2016). The Causal Nature of Modeling with Big Data. Philosophy & Technology, 29(2), 137–171. doi:10.1007/s13347-015-0202-2

12 - Epistemic risks in AI-based science

Inkeri Koskinen

Recently the increasing use of AI-applications in science has raised many worries related to the possible epistemic risks arising with this development. For instance, risks related to biased data sets have become a much discussed topic. However, in this paper we argue that there are other important of epistemic risks related to the use of software in science.

In this paper we suggest a preliminary taxonomy that will help us distinguish between different types of epistemic risks in AI-based science. Such a taxonomy is needed, as the different types of risks require very different risk mitigation strategies.

An epistemic risk is "any risk of epistemic error that arises anywhere during knowledge practices" (Biddle and Kukla 2017, 218). Justin B. Biddle and Rebecca Kukla distinguish several different types of epistemic risks, starting from "analytic risk", the risk of reasoning errors, and ending with complex "phronetic" risks – for instance, unavoidable risks related to operationalization or model choice. Researchers cannot proceed in their work without continually taking such risks.

Scientists have of course developed effective strategies for mitigating such risks. However, as AI systems are used increasingly in science, the issue of epistemic risks is complicated further. Reliance on AI-systems may bring about new, unforeseen types of epistemic risks.

Many of the epistemic risks familiar to us arise from our own imperfections as epistemic agents. For instance, as human beings we are particularly prone to certain types of reasoning errors and biases. Also in AI-based science many of the epistemic risks we encounter arise from our own imperfections as epistemic agents.

However, as AI systems are used increasingly in science, and as the ways in which AI works are not transparent to humans, an AI system may gain the role of an epistemic agent in scientific research. This broadens the range of epistemic risks related to the use of software in science. Furthermore, they may bring about new, unforeseen ways of being imperfect as an epistemic agent – and, as a result, entirely new types of epistemic risks. We currently have have no effective strategies for mitigating such risks, or even recognising them. To start developing such strategies we must pay attention to the different potential sources of epistemic risks.

In this paper we develop a preliminary taxonomy for different types epistemic risks in AI-based science. We start by distinguishing four risk types:
i) Risks related to the biased data
ii) Risks related to the biased algorithms
iii) Risks related to the non-transparent algorithms
iv) Risks related to biased AI-architectures.
We then analyse these risk types, taking into account the source of the risk. Finally we identify in each risk type distinct characteristics that must be taken into account when developing risk mitigation strategies.

Biddle, J. B. and Kukla, R. 2017. The Geography of Epistemic Risk. In K. C. Elliott and T. Richards (eds.). Exploring Inductive Risk: Case Studies of Values in Science. New York: Oxford University Press, 215–237.

13- Learning to doubt. Epistemic opacity in computer simulations and the analogy from expert testimony

Jon Leefmann, Steffen Lesle

Epistemic opacity represents a problem to epistemologists since computer simulations have advanced to new complexities: The impossibility to control all epistemically relevant elements of advanced computer simulations leaves human users unjustified to belief in the simulations' output.1 Given the importance of computer simulations to contribute to the division of epistemic labour users must inevitably regard epistemically opaque systems as doubtful sources for the formation of new beliefs. This problem is, however, well known from the context of expert testimony in which a lively philosophical debate has considered several solutions to the question of how non-experts can acquire justified belief from expert's assertions.

In our paper we compare a simulation's epistemic opacity to cases of "pragmatic epistemic opacity", which define enquiring situations related to expert testimony. In enquiring situations non-experts ask experts for information despite poor individual capabilities to comprehend the epistemic relevant elements motivating the experts' belief. Nevertheless, non-experts can be quite successful in acquiring knowledge from expert testimony. Checking for indicators of expertise, deferring one's judgement to the assessments of meta-experts or building a trusting relationship with the expert are well established epistemic strategies.2 In the case of epistemic opacity in computer simulations these strategies are, however, rendered useless: Checking the output of a complex computer simulation in a satisfying way is as impossible as having a trusting relationship with a machine; and plainly relying on the programme's output simply ignores the question of how to deal with epistemic opacity.3

In contrast to these strategies we propose a better solution to the problem of acquiring justified beliefs epistemically opaque computer programs. Drawing on the analogy with expert testimony, we argue that it is a promising approach for non-experts to doubt the experts' testimony and to acquire reasons by learning from the experts' refutation of the doubts. Two reasons speak in favour of transferring this strategy to the context of computer simulations. First, defying doubts takes far less resources than explaining the epistemic processes that lead to a belief. As doubts always represent concrete alternatives to a testified belief only these alternatives need to be ruled out to acquire a reason to believe. This can easily be accomplished by complex computer programs. Second, the shown incompatibility of a doubt with concrete alternatives can be reproduced in order to defy similar doubts of users with similar beliefs. Users could then refer to these reasons and justify the formerly doubted belief.

