Verifiability as a Complement to AI Explainability: A Conceptual Proposal

Kaustubh R. Patil
k.patil@fz-juelich.de
Forschungszentrum Jülich
Institute of Neuroscience and Medicine: Brain and Behaviour (INM-7)
52425 Jülich, Germany
and
Heinrich-Heine University Düsseldorf
Medical Faculty – Institute of Systems Neuroscience
40225 Düsseldorf, Germany

Bert Heinrichs
b.heinrichs@fz-juelich.de
Forschungszentrum Jülich
Institute of Neuroscience and Medicine: Ethics in the Neurosciences (INM-8)
52425 Jülich, Germany
and
Rheinische Friedrich-Wilhelms-Universität Bonn
Institute of Science and Ethics
53113 Bonn

Abstract
Recent advances in the field of artificial intelligence (AI) are providing automated and in many cases improved decision-making. However, even very reliable AI systems can go terribly wrong without human users understanding the reason for it. Against this background, there are now widespread calls for models of "explainable AI". In this paper we point out some inherent problems of this concept and argue that explainability alone is probably not the solution. We therefore propose another approach as a complement, which we call "verifiability". In essence, it is about designing AI so that it makes available multiple verifiable predictions (given a ground truth) in addition to the one desired prediction that cannot be verified because the ground truth is missing. Such verifiable AI could help to further minimize serious mistakes despite a lack of explainability, help increase their trustworthiness and in turn improve societal acceptance of AI.

Keywords
Artificial intelligence (AI), machine learning, explainability, verifiability, reliability

1. Introduction

Recent advances in the field of artificial intelligence (AI) are providing automated and in many cases improved decision-making in critical areas, including medical diagnostics and therapy [1–3], but also in many other areas. This raises high hopes for positive impacts on society. However, these hopes are countered, among other things, by fears associated with the rare but unintelligible mistakes generated by AI. As some prominent examples have shown even very reliable AI systems can go terribly wrong without human users understanding the reason for it (Heaven 2019). Accordingly, these are systems that are very difficult or impossible for humans to assess. Against this background, there are now widespread calls for models of "explainable AI" (XAI), which should make it possible to make AI decision-making transparent and comprehensible for human users. This demand is generally correct. However, it leads to some follow-up problems: On the one hand, explainable AI will often have a lower accuracy, so that difficult trade-offs between the two objectives will have to be found. Secondly, current explanations already overload human understanding [4] and the growing complexity of AI could exceed the limits of human comprehensibility, so that especially high-performance systems cannot fulfil the requirement of explainability. This demand for XAI, albeit with its own merits, seems somewhat paradoxical against the historical backdrop of the failure of "expert systems"—an older kind of AI-–which was partly responsible for the AI winter in the late 1980s[5]. One of the challenges faced by expert systems was the difficulty in knowledge acquisition, i.e., extracting and capturing the domain knowledge from an expert to codify it as a part of an expert system. Experts found it difficult to describe their knowledge and decision processes, especially in a way that could be automated. But now, ironically, we are demanding that AI systems to provide explanations regarding their decisions and their inner workings in a human understandable form. In other words, we are expecting AI to provide explanations to us that we humans cannot provide to ourselves. Consequently, we should consider that the

limitations of human reasoning extend to AI systems as they "are not significantly more opaque than human brains/minds" making any decision—human or AI—nontransparent [6].

We therefore propose another approach as a complement, which we call "verifiability". In essence, it is about designing AI so that it makes available multiple verifiable predictions (given a ground truth) in addition to the one desired prediction that cannot be verified because the ground truth is missing. Such a verifiable AI could help to further minimize serious mistakes despite a lack of explainability, help increase their trustworthiness and in turn improve societal acceptance of AI.

2. Reliability is a necessary but insufficient criterion

Today's AI is often very reliable, i.e., the systems only rarely generate results that turn out to be wrong in retrospect. If you compare AI with other technologies, you will hardly be able to say that it is a particularly error-prone technology. The problem does not really lie in the frequency of the errors, but rather in the fact that the errors that occur are incomprehensible to us.

