The Reasoner

Volume 18, Number 2
March 2024

thereasoner.org
ISSN 1757-0522

Contents

Editorial 9

Features 9
Interview with Vaishak Belle 9
Conditional beliefs aren't conditional probabilities . . 12

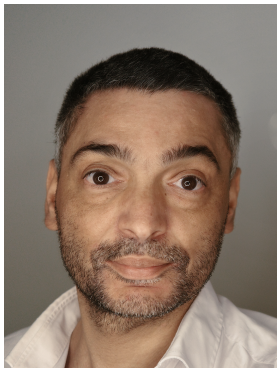
The Reasoner Speculates 13
Benefits of cybernetic models in philosophy 13

Dissemination Corner 14
SMARTEST 14
BRIO 15

Editorial

Dear Reasoners,

I am very happy to welcome you to this new issue of The Reasoner. It opens with my interview with Vaishak Belle and then features Jon Williamson arguing that conditional beliefs aren't conditional probabilities, Ferenc András suggesting the benefits of cybernetic models in philosophy, and ends with the dissemination of two exciting research projects.



As a short introduction to my guest, Vaishak is Reader at the University of Edinburgh, an Alan Turing Fellow, and a Royal Society University Research Fellow. He is one of the key contributors to the field known as neurosymbolic AI which, as you will see, tries to make the most of the two traditions in the field.

Since this topic is of great interest to many readers of The Reasoner, we are soliciting a *Focussed Issue* on it (see for instance here and here for two examples). Please send short proposals with the list of contributors to hykel.hosni@unimi.it.

Before leaving you to the interview, I'd like to thank warmly Vaishak Belle for his time and for the generosity with which he shared his views with us.

Hykel Hosni
Logic, Uncertainty, Computation and Information Lab,
University of Milan

Features

Interview with Vaishak Belle

Hykel Hosni: You are an expert in neurosymbolic AI, which is very much in the news these days.

Vaishak Belle: I'm not sure I would consider myself an expert, but I do find myself very interested in the area. One of the reasons I suppose it is difficult for anybody to declare themselves an expert in this field is because it is rapidly changing.

HH: Can you tell us what it is all about?

VB: In the early days, the term "neurosymbolic AI" was usually referred to formalisms that allowed neural architectures in logical languages: representations that combine some aspects of neural networks in a logic, especially fuzzy logic, which allows for real-valued truth. However, recently, neurosymbolic AI is better understood as formalisms that combine aspects of logical reasoning with deep learning.

HH: And since deep learning encompasses a variety of methods...

VB: ... there is no single agreed-upon definition, indeed! This obviously opens up the space for a wide range of solutions. For example, perhaps the most common kind of solution typically seen in robotics applications is when you have deep learning systems for vision and audio and language that are in-

terpreted using some kind of control framework – e.g., a symbolic automated planning framework might interact with one of these deep learning outputs to help the robot operate purposefully in its domain. This is often regarded as a loose coupling between logic and deep learning because it only allows a limited sense in which the semantics of the logical language captures what is happening inside the deep learning system.

HH: Can you give an example in which the connection between logic and deep learning is tighter?

VB: Sure. A deeper integration involves exploring ways to enable joint training or reasoning between logical systems and deep learning architectures. For instance, a popular area of inquiry which has recently been attracting interest is based on the idea of modifying the loss function of deep learning systems with logical formulas. This modification allows the distributions learned by the neural networks to capture the semantics of those formulas and constraints. Consequently, predictions can be ensured to adhere to physical and geometric properties of the domain. Another type of coupling involves extracting facts and relations from the web, say, using a deep learning system. These facts can then populate a database or an ontology over which a logical query-driven engine is built. Yet another area of inquiry is investigating the possibility of extracting symbolic structures, such as computer programs, from neural architectures. The idea is that these programs could be interpreted by humans and therefore become, in some sense, explainable. HH: What varieties of logic and deep learning are relevant in those applications?

VB: All the examples I just mentioned involve an interesting and often complicated mixing of model theory on the logic side and statistical learning and geometry from the deep learning side. Therefore, there is a very real possibility that neurosymbolic AI lays the foundations of a new type of AI that involves the best of both worlds.

HH: That may come as a surprise to (classical) logicians in the first place!

VB: It should be noted that the learning of logical formulas and the use of logic in machine learning are long-standing areas of research in their own right. Perhaps the most popular representation of this is statistical relational learning, which combines machine learning and probabilistic logical languages, such as relational Bayesian networks and probabilistic logic programs. This is yet another facet of how logic and deep learning can combine: by using a probabilistic logical formalism, distributions learned by deep learning models could be directly embedded in a logical language.