We argue that this novel approach allows to justify belief in the output of epistemically opaque computer simulations. However, this strategy requires changing the way one approaches information from complex computer programmes: We need to start doubting and to start teaching the programmes how to defy our doubts.

1 Humphreys, P. (2008): ZiF Mitteilungen. Technical report, Zentrum für interdisziplinäre Forschung.

2 Goldman, A. I. (2001): Experts. Which ones should you trust? In: Philosophy and Phenomenological Research 63 (1), S. 85–110.
3 Knight, W. (2017): The Dark Secret at the Heart of AI. In: MIT Technology Review 120 (3).

14 - Why Justification for Machine Learning is not Undermined by Epistemic Opacity

Dilectiss Liu

We argue that the worry about epistemic opacity in machine learning is misplaced. Our argument consists of four steps.
The first step shows that the criteria for being epistemically justified is dependent on the goals of the epistemic domain. In mathematics, the current standard of justification is that of proofs, or more realistically — natural language texts that are supposedly translatable into gap-free formal deductions. In the natural sciences, a popular view is that valid inferences are grounded on the probability calculus. In particular, Bayesian confirmation theory provides such an explication. In everyday life, we use testimony, perception, memory as means to justify our epistemic practices such as forming beliefs, making assertions etc.
The second step argues that given a domain, the method that is justified is the method that would most likely satisfy the goals of that domain. Justification comes in degrees, testified by the continual refinement of our methods of enquiry. In mathematics, we have developed more sophisticated methods for proofs over the centuries, which has in turn raised the standards of mathematical justification. This is likewise the case in the sciences and in everyday life. Justification is therefore not a static, but a dynamic property. What counts as being justified in a domain is dependent on how well a method performs in that domain with respect to its competitors.
The third part argues that the primary purpose of machine learning is to give reliable outputs, given a reliability measure. In so far as most automated systems share this goal, machine learning and, say, logic-based A.I. are to be considered as being in the same epistemic domain. The measure of justification should thus be construed in an externalist fashion — that of reliability. What justifies a speaker's use or knowledge of her native tongue is simply that she can reliably employ the language correctly. Whether there is any explicit reasons she can provide for doing so is hardly relevant.
As a side remark, we deal with cases where the epistemic goal for automated systems is not to solve problems, but to represent certain processes. Nonetheless, we maintain that such goals are peripheral in the actual applications of automated systems.
The final part argues that the kind of epistemic opacity in machine learning affects not the efficacy with which machine learning achieves its goals in comparison to other automated systems. Machine learning is our most effective method — on multiple salient measures — for the problems that it tries to tackle; several applied examples verify this. Therefore, epistemic opacity in machine learning does not undermine the fact that the method of machine learning is epistemically justified.
We end with a clarification of what the issue of epistemic opacity really amounts to. It turns out that the kind of epistemic opacity in machine learning is not as novel nor as problematic as popular conception has portrayed it to be.

15- Ethical Implications of Opaque Simulations

Christoph Merdes

As John Stuart Mill's defense of freedom of expression on the basis of human fallibility (Barreteau et al., 2003) ill prominently showcases, epistemological concerns, and limitations to human epistemic capacities tend to have practical consequences. Those practical consequences encompass not only potential nonmoral normative requirements on the working scientist and engineer, but they also raise genuinely ethical questions. The opacity of computational science constitutes an instance of this general claim. To establish that, I shall discuss two instances of such ethical implications. By the opacity of computer simulations, following Humphreys (2009), I understand the impossibility for a human agent to follow the execution of a computational model in principle, be that due to the sheer number of instructions, the structural complexity of the model or the diversity of the knowledge embodied in the model. It is not necessary to decide here whether epistemic opacity is a specific feature of computer simulations, but computational models provide the most compelling examples.