In a news feature article in Nature, Douglas Heaven compiled a whole series of studies showing how easy it is to mislead Deep Learning systems [7]. The amazing thing is not that highly complex systems fail at certain tasks, but much more that we cannot understand how they can err so thoroughly at tasks that seem completely trivial to us. One of the by now probably quite well-known examples is about the recognition of street signs. A specially trained deep neural network fails to recognize a stop sign when it has some small rectangular black and white areas on it. Instead of identifying it as a stop sign, the system classified the sign as a speed limit (45). No child who knows road signs would make this mistake. It is completely incomprehensible that the system could be so wrong. Countering this by pointing out the rarity of the errors the system makes does not seem to be enough. A system should simply never make such errors.

Or, to put it the other way around, a system that makes such errors seems highly dubious to us, even if they are rare.

To put it bluntly, one could say that the use of AI is a kind of Russian roulette. In rare cases it hits someone, and you never know when and why. This is obviously not a very attractive prospect. So, we should try to design AI in a way that avoids the problem in the first place.

3. Explainability as an attempt to solve the problem

One currently much-discussed approach to solving the problem described is to make AI explainable [8]. The rationale behind this approach is as simple as it is obvious: if the epistemic opacity of AI means that we cannot understand errors (as well as correct outcomes), then we need to overcome opacity. The "right to explanation" is also enshrined in the European Union's General Data Protection Right, though its scope and impact yet remains to be seen [9]. Indeed, many technical solutions have been proposed to achieve explanations of different kinds [10–13].

An example of XAI is prediction of schizophrenia symptoms using neuroimaging data where the predictive ability of literature derived networks (i.e., a collection of brain regions whose interaction is then derived from neuroimaging data) such as theory-of-mind and default mode network was used to ascribe network-level interpretation that neuroscientists and psychiatrists are familiar with [14]. The same work could then extend the interpretation by identifying predictive importance of network edges as it was based on a "conventional" machine learning algorithm called relevance vector machine [15] which is considered intrinsically interpretable [16]. These two types of interpretations—network-level and edge-level—demonstrate the complexities associated with an apparently simple application using interpretable models. Deep neural networks, on the other hand, are not intrinsically interpretable which makes it harder to understand their internal architecture and several methods have been proposed to extract explanations—most of them post-hoc—each with its own pros and cons. A detailed review of these methods is out of scope, and we refer the reader to recent reviews [11,13].

As convincing as XAI may be as a research program, it raises two fundamental problems: Firstly, it must be assumed that transparency is always bought at the expense of performance or accuracy. If one decides in favor of explainable AI, then one will at the same time decide in favor of less efficient systems. On the other hand, it is not unlikely that as the amount of data and computational capacity grows very powerful systems will become so complex that the methods of explainability will quickly reach their limits. The billion parameter natural language models that were state-of-the-art just a few years ago, e.g., the GPT-3 model with 175 billion parameters (already a big leap compared to its predecessor GPT-2 with 1.5 billion parameters [17] trained by OpenAI researchers on 570 GB of clean data [18] now look miniscule in front of trillion parameter models. The trillion-parameter race is on and the natural language model "with outrageous numbers of parameters", 1.6 trillion [19] is now surpassed by the multimodal (learning from both text and images) model with 1.75 trillion parameters [20]. Technological and conceptual advances make it possible to train such large models and this trend is likely to continue. The emphasis of these models is on better performance by learning complex representations that can be only learnt by complex models. Explaining such models is going to be clearly difficult if not impossible even for expert human audience.