HH: Regular readers of The Reasoner will recognise this, as Felix Weitkämper has been running a column on statistical relational learning for quite some time!

VB: That's great! Although there are plenty of academic communities working on neuro-symbolic AI, the industry has been taking a very serious interest as well, especially considering that deep learning on its own seems to be data-hungry and often struggles in safety-critical applications, owing to issues such as distribution drift, and generally the lack of guarantees that comes with that. Thus, verifying the robustness of neural networks is an important topic, as is explainability owing to its black-box nature. Finally, because the use of machine learning in the real world doesn't often immediately reduce simply to prediction, there is an inherent need to combine structures and symbolic grammars with neural networks.

HH: Many are tempted by the analogy with dual systems of cognition where deep learning embodies the fast and highly fallible "system 1", whereas logic is asked to play the role of the slow and arguably reliable "system 2", in the terminology made popular by Daniel Kahneman.

VB: With the advent of large language models and their capacity for confabulations, the idea that perhaps one could use symbolic reasoners as a post hoc solution for correctness and consistency has been circulating. For example, Wolfram Alpha recently started to feature an integration with ChatGPT so that mathematically correct answers can be provided for questions of a mathematical or computational nature. The general idea is that whatever is uttered in natural language is processed by ChatGPT and converted to a form that can be interpreted by Wolfram Alpha, after which the symbolic solver returns the solution.

HH: One characteristic feature of the current AI spring is that it is driven by private companies who nonetheless appear to make significant scientific contributions. Of course I am thinking of DeepMind...

VB: Indeed! The recent AlphaGeometry approach by Google DeepMind, which made sensational news in the New York Times (17 January), attempts to solve geometry problems from the International Mathematical Olympiad. The key idea here too is to use the language model to create formal constructs and have a symbolic engine interpret these constructs to not only solve them but potentially give signals back to the language model for more effective problem-solving.

HH: Do you agree with those who think that this is yet another game changer from DeepMind?

VB: It should be noted that for AlphaGeometry to work they needed to generate a hundred million synthetic data examples. Such an effort might not be possible for everyone. However, as more and more of such synthetic samples are generated for numerous domains on which the language models are trained, it might eventually be possible to use one of these models in different settings, provided you have an appropriate symbolic reasoner to ensure that the responses are correct. Thus, neurosymbolic AI has a promising future, it seems.

HH: I can see expectations being really high! Can you tell us about your background?

VB: I completed my undergraduate degree in India in a field that could be considered closer to software engineering than computer science. I then pursued my master's degree as part of an Erasmus Mundus program between Germany and Italy. This was perhaps my first exposure to formal approaches.

HH: Was it the classic love at first sight?

VB: Not sure! Initially, I wasn't entirely convinced of their applicability in the real world. In India, the emphasis was often more on software engineering, as graduates were being trained for services-oriented software companies. It took me quite some time to rewire my way of thinking to develop an appreciation for theory.

HH: But I guess that happened quite quickly. Were you set to pursue the academic path after graduation?

VB: At that point I wasn't necessarily keen on an academic career per se. To be honest, I didn't quite know what it entailed,

but I did entertain the notion that a job involving writing and thinking all day was a fun career, if such a thing was possible at all.

HH: That sounds very familiar!

VB: I was also into science fiction, so in some sense, I was interested in artificial intelligence fairly early on. It was with my master's degree and the start of a Ph.D. that I slowly transitioned to becoming familiar with logic. Somewhat bizarrely, because I lacked a formal background, I ended up teaching myself about modal logic first, and never covered propositional or first-order logic in any course. Interestingly, in contrast, my master's thesis was on face recognition. Logic appealed to me, but when I began working on it, I still recognized the value of the machine-learning way of thinking, especially in the sense of extracting patterns from data through the training process.

HH: It is interesting to see how the hybrid approach to AI you are pursuing in your research is rooted in your very personal trajectory. So, after your masters, you started a PhD in Germany. What was its topic?

VB: At the beginning of my Ph.D.I was quite interested in interactive epistemology as it was making its way into game theory. Then I began to consider whether those kinds of formalisms could be useful in AI. Ultimately, this led me to work on epistemic and dynamic logic for my Ph.D. And few years into my Ph.D., I began to wonder if it would be useful to examine languages that combined the capabilities of logic and probabilistic reasoning.