For the first example, assume a basic contractualist framework[3], where ethical judgment depends on the agreement of reasonable agents. This conception of justification can be recognized in the companion modeling approach in simulation validation (Barreteau et al. , 2003): There, the acceptability of a simulation model and its results is built on the acceptance by stakeholders. It is easy to find examples of direct ethical relevance, as when city planning is supported by simulations the citizenry is allowed to contribute. Model opacity than raises the question whether a reasonable agent would agree to a procedure, in which a substantial portion of inferences are black boxed, not only contingently, but in principle. All use of experts entails a degree of black boxing; the difference is whether a black box can in principle be looked into and understood, as is usually possible for arguments backing expert judgment, or if instead part of the procedure is inherently epistemically opaque. A reasonable agent, however, is not compelled to accept the inferences provided by a black box, and therefore, no contractualist justification of the outcome of such stakeholder-based procedures can be constructed; if the dispute on a project of, e.g. economic development, depends on an opaque epistemic tool, people can reasonably disagree and therefore block the ethical justification of the project.

Second, consider the opacity of deep neural networks, usually referring to networks with multiple inner layers of artificial neurons. Very little is known analytically about the properties of these networks, which introduces a problematic new dimension into ethical decision-making. For the sake of simplicity, imagine the utilization of such an opaquely operating network in classifying tissue sample images for the purpose of medical decision-making. Unlike in classical statistics and many other numerical algorithms, there are no known error bounds, at least partly as a consequence of the opacity of the method. From an ethical viewpoint, this turns risk, that is, a problem with known probabilities, into uncertainty. Utilitarian decision theory provides very different methods to treat problems of uncertainty2, and in general, those are much weaker. Hence, the opacity of the computational method in question limits the strength of our reasons to choose from various options, constituting a practical, ethical disadvantage of the method.

These examples expound the ethical relevance of a particular epistemic feature of computational methods; as mentioned in the beginning, such connections have been discussed for centuries, but the peculiar nature of computer simulations adds a new and possibly very difficult class of examples to the conversation. The problems raised here also call into question a suggestion by Humphreys

[3] See for example Rawls (1958).

(2009) calling for an epistemic ideal different from transparency, since simply accepting fundamental opacity has consequences far beyond the epistemic realm.

References

Barreteau, Olivier, Antona, Martine, D'Aquino, Patrick, Aubert, Sigrid, Boissau, Stanislas, Bousquet, Franc¸ois, Dar´e, William's, Etienne, Michel, Le Page, Christophe, Mathevet, Rapha¨el, et al. . 2003. Our companion modelling approach. Journal of Artificial Societies and Social Simulation, 6(1).
Humphreys, Paul. 2009. The philosophical novelty of computer simulation methods. Synthese, 169(3), 615–626.
Mill, John Stuart. 1966. On liberty. Pages 1–147 of: A Selection of his Works. Springer.
Rawls, John. 1958. Justice as fairness. The philosophical review, 67(2), 164–194.
Savage, Leonhard J. 1972. The Foundations of Statistics. Dover Publications.

16 - Epistemic opacity and understanding: Herbert Simon on complexity and hierarchies

Henri Salha

In this communication we explore the concept of "epistemic opacity" at the light of scientific understanding. According to [Humphreys 2004] epistemic opacity "can result in a loss of understanding because in most traditional static models our understanding is based upon the ability to decompose the process between model inputs and outputs into modular steps, each of which is methodologically acceptable both individually and in combination with the others." We feel we understand a model or theory when we are confident that we can follow or replicate the steps of the process or reasoning.
This assertion seems however challenged by a recent account on the nature of scientific understanding proposed by [de Regt 2017, p102]: "A scientific theory T is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations". Here understanding seems on the contrary attained when the epistemic agent can predict results of the model or theory without her needing to perform any detailed computation.

A first explanation of the discrepancy could be that we are indeed confronted to two different traditions on the nature of understanding, a "formalist" one (Descartes, Hilbert) vs. an "intuitionistic" one (Maxwell, Schrödinger). We will aim to show that such an opposition is unnecessary and that the two views can be reconciled.

For this purpose we will convoke the thinking of Herbert Simon on complexity and on hierarchies. Simon believed indeed that complex situations could only be grasped by the epistemic agent if she was able to find the right "decomposition" of the problem in smaller sub-problems. Such hierarchical decomposition takes often the form of patterns recognition, which seems immediate – like a grandmaster grasps a chess situation by recognizing 5-6 configurations on the chessboard – but is actually grounded on a deep internalization of a lexicon of these patterns.