Furthermore current XAI is limited and faces many challenges, biases deter learning and explanations [2], information overload might preclude interpretation [4], techno-scientific explanations might not be desirable [6], human subjectivity in background and expertise in AI [21] and cognitive biases [22] make it challenging to design and consume XAI, the explanations might lack robustness [23,24], and human skill might impact modern deep learning models in ways that are hard to explain [25]. Such considerations have led to calls for abandoning explaining complex models in favor of simpler interpretable models for high stake decisions [26]. It is worth noting that AI, as in deep neural networks, made great headway in the last decade because of their complexity and technical advances devoid of an explicit requirement for expandability which only became a part of the narrative at later stages. Explanations, if available, are a welcome

outcome but imposing them as a requirement can slow the rate of development of AI systems, which still have a long way to go.

These considerations of explainability are not intended to fundamentally call the concept into question. We merely want to point out inherent problems of the concept and argue that explainability alone is probably not the solution. But what then?

4. Verifiability as a new method for safeguarding AI

Let us say we developed an AI to predict a mental disease and trained it on large amounts of data. Let us further assume that the system has achieved a remarkably high level of accuracy in tests. Finally, let us assume that the system is so complex that current methods of explainability fail, i.e., we do not really understand how the system arrives at its (predominantly correct) results. This is certainly not a science fiction scenario, but rather corresponds to the current state of research. What could help us to accept the results of the system and, perhaps more importantly, what could help us to detect errors in the system? One obvious answer is: further forecasts provided by the system, which we ourselves can check directly. Let us assume that we would not only let the system predict the occurrence of the mental illness (which of course we cannot check ourselves, that is what we need the AI for), but further characteristics of the person that are known to us (e.g., age and sex), but are not explicitly contained in the data about the person (e.g., MRI-based neuroimaging) with which the AI was fed. If all these other predictions were accurate, then they would in a way verify the actual prediction we are interested in. Of course, it would not be a verification in the sense of a watertight proof. It could be that the AI is simply wrong in all predictions equally. If the predictions were somewhat independent of each other, however, the probability of this would be very low. So, it might be promising to develop just such a verifiability, at least in such cases where explainability does not work or is too limiting.

As a welcome side effect, the verifiable AI will be resilient to biased decision making as suspected biases (e.g., sex differences) could be incorporated as verification tasks (e.g., sex prediction). Furthermore, verification can also serve as an explanation. Take the classic example of the dog versus cat classification. If we have a verification task of "color" of the animal in addition to the species, then a human can verify the color (assuming species itself is not verifiable). This will result in a more robust classier as it also needs to provide an explanation that is the color in this case.

The issue of confounding [27] with which the scientific community as well as more recently the machine learning community has grappled with can also be handled within the verifiable AI framework. Broadly speaking, confounding refers to unwanted biases in the data that get encoded in the model biasing what they learn and their predictions. If such confounding information is available, it can be conceptualized as a “non-learning task” where the prediction should be (on average) wrong—essentially verifying chance-level predictability of a confounding task. Luckily, the AI community has already provided solutions to this effect, for instance the Fader Network architecture [28].

A more complex output-space of AI is also advocated as a response to requirements of a complex task, e.g., for use of language in a social context [29]. Such synergies and associated technological developments can thus help the field to grow. In a sense, we propose a more “general” AI than a single-task AI but still restricted nonetheless and far from the goal of human-like AI which has been argued cannot be realized [30].

Taken together, existing technology can be indeed used to train a verifiable AI if the verification tasks are defined, and corresponding data is available. Essentially, verifiable AI will need to learn representations that are common across the several tasks it is expected to perform as depicted in Figure 1.