HH: Which you took forward in your postdoctoral years.

VB: Exactly. After the PhD my work focussed on integrating probability and logic and, ultimately, on learning and logic.

HH: Which lead you to venturing in neurosymbolic AI. Many PhDs with this background would be attracted to industry careers. Did you consider that?

VB: I did, briefly. However, I was fortunate enough to obtain a postdoc, which seemed like a more natural choice. We –my partner and me– had to move to Canada for this, but ultimately it was the start of a wonderful adventure.

HH: You spent two years in Canada. What happened next?

VB: I held a postdoc fellowship in Belgium. After a little more than a year of that, I began applying for academic positions and was fortunate to obtain one here in Edinburgh, where I have remained since.

HH: What is the most exciting problem you are working on at the moment?

VB: At the moment, I am very interested in mechanisms for extracting logical knowledge using neural architectures, as well as the ways in which logical knowledge can be embedded as constraints in neural architectures. In some sense, both of these are begging the question: what kind of semantics and formal machinery best allows the representation of neural computations with logical knowledge? How does this affect using logical solvers as part of this architecture? And where should we draw the line, from a scalability point of view, to either rely completely on neural computations or completely on logical computations? There clearly needs to be a boundary that allows us to go back and forth to have the most effective way to reason about neurosymbolic computations. And that's a broad open challenge that I find very interesting. Ultimately, I suppose, it really is a way to get at the dichotomy between deduction, abduction and other kinds of deliberative computation versus reactive complications such as predictions from a neural network.

HH: Fascinating. We have covered a lot, but I am sure there is more in the pipeline! Can you tell us about your plans for the future?

VB: I have a couple of projects related to large language models and logic that I am looking into. But I suppose what is really keeping me occupied right now is organizing some of the ideas I mentioned in a kind of unified framework and seeing how this evolves in the next few months.

HH: Sure. Is there any advice you would like to give to PhD students who just started or are about to start?

VB: Two things stand over the others: do good science, and trust the process. To get started on doing good science, the nature of which can vary wildly from area to area, we need to have an understanding of the background literature and the foundations (e.g., keep a few textbooks in hand, and not just the latest works to study the lineage), and keeping the motivation and need for this result in mind, are the best ways to have a clear-cut goal, from which you can define a path.

HH: I can imagine them now being impatient to hear how they turn this into exciting research

VB: Of course, this is only the beginning! The results will come gradually, as long as we put in the work in a disciplined manner, and are consistent, and take a scholarly approach to the related work. It is important to be honest about the kind of results we desire and to acquire the necessary skills along the way. The nature of research is that it is often unfamiliar, and mistakes will be made. However, by learning from these mistakes, acquiring new knowledge, and constantly ensuring that we are not repeating past mistakes or reinventing the work of others, we can ensure satisfaction with the end result, whether positive or negative. The related work can provide guidance and feedback too, if studied properly, similar to that of a supervisor. So, it is important to see how others in the community approach the problems, their intuition, expertise, and knowledge, and work along those lines with attention to detail.

HH: You mentioned trusting the process.

VB: The process should be as enjoyable as the outcome. After all, science is supposed to be fun, so make sure that curiosity is not hindered and that the enjoyment of the process is as rewarding as the end result. Even if the desired outcome is not achieved, a lot of valuable lessons will have been learned along the way, making it easier to tackle the next problem.

HH: Indeed! Finally, can you share any reading suggestions for anyone serious about neurosymbolic AI?

VB: There are a couple of edited volumes on neurosymbolic AI, which, although not immediately accessible for a reader who isn't working on AI, still give away the most important ideas emerging in the space right now. But to me, to really get to the heart of Neurosymbolic AI, it might be helpful to look at some major books discussing common sense and the need for combining logic and learning more generally. For instance, a book by Gary Marcus and Ernie Davis titled *Rebooting AI*, and *Machines like us* by Hector Levesque and Ron Brachman. I think they capture the essence of what is required for a commonsensical AI agent to perform in a way that is reasonable with our view of the world. And even though they don't directly speak about current developments in Neurosymbolic AI, I believe they are relevant. From a technical perspective, they strongly advocate for why people should be considering the integration of logic and learning.

Conditional beliefs aren't conditional probabilities

The claim that conditional rational degrees of belief are conditional probabilities is falsified by the following simple counterexample.