This idea, which still finds currency in recent cognitive science, is in line with the epistemic opacity concept, while making room for the "intuitionistic" perspective of de Regt. In Humphreys' quote

above, understanding is equated with an "ability to decompose … in modular steps", which echoes Simon's hierarchy. It also helps to better circumscribe the true domain of epistemic opacity: very complex chains of reasonings or calculations, but for which a hierarchy of verification procedures is available, should not pose any problem of opacity. Each node in the verification tree can be independently verified or warranted by third-parties [Barberousse & Vorms, 2014], therefore building-up the transparency of the whole. The core crux of epistemic opacity lies in processes or reasonings for which any hierarchization strategy is intrinsically unavailable – like Marr's type 2 theories. Interestingly, intuition is also failing us faced to these "true complexity" situations.

We will conclude on Simon's difficult anticipation of epistemic opacity . A rationalist mind, it took him a long time to believe that such opaque situations could be possible. He writes in 1969: "If there are important systems in the world that are complex without being hierarchic, they may to a considerable extent escape our observation and understanding". For the early Simon, epistemic opacity would be so deep that we would not even notice it, like black matter.

17 - Degrees of Epistemic Opacity

Iñaki San Pedro

The paper distinguishes two senses of "epistemic opacity" in computer simulations, namely a qualitative sense and a quantitative sense, and explores their relation to actual simulating and modelling practices.

From a qualitative point of view, the notion of "epistemic opacity" in computer simulation seems to have the same significance and implications for any computer simulations. That is, from a qualitative point of view, computer simulations seem to be equally opaque —i.e. we open the black box, and find it (always) dark! In this sense "epistemic opacity" expresses the fact that when a computer simulation is performed there is an "epistemic leap" associated to it. This kind of epistemic leap is characteristic rather than of a specific model or simulation, of the fact that a simulation is performed.

On the other hand, "epistemic opacity" can also be approached from a quantitative point of view. The questions to be asked then are rather different, e.g. is the "epistemic leap" noted above always of the same size? or are all computer simulations equally opaque, i.e. when we open the back box and find it dark, is it always as dark? The paper argues that (from this quantitative point of view), computer simulations display degrees of "epistemic opacity" (with the limit of non-opacity set in analycity). I will not discuss here whether these degrees of "epistemic opacity" can be measured (i.e. exactly quantified), or attempt provide a method for doing that. I will claim nevertheless that actual degrees of "epistemic opacity" are tightly related to what we can call the "complexity of the computational process", which is associated for instance to the particular design of the computing software at work, specific computer settings, or to hardware limitations. With this idea of complexity in mind, I will claim, the more complex a computational process is, the more (quantitatively) epistemically opaque will the simulation result.

I will note finally that a good deal of methodological decisions taken by scientist and modellers when performing computer simulations —i.e. typical tricks-of-the-trade such as parametrisation, use of expert knowledge, scaling, etc.—, which constitute an important part of current scientific practices

in the field, are precisely aimed at reducing such complexity. I will conclude thus that actual scientific practices (or part of these, at least) in fact reduce (quantitative) "epistemic opacity." This opens new and interesting questions such as whether actual scientific practices can manage to reduce "epistemic opacity" to the limit of analycity (thus eliminating "epistemic opacity" also in a qualitative sense), whether specific scientific practices can be said to reduce in some (qualitative) sense some of the uncertainties that computer simulations involve, or whether they have an impact on the reliability or confidence of specific computer simulations (possibly of the very same system).