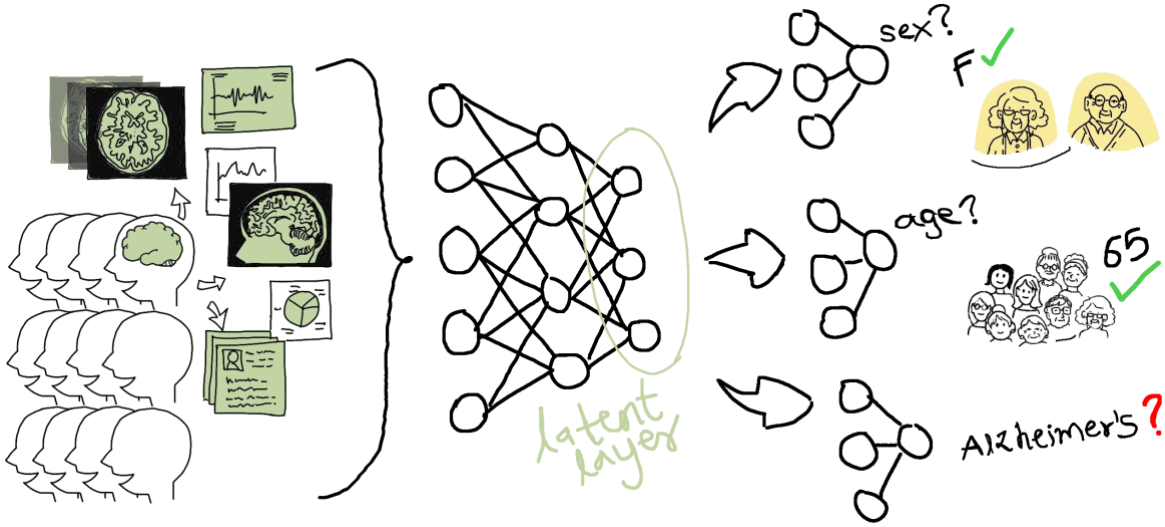

Fig 1. Proposed verifiable AI framework. The "big data" is fed into a deep neural network which first learns a latent representation. This latent representation is then used for prediction of different tasks. Here the sex and age are verification tasks while predicting Alzheimer's is the task of interest.

The outputs of a verifiable AI model can be either directly consumed by a human or they might be fed into machine-learning models to quantify confidence in a prediction. Such a model, of course, needs to be trained using historical gold-standard data.

5. Objections

The verifiable AI framework raises several (obvious) questions. Primarily, what the verification tasks should be and what is their sensitivity and specificity for the goal of verification. For instance, do different primary tasks (e.g., prediction of AD status or prediction of Schizophrenia severity from neuroimaging data) require different verification tasks? The answer is obviously yes, but that leads to other the hard-to-answer question, which ones? So, if not properly conceived and implemented, verifiable AI will be just a gimmick rather than a useful tool. It can also cause a loss of accuracy as including additional tasks that do not share similar latent representation will degrade the performance.

Furthermore, one could object that the proposed concept does not solve the actual problem of opaqueness and inexplicability, but merely conceals it. It is true, of course, that a verifiable AI

would still be opaque. However, we do not even want to claim that this problem is solved. On the contrary, our argument is rather that it could be that - with increasing complexity of the systems - it is unsolvable. The concept of verifiability proposed here is an attempt to deal with inexplicable AI. If we have to decide whether or not to use unaccountable systems, then perhaps verifiability could support a positive response in favor of AI.

6. Conclusion

In the foregoing, we have made a conceptual proposal of how AI might be developed. We assumed that the growing complexity of AI will lead to the fact that the - basically correct - approaches for "explainable AI", which are currently pursued by many, will reach their limits. Verifiability could provide a way to enable the estimation of errors even when a system is not understood. This could make the use of AI defensible even when it is opaque to us.

Acknowledgements

KRP was partly supported by the Deutsche Forschungsgemeinschaft (DFG, PA 3634/1-1) and the Helmholtz-AI project DeGen.

Bibliography

1. Gordon, L., Grantcharov, T. & Rudzicz, F. Explainable artificial intelligence for safe intraoperative decision support. *JAMA Surg.* **154**, 1064–1065 (2019).

2. Ghassemi, M. & Nsoesie, E. O. In medicine, how do we machine learn anything real? *Patterns (N Y)* **3**, 100392 (2022).