Red Faces. Suppose that a fair six-sided die is to be rolled (proposition X) and that each face of the die is coloured red, blue or green (E). Consider the outcome that the number rolled will be three or greater (A). It is reasonable to believe A to degree $\frac{2}{3}$: given that the die is fair, each number has chance $\frac{1}{6}$ of being rolled, and four of the 6 numbers on the die are greater than or equal to three. So, if conditional degrees of belief are conditional probabilities, it is rationally permissible to set:

$$P(A|XE) = 2/3.$$

Now consider an alternative outcome: that the colour rolled (i.e., the colour of the uppermost face) is red (R). It is clearly reasonable to believe R to degree $\frac{1}{3}$, on the grounds that red is one out of the three possible colours and there is no evidence that favours one of these colours over any of the others. Thus it is permissible to set:

$$P(R|XE) = 1/3.$$

Now suppose in addition that the red faces are precisely those that are numbered three or greater, i.e., $A \leftrightarrow R$. Given that the die is fair, it is again clearly rationally permissible to believe A to degree $\frac{2}{3}$:

$$P(A|XE(A \leftrightarrow R)) = 2/3.$$

(Note that for these conditional probabilities to be well defined, it must be rationally permissible to set $P(XE(A \leftrightarrow R)) > 0$, i.e., to assign some positive credence to the claim that the die is fair, faces 3-6 are red and faces 1-2 are blue or green.)

It turns out, however, that these assignments of degree of belief are inconsistent: there is no probability function that satisfies them all (Wallmann & Williamson 2020: *EJPS* 10(3); Williamson 2023: *IJB* 19(2), 295–307). It is thus not possible to use conditional probabilities to validate the above judgements about rational permissibility: i.e., conditional beliefs are not always identifiable with conditional probabilities.

Consequences. Let belief function B represent a rationally permissible assignment of conditional degrees of belief: $B_C(A)$ is the degree to which proposition A is believed under condition C, for all A and C in a given domain of propositions. The claim that conditional beliefs are conditional probabilities can be formulated as follows:

CBCP. For any belief function B, there is some probability function P such that $B_C(A) = P(A|C)$ for all A and C.

In the red faces counterexample we have an assignment of degrees of belief that is clearly rationally permissible, yet cannot be captured by a conditional probability function. Hence CBCP is false.

This has two important consequences.

Firstly, if CBCP is taken to be constitutive of Bayesianism, as is standardly the case, then Bayesianism is untenable. The red faces problem threatens the tools of Bayesianism as well as its philosophical foundations. Bayes' Theorem is only of use if conditional probabilities are themselves of use, but this requires some connection between conditional probabilities and rational belief such as CBCP. Bayesian conditionalisation also apparently rests on CBCP: why update by means of conditional probabilities unless those conditional probabilities represent degrees of belief conditional on new evidence? Without Bayes' Theorem or Bayesian conditionalisation, Bayesianism would seem very impoverished.

Second, the 'new paradigm' in the psychology of reasoning, which seeks to understand our reasoning by appeal to conditional probabilities, is untenable without CBCP or something like it (Oaksford & Chater 2020: *ARP* 71(1), 305–330). For instance, the new paradigm analyses our use of conditional propositions in terms of conditional probabilities. This analysis involves two steps: an appeal to conditional beliefs to analyse cognition involving conditional propositions and then an application of CBCP to connect to conditional probability. Without CBCP, this analysis cannot succeed.

A Potential Resolution. The red faces problem shows that conditional beliefs can't always be construed as conditional probabilities. On the other hand, the successes of Bayesianism and of the new paradigm show that it can sometimes be helpful to identify conditional beliefs with conditional probabilities. What we need is a more fundamental theory to explain the successes and failures of CBCP.

There is a non-standard approach to Bayesianism that might help here (Williamson 2010: *In defence of objective Bayesianism*, OUP). This version of Bayesianism identifies conditional beliefs with probabilities, but not conditional probabilities:

CBP. For any belief function B and proposition C, there is some probability function P_C such that $B_C(A) = P_C(A)$ for all A.

How are these unconditional probabilities obtained? Firstly, P_C must satisfy constraints imposed by C—in particular, constraints imposed by calibration to chances: if one establishes from C that the chance of A is x then $P_C(A) = x$, as long as C doesn't imply anything that defeats this ascription (e.g., proposition A itself). Second, P_C should be maximally equivocal with respect to propositions whose probability isn't determined by constraints imposed by C. This is typically explicated by setting P_C to be the function, from all those that satisfy constraints imposed by C, that has maximal entropy.