18 - Neural network as an architectural principle

Gašper Štukelj

The opacity of an artificial neural network (aNN) often designates the absence of any law-like relationship between its interior organization and its external behavior. This seems to qualify aNN as an example of an "opaque machine". However, such conclusions hinge on viewing aNN solely as an (epistemologically opaque) algorithm, or "mathematical technique" [1]. To the contrary, a biological neural network, to which an aNN relates at least superficially, represents a physical layer of computation and not an intentional computational procedure. It was argued in [2] that representations and algorithms describing the behavior of biological networks are not their target properties, but rather mere by-products of their organizational principles (recurrency, mutual inhibition, etc.). These principles have evolved under evolutionary pressure to produce efficient, robust, and widely applicable control mechanisms. Humphrey's terminological proposition to use "non-linear prediction and classification techniques" instead of "neural nets" [1] suggests that an aNN should be contrasted with a "linear prediction and classification technique". However, mechanistic understanding of the biological networks puts them on a par with something like a von-Neumann architecture. Indeed, this is one of the core postulates of neuromorphic engineering (NE). The design of neuromorphic electronic systems draws upon our understanding of organization in neurobiological systems with an explicit goal of providing an alternative to the von-Neumann computer architecture [3]. Due to the development of NE, research in biological and artificial networks, which were before connected only superficially, have gained a deeper link. Training aNN on dedicated neuromorphic devices (ND) is more likely to involve specialized computation that occurs due to the physics of the hardware, rather than as a simulation with digital circuits. Trading off the flexibility of a digital computer, ND come with many perks; e.g., they provide greatly increased energy efficiency, and are better suited for parallel computing [3]. I argue that adequate philosophizing about aNN will have to take into account the difference between aNN as an architectural principle, and aNN as an algorithm, which can furthermore be implemented on a neuromorphic, or on digital hardware. It's not clear whether opacity is a concern when aNN is applied as an architectural principle to achieve computation with certain properties (e.g., robustness, low energy consumption). Similarly, to understand the computation, which is the source of epistemic opacity in the digitally simulated aNN, it's sufficient to provide a mechanistic explanation in the case when aNN is implemented on a ND. The epistemic status of the latter is similar to the epistemic status of a simulation of a quantum system on a quantum computer; with the borders between experiments and simulations blurred, the idea of epistemic opacity might prove to be over-pessimistic.

References

[1] P. Humphreys, “Data analysis: Models or techniques?,” Foundations of Science, vol. 18, no. 3, pp. 579–581, 2013.
[2] L. T. Hunt and B. Y. Hayden, “A distributed, hierarchical and recurrent framework for reward-based choice,” Nature Reviews Neuroscience, vol. 18, no. 3, pp. 172–182, 2017.
[3] C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, and J. S. Plank, “A survey of neuromorphic computing and neural networks in hardware,” CoRR, vol. abs/1705.06963, 2017.

19 - Strategies for Handling Epistemically Opaque Big Data

Marius Wälchli

Non-standardized big data of unknown quality exhibits a stronger epistemic opacity than standardized data normally used in science. However, due to the increasing ubiquity of data from cheap sensors, questions arise concerning its usefulness for scientific inquiry. Measurements and consequently datasets can be viewed as model-based representations (Tal 2013, 2017) and it is not possible to know all epistemically relevant elements of complex modelled datasets (Humphreys 2009). In climate science and meteorology, datasets are generated through complex networks of calibration and subsequent processing (Lloyd 2018). To assess whether such datasets are an adequate representation of the state of the climate system, the generation needs to be transparent in terms of technologies, calibration procedures, and theories used. Hence, strategies to resolve epistemic opacity of climate data include peer-reviewed publications backed up by well-established theoretical knowledge (see Morice et al. 2012; Kennedy et al. 2011a, 2011b). For non-standardized big data, transparency in data generation is not always possible, and hence, pragmatic accounts to deal with its epistemic opacity need to be developed.
In this paper, I use a case study to introduce a strategy for using non-standardized datasets despite their larger epistemic opacity. The case study is based on hourly temperature data of 3’589 non-standardized private weather stations (PWS) in seven cities on four continents for an entire year. An artificial neural network (ANN) is used as calibration tool, which accounts for biases and non-systematic errors. The ANN is trained to predict temperature values of standardized stations of the world meteorological organization (WMO) based on the PWS data. Other WMO-conform stations within the same and across different cities are used for out-of-sample evaluation. An accurate predictive model serves as proof of concept that the information in the PWS is sufficient to accurately predict WMO station measurement values. The calibrated ANN can then be used to predict temperature values across a city. However, this requires scientific background knowledge, e.g., about the urban heat island effect, in order to understand the limitations of such a data-based interpolation approach.

I show that machine learning techniques such as ANN can in principle be used as a pragmatic strategy to deal with epistemically opaque non-standardized big data in order to make it useful for scientific inquiry. This works if (i) standardized reference data is available for calibration, (ii) new observations allow the constant re-calibration of the machine learning algorithm, and (iii) the algorithm has a good predictive success. However, the factors limiting the confidence in these interpolations need be derived from scientific background knowledge.