3. Heinrichs, B. & Eickhoff, S. B. Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Hum. Brain Mapp.* **41**, 1435–1444 (2020).

4. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. & Wallach, H. Manipulating and Measuring Model Interpretability. in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI, 2021).

5. Haenlein, M. & Kaplan, A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review* **61**, 5–14 (2019).

6. Maclure, J. AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. *Mind. Mach.* **31**, 421–438 (2021).

7. Heaven, D. Why deep-learning AIs are so easy to fool. *Nature* **574**, 163–166 (2019).

8. Shaw, J., Rudzicz, F., Jamieson, T. & Goldfarb, A. Artificial intelligence and the implementation challenge. *J. Med. Internet Res.* **21**, e13659 (2019).

9. Goodman, B. & Flaxman, S. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AIMag* **38**, 50–57 (2017).

10. Karanasiou, A. P. & Pinotsis, D. A. A study into the layers of automated decision-making: emergent normative and legal aspects of deep learning. *International Review of Law, Computers & Technology* **31**, 170–187 (2017).

11. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy (Basel)* **23**, (2020).

12. Hancox-Li, L. Robustness in machine learning explanations. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

13. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I. & Atkinson, P. M. Explainable artificial intelligence: an analytical review. *WIREs Data Mining Knowl Discov* **11**, (2021).

14. Chen, J. *et al.* Intrinsic Connectivity Patterns of Task-Defined Brain Networks Allow Individual Prediction of Cognitive Symptom Dimension of Schizophrenia and Are Linked to Molecular Architecture. *Biol. Psychiatry* **89**, 308–319 (2021).

15. Tipping, M. E. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research* 211–244 (2001).

16. Molnar, C. *Interpretable Machine Learning*. (2019).

17. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. *undefined* (2019).

18. Brown, T. *et al.* Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* (2020).

19. Fedus, W., Zoph, B. & Shazeer, N. [2101.03961] Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv* (2021).

20. Zou, X. *et al.* Controllable Generation from Pre-trained Language Models via Inverse Prompting. in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* 2450–2460 (ACM, 2021). doi:10.1145/3447548.3467418.

21. Ehsan, U. *et al.* The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *arXiv* (2021).

22. Brynjolfsson, E. Where Humans Meet Machines: Intuition, Expertise and Learning. *MIT IDE | MIT Initiative on the Digital Economy | Medium* https://medium.com/mit-initiative-on-the-digital-economy/where-humans-meet-machines-intuition-expertise-and-learning-be639f00bade.

23. Zhou, Y., Booth, S., Ribeiro, M. T. & Shah, J. Do Feature Attribution Methods Correctly Attribute Features? in *AAAI-22* (2022).

24. Alvarez-Melis, D. & Jaakkola, T. S. On the Robustness of Interpretability Methods. in (WHI2018, 2018).

25. Anand, K., Wang, Z., Loog, M. & van Gemert, J. Black Magic in Deep Learning: How Human Skill Impacts Network Training. in *British Machine Vision Conference* (2020).

26. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).

27. More, S., Eickhoff, S. B., Caspers, J. & Patil, K. R. Confound removal and normalization in practice: A neuroimaging based sex prediction case study. in *ECML PKDD 2020: Demo Track* (eds. Dong, Y., Ifrim, G., Mladenić, D., Saunders, C. & Van Hoecke, S.) vol. 12461 3–18 (Springer International Publishing, 2021).

28. Lample, G. *et al.* Fader networks: manipulating images by sliding attributes. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 5969–5978 (2017).

29. Kovač, G., Portelas, R., Hofmann, K. & Oudeyer, P.-Y. SocialAI: Benchmarking Socio-Cognitive Abilities in Deep Reinforcement Learning Agents. *arXiv* (2021).

30. Fjelland, R. Why general artificial intelligence will not be realized. *Humanit. Soc. Sci. Commun.* **7**, 10 (2020).