This version of Bayesianism is immune to the red faces problem: it will consistently set $P_{XE}(A) = 2/3$ (by calibrating to the chance information in X), $P_{XE}(R) = 1/3$ (equivocating between the three possible colours), and $P_{XE(A\leftrightarrow R)}(A) = 2/3$ (by calibration to chance again).

The theory can also help to explain when it is safe to conditionalise. If (i) learning D only imposes the constraint $P(D) = 1$, (ii) $P_C(D) > 0$, and (iii) $P_C(\cdot|D)$ satisfies all the constraints imposed by C, then it is safe to conditionalise on D, i.e., $P_{CD}(\cdot) = P_C(\cdot|D)$; see Result 1 of Seidenfeld (1986: Entropy and Uncertainty, *PoS* 53: 467–491) and Theorem 5.16 of Williamson (2017: *Lectures on inductive logic*, OUP). In the red faces problem, it is not safe to conditionalise on $A \leftrightarrow R$

because $P_{XE}(\cdot|A \leftrightarrow R)$ does not satisfy all the constraints imposed by XE. In particular, as the Appendix of Williamson (2023) shows, $P_{XE}(A|A \leftrightarrow R) = 1/2 \neq 2/3$, the value required by calibration to the chance information in XE.

Thus, although this version of Bayesianism may seem unorthodox, it is explanatory. In any case, a significant departure from Bayesian orthodoxy is required to avoid *red faces*.

Jon Williamson
University of Kent

The Reasoner Speculates

Benefits of cybernetic models in philosophy

A common research method among philosophers is the usage of thought experiments. Take for example John Searle's 'Chinese Room' or Frank Jackson's 'Mary's Room' argument. David Lewis goes further by using neuron diagrams to represent causality in his counterfactual theories of causation. His method has since been further refined and developed. Interestingly, the usage of logical circuits or finite automaton to represent causal relations has not yet been considered. As an advantage, the latter can be visualised in cyberspace using spreadsheets and tested in practice. Furthermore, it is not only in the problem of causality that cybernetic models can fruitfully be used to provide a philosophical explanation, they can also be utilised to represent logical semantic problems. Let us consider an example for this.

Many logic handbooks allude to the obvious connection between propositional logic and logic circuits. Truth functions in logic can be represented by logic circuits in which the high or low voltage levels of the circuits correspond to the true and false logic values, respectively. At the propositional logic level, the logical connectives of propositions can be simulated by logic circuits as follows: the true or false logical evaluation of atomic propositions corresponds to the high or low level of the circuit input and the truth value of compound propositions corresponds to the circuit output state. A high circuit output signifies that the compound sentence is evaluated as true, whereas a low output indicates that the compound sentence is false. It is well known that in the world of logic circuits, the AND connective in logic corresponds to the AND gate, the OR connective to the OR gate and the negation operation to the inverter. The output of a circuit equivalent to contradiction is always low and that of a circuit corresponding to tautology is always high irrespective of the input state. The remaining compound formulas correspond to logic circuits with a high output level for some inputs and low output level for other inputs. However, what logical circuit can model a circular sentence?

Indeed, every formula in propositional calculus can be modelled based on an equivalent logic circuit, specifically referred to as a combinational logic circuit. However, not all logic circuits are combinational logic circuits. The range of logic circuits is wider than that of the combinational logic circuits. It includes logic circuits whose input states do not determine unambiguously their output states, i.e. the output is not a function of the input. This is because the circuit has feedback. Circuits that contain feedback are called sequential logic circuits. Although every formula in propositional calculus can be modelled based on an equivalent combinational logic circuit, it remains unclear whether the converse theorem is valid. Can every sequential logic circuit be equivalent to a formula in propositional calculus? Does any formula at the propositional logic level correspond to sequential logic circuits?

Sequential logic circuits have memory owing to feedback mechanisms. (The operation of these circuits is mathematically isomorphic to that of a finite automaton. Examples of such circuits include flip-flops, registers, counters, clocks and memories.) The output state of sequential logic circuits is not a function of the input states but depends on previous input states. In contrast, the truth value of formulas in propositional calculus is a function of the evaluation of atomic formulas, without considering previous evaluations of these formulas. Therefore, the answer is negative; logical formulas cannot be simply matched with sequential logic circuits at the propositional logic level. However, logical relations between sentences may exist beyond propositional logic, corresponding to the operation of certain sequential logic circuits. What type of logic relationships can sequential circuits model? In the following text, I will provide a simple example of this.