References:
Humphreys, Paul. 2009. “The Philosophical Novelty of Computer Simulation Methods.” Synthese 169 (3): 615–26. https://doi.org/10.1007/s11229-008-9435-2.
Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby. 2011a. “Reassessing Biases and Other Uncertainties in Sea Surface Temperature Observations Measured in Situ since 1850: 1. Measurement and Sampling Uncertainties.” Journal of Geophysical Research 116 (D14). https://doi.org/10.1029/2010JD015218.
———. 2011b. “Reassessing Biases and Other Uncertainties in Sea Surface Temperature Observations Measured in Situ since 1850: 2. Biases and Homogenization.” Journal of Geophysical Research 116 (D14). https://doi.org/10.1029/2010JD015220.
Lloyd, Elisabeth A. 2018. “The Role of ‘Complex’ Empiricism in the Debates About Satellite Data and Climate Models.” In Climate Modelling, 137–73. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-65058-6_6.
Morice, Colin P., John J. Kennedy, Nick A. Rayner, and Phil D. Jones. 2012. “Quantifying Uncertainties in Global and Regional Temperature Change Using an Ensemble of Observational Estimates: The HadCRUT4 Data Set.” Journal of Geophysical Research: Atmospheres 117 (D8): D08101. https://doi.org/10.1029/2011JD017187.
Tal, Eran. 2013. “Old and New Problems in Philosophy of Measurement.” Philosophy Compass 8 (12): 1159–73. https://doi.org/10.1111/phc3.12089.
———. 2017. “A Model-Based Epistemology of Measurement.” In Reasoning in Measurement, edited by N. Mössner and A. Nordmann, 233–53. London: Routledge.

20 - Opacity, Marr, and the Norms of Explainable AI

Carlos Zednik

Machine Learning (ML) methods are a major catalyst for progress in Artificial Intelligence (AI). Unfortunately, computers programmed using ML methods such as deep and reinforcement learning are becoming increasingly opaque: it is difficult to “look inside” to know why or how they do what they do. The challenge posed by this opacity is known as the Black Box Problem in AI; in this talk I will concern myself with the extent to which solutions to the Black Box Problem proposed within the Explainable AI research program are adequate.

In order to properly understand the adequacy of proposed solutions, it is necessary to better understand the problem—that is, to clarify the meaning of ‘opacity’. I will argue that David Marr’s (1982) “levels of analysis” framework for explaining the behavior of cognitive systems can be redeployed to better understand the opacity of ML-programmed computers in AI. As on Humphreys’ (2009) influential analysis a system’s opacity is relative to a particular agent, on Marr’s framework, a system’s opacity is relative to one or more levels of analysis. Thus, the paradigmatic case of a deep neural network is one in which, although the implementational level may be fairly well-understood, questions remain at the algorithmic and possibly also computational levels.

Thinking about opacity in Marrian terms can also be used to better understand what is required in order to render ML-programmed computers transparent, and thus, to specify the norms of Explainable AI. Rendering an ML-programmed computer transparent requires finding answers to questions about “what” that system does (and “why” it does it) at the computational level, answering questions about “how” it does what it does at the algorithmic level, and answering questions

about “where” the relevant algorithms are carried out at the implementational level—and ideally, also answering questions about the way in which the “what”, “why”, “how”, and “where” relate.

So to what extent is recent work in Explainable AI poised to answer any or all of these questions, and thus, to provide an adequate solution to the Black Box Problem? Two distinct research strategies must be considered. On the one hand, the analytic strategy deploys mathematical techniques to analyze and visualize the activity within deep neural networks, but also invokes experimental techniques co-opted from psychology and neuroscience to infer the algorithms acquired via Machine Learning. This approach seems well-suited for answering “how” questions at the algorithmic level—analogous to the way these questions are answered in psychology and neuroscience. On the other hand, the synthetic strategy involves the development of new learning algorithms that do not only provide a problem-solving output, but that also “self-report” on their “reasons” therefore. This research strategy is arguably well-suited for answering “what” and “why” questions at the computational level. In order to combine the “what” and “why” with the “how”, it may be necessary to combine the analytic and synthetic strategies.