Jean Buridan's paradox sentence As an influential medieval French philosopher of his age, Jean (John) Buridan (c. 1295–1358) presented a puzzle with the following essence:

> Twelfth sophism: God exists and some conjunction is false.

John Buridan (2001: *Summulae de dialectica* (translated by Gyula Klima), Yale University Press, c.8, p.980)

Or in other words:

> God exists and none of the sentences in this pair is true.

What do you think about the truth value of these two sentences? Which of these two is true?

> p := God exists.
>
> q := Neither sentence p nor q is true.

'p' is true if God exists and false, otherwise. 'q' is true if neither p nor q is true.

Sentence q asserts a 'Not-OR' relation because 'neither p nor q' is equivalent to 'not (p or q)'. One component of the 'or' relation is an existential proposition, while the other is the 'or' relation itself. It is a peculiar sentence because it has a truth value, if it has any at all, which depends on itself. Therefore, it certainly cannot be translated into the classical first-order logic language.

Let us examine the logical possibilities. If p is true (i.e. God exists), then q is false because one of its components is true and the other is false. Consequently, the two together are false (i.e. q is false). The situation is not that simple if p is false (i.e. we deny God's existence). Suppose that q is true. This is possible only if both members are false. This is not, however, the case because the first member is false and the second member is true; hence, the result is false together and q cannot be

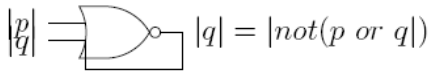


Figure 1: NOR gate

true. Let us now assume the opposite that is, q is false. Both members of q (q = not p and not q) are false because we denied God's existence. Consequently, q must be true, contrary to our assumption. Again, we are stuck in a contradiction. q can be neither true nor false. We alternately evaluate it as true or false. We are caught in a trap, from which the only way out is to assume that God exists.

Simulation of paradox applying logical circuits Buridan's sentence has a constant truth value only if p is true, that is, we assume that God exists. In which case, q is false. This paradox cannot be expressed in the classical formal logic language, but can be presented using logical circuits.

Logical circuits have either a high or low state. High and low levels correspond to true and false, respectively. The NOR (Not-OR) gate output is low if any of the inputs is high. In our case, the NOR gate has p sentences at one input and q sentences at the other input. The NOR gate output is also a q sentence. With this solution, the NOR gate output is fed back to one of its inputs. This feedback simulates the self-dependence of the truth value of Buridan's sentence. The logical circuit will work exactly considering that we have examined the logical possibilities. (See Figure 1. $|x|$:= truth value of sentence x)
The NOR gate output was connected to one of its inputs. This way, we can use the feedback to simulate the circularity of the truth value of a sentence q. The input p is high if God exists, but low if God does not exist. The output state of the automaton is fed back to the other input, thereby corresponding to the sentence 'neither one nor the other is true.' If the first input has a high level (i.e. God exists), then the NOR gate output has a low level (i.e. the sentence q is false). Conversely, its first input is low (i.e. God does not exist), the NOR gate output will alternate between low and high, such that sentence q will not have a constant truth value. The cybernetic model exactly simulates the logical paradox. The logical circuits can be represented using spreadsheet software and tested in operation. Please see: https://sht.andrasek.hu/buridan3ext.xlsx

As mentioned, Buridan's sentence is a paradox. It cannot be translated into the classical formal logic language; however, the operation of the cybernetic model that simulates the paradox can be. Its operation is consistent, not paradoxical, which is the benefit of such models.

Ferenc András
Pomaz, Hungary

Dissemination Corner

SMARTEST

Simulation of Probabilistic Systems for the Age of the Digital Twin

Our general world-view, our scientific understanding and our practical daily interactions with reality are today fully mediated by digital technologies. From our smartphones to simulation technologies, from digital personas to services and tools in the workplace. Maybe the pinnacle of this hybrid reality is embodied by the concept of the digital twin (DT) – a term coined in 2001 by pioneering technologist Michael Grieves – denoting digital replicas of physical systems. They serve to test and understand how systems and products might behave during their lifecycle. To this purpose, virtual environments and digital simulations are used, aided by data-driven algorithms. The DT strategy is largely endorsed at the European level, with projects as diverse as mapping the Earth and the Ocean, to the support of SMEs, and it is at the basis of the Italian Recovery and Resilience Program (PNRR) for the Transition 4.0 area.

In the light of this new human condition, ripe with opportunities and excitement, one ground question arises: o what extent does a digital artefact faithfully represent the reality it intends to emulate? This question is even more pressing in the light of essential uncertainty and non-determinism dictated by Machine Learning (ML) and Deep Learning (DL) algorithms intervening in the simulation process, and with huge amounts of data used to draw possibly weak correlations about potential behaviours of the system under analysis.

An answer to our question can be cast in formal, epistemological and ontological terms. From a formal point of view, the quest for safe and reliable digital artefacts simulating real entities is one of identity criteria under algorithmic feasible preservation of relevant properties: one wants to model formally the entities under observation and to compute and verify how much interesting properties like safety, reliability and accuracy of one are preserved by the other. This task can be developed both semantically and syntactically and techniques abound in the literature: but the combination of characteristics dictated by AI elements of these systems (from uncertainty to partiality and high bias) are new to the problem and require a significant change of paradigm. From an epistemological perspective, one aims at establishing whether the empirical and experimental nature of the knowledge inferred from the digital simulation of artefact is robust enough to grant those characteristics that we usually ascribe to verified contents, and what errors can occur. Finally, from an ontological point of view, the aim is to provide the ground criteria for the construction of reliable digital artefact, through the analysis of those notions already available in the philosophy of technology, like copia and replica.

All these questions are at the core of SMARTEST (Simulation of Probabilistic Systems for the Age of the Digital Twin), a new project funded by the Italian Ministry of University and Research (MUR) through the scheme PRIN 2022. The research consortium consists of 3 units, for a total of 4 permanent researchers and 3 postdocs to hire. The Laboratory for Applied Ontology (LOA, Trento) part of the Institute of Science and Technologies of Cognition of the National Research Center, has extensive expertise in formal ontology, mathematical logic, and epistemology with more than 15 years experience in formal and computational modelling in the engineering domain. This Unit will be led by dr. Roberta Ferrario. The Department of Cognitive, Psychological, Pedagogical Sciences and Cultural Studies at the University of Messina contributes with expertise in the

fields of philosophy of computing, epistemology of computer simulation, and formal verification. This Unit will be led by dr. Nicola Angius. The LUCI (Logic, Uncertainty, Computation and Information) Lab, part of the PhilTech Research Center at the Department of Philosophy of the University of Milan hosts one of the most active Italian and European research groups for formal and applied logics, with a particular focus on modelling and verification of real systems, including both human and AI systems and their interaction. This unit will be led by Prof. Giuseppe Primiero, who also acts as PI of the project.

The project will contribute to a better understanding and safer deployment of the digital transition underpinned by the latest European political and economical development plans. Follow our updates in future Columns in this venue, or on our website https://sites.unimi.it/smartest/.

Giuseppe Primiero
Logic, Uncertainty, Computation and Information Lab,
University of Milan

BRIO

Ascribing trustworthiness to AI systems

The notion of Trustworthy AI (TAI) has been playing an increasingly central role in discussions on the responsible and ethically acceptable development and deployment of AI systems. Analysing the epistemological and normative grounds of the notion of TAI is among the aims of the research project *BRIO – Bias, Risk and Opacity in AI* (PRIN MUR). In particular, BRIO's research unit based at Politecnico di Milano has been working on conceptual issues related to the attribution of trustworthiness to AI systems.

The road to TAI is long and obstacles abound. First, there is no agreement on the *determinants* of trustworthiness in AI – *i.e.*, what makes an AI system trustworthy. What is more, it is dubious that some features that are typically deemed necessary for TAI are actually feasible for all AI systems. A prominent case involves explainability, which is systematically taken to be a fundamental ingredient of TAI and yet is hardly achievable in many systems. Upstream of all of this, however, there is an additional, foundational problem: from a conceptual point of view, it is unclear whether the very ascription of trustworthiness to AI systems could be a legitimate move.

In a nutshell, the problem is the following. Our understanding of trust is shaped by the dynamic of relationships between *persons*, that qualify as full moral agents capable of having intentions and motivations as well as adhering to moral norms. And indeed, we consider trustworthy a person that not only is able to perform the task we delegate them with but also does so willingly, motivated by benevolent intentions, and acts in accordance with the moral obligation to fulfil their commitment. On the contrary, we would not consider trustworthy someone who performs a duty only because they have no choice and would immediately betray us, if only they had the chance.

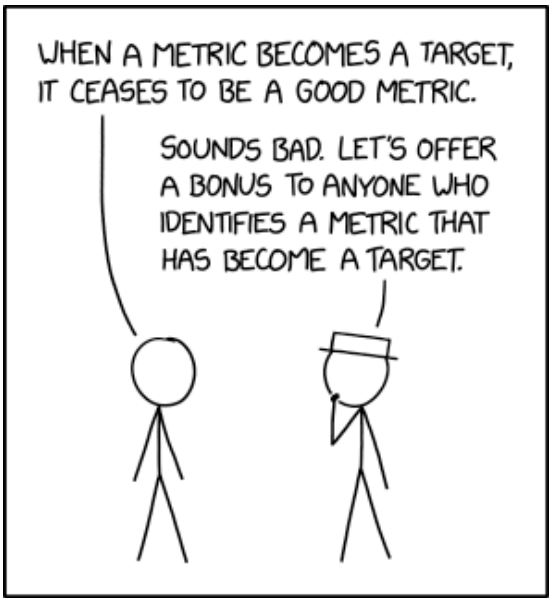
More generally, there is an ineliminable moral dimension when it comes to trustworthiness, which is typically conceived as inherently depending on the trustee's interests, motivations, and moral obligations. Given these premises, we can easily anticipate the short circuit in the ascription of trustworthiness to AI systems, which simply do not possess motivations and intentions, and cannot adhere to moral obligations – at least, not in the wilful and conscious way a full moral agent can do it. On these grounds, the notion of TAI has been criticized from various quarters as "conceptual nonsense" (Metzinger, 2019), highlighting how it would reinforce our tendency to unduly anthropomorphise AI systems and would veil their developers' and users' responsibility (DeCamp & Tilburt, 2019; Fossa, 2019; Hatherley, 2020; Ryan, 2020).

As a research unit of the BRIO project, we have tackled these issues in a recently published article (Zanotti, Petrolo, Chiffi, Schiaffonati. Keep trusting! A plea for trustworthy AI. *AI and Society*. https://doi.org/10.1007/s00146-023-01789-9). More precisely, our analysis was aimed at answering two strictly interrelated questions: *why* we should want to employ the notion of trustworthiness to characterize AI systems and *how* we could do it without making conceptual errors.

To answer the first question, we took our starting point from the role the notion of TAI plays within the European *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019). Now, these guidelines can hardly provide an out-of-the-box answer to the conceptual issues at stake here. Still, they can offer a glimpse on a productive use of the notion of TAI. In particular, we emphasized how they identify at least two kinds of requirements AI systems should have to be deemed trustworthy. To begin, they should possess features that, consistently with the literature in the philosophy of technology, we subsumed unto the umbrella term of *reliability*. Most notably, they should be accurate in their predictions, classifications, and decisions as well as robust in their functioning. However, this is only half of the story. In addition to this, trustworthy AI systems must also meet a series of requirements having to do with the respect of human autonomy and privacy, fairness, transparency, accountability, and societal and environmental well-being.

In our view, the integration of aspects related to algorithmic performance and the ethical dimension of the development and use of AI systems is what makes the notion of TAI so pivotal and worth employing. This is especially true if the alternative is sticking to the mere notion of *reliable* AI, as some of the detractors of the concept of TAI suggest. Although we maintain that reliability is a crucial ingredient of trustworthiness, we argue that it is not enough. As we have seen, trustworthy people need to be both competent with respect to the delegated task and have the right moral profile, so to say, when it comes to their motivations and their adherence to moral obligations. Analogously, trustworthy systems need to be reliable but also ethically developed and used.

Having identified the crucial role of trust in encompassing both these aspects, we provided a possible solution for ascribing trustworthiness to AI systems without falling into categorical errors. The literature on TAI often takes for granted that the notion of trustworthiness in AI should be *uncompromisingly* modelled on its personal counterpart, thereby inheriting the focus on the trustee's motivations, interests, and moral obligations. We proposed to abandon this methodological assumption – which, by the way, is typically left unwarranted – in favour of a different approach that leaves room for two distinct notions of trustworthiness: one for persons and one for

AI systems. Needless to say, there is a common ground: in both cases, trustworthiness results from the interplay between the trustee's reliability and ethical and value-laden aspects – we also touch upon the relation between trustworthiness and risk, but this story shall be told another time (stay tuned!). What changes is the way the ethical dimension of trustworthiness is realized. Unlike in the case of persons, where motivations and moral obligations matter, the ethical dimension of AI systems' trustworthiness has to do with the respect for human autonomy, fairness, and so forth. This conceptual distinction paves the way for a legitimate and meaningful use of the notion of TAI.

GIACOMO ZANOTTI
Politecnico di Milano