Models with verbally enunciated explanations: Towards safe, accountable, and trustworthy artificial intelligence

Models with Verbally Enunciated Explanations: Towards Safe, Accountable, and Trustworthy Artificial Intelligence

Mattias Wahde[a]
Chalmers University of Technology, SE-412 96 Gothenburg, Sweden

Keywords: Artificial Intelligence, Interpretability, Accountability and Safety.

Abstract: In this position paper, we propose a new approach to artificial intelligence (AI), involving systems, abbreviated MOVEEs, that are capable of generating a verbally enunciated explanation of their actions, such that the explanation is also correct *by construction*. The possibility of obtaining a human-understandable, verbal explanation of any action or decision taken by an AI system is highly desirable, and is becoming increasingly important at this time when many AI systems operate as inscrutable black boxes. We describe the desirable properties of the proposed systems, contrasting them with existing AI approaches. We also discuss limitations and possible applications. While the discussion is mostly held in general terms, we also provide a specific example of a completed system, as well as a few examples of ongoing and future work.

1 INTRODUCTION

Models based on deep neural networks (DNNs) have revolutionized artificial intelligence (AI), giving increased performance in many relevant tasks such as, for example, image interpretation and classification (Gupta et al., 2021), data classification in general (MacDonald et al., 2022), speech recognition (Li et al., 2022), and conversational AI, the latter currently being dominated by large language models (LLMs), such as, for example, ChatGPT and GPT-4 (OpenAI, 2023).

Due to their large size as well as their non-linear, distributed computational nature, DNNs are essentially black boxes. That is, their reasoning is normally not human-understandable. In low-stakes applications, the black-box nature of DNNs is of little concern. However, in high-stakes situations involving, say, healthcare, automated driving, or personal finance, being able to understand *how* an AI-system generated a decision may be of utmost importance, and may also soon be a legal requirement (Bibal et al., 2021). Thus, despite the success black-box AI applications over the last decade or so, there are legitimate reasons for concern when such models are used in high-stakes applications.

There are two main paths available for overcoming the potential problems associated with black-box models: Either attempt to provide post-hoc human-understandable explanations for the decisions taken by a black-box model, or instead use a more transparent (interpretable) type of model in the first place[1], also referred to as a *glass-box model*. At present, given the current dominance of DNN-based models in AI, the vast majority of research in this field is geared towards the first of those two options (referred to as explainable AI), even though research is also being conducted on interpretable models (Rudin, 2019).

While DNNs certainly are black boxes, it is *not* so that the decision-making in all supposedly interpretable models is easy to decipher. Interpretability is per definition a subjective concept: A system that is clearly interpretable to one person may be very hard to interpret for another (Virgolin et al., 2021). Thus, here we propose a novel approach, referred to as a *model with verbally enunciated explanations* (MOVEE), which is interpretable in *principle*, but is also augmented with the ability to provide (when prompted) a clear, *verbally* enunciated, correct-by-construction explanation of its decision-making, rendering such a system interpretable also in *practice*. The primary aim of this position paper is to propose the idea conceptually and to describe possible applications and limitations.

[a] https://orcid.org/0000-0001-6679-637X

[1] Note that many researchers use the terms *explainable AI* and *interpretable AI* more or less interchangeably. We do not, as discussed in Section 2 below.

Wahde, M.
Models with Verbally Enunciated Explanations: Towards Safe, Accountable, and Trustworthy Artificial Intelligence.
DOI: 10.5220/0012307100003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 101-108
ISBN: 978-989-758-680-4; ISSN: 2184-433X

We start by a description of various types of AI systems (Section 2), and then proceed with a definition of the MOVEE concept in Section 3. In Section 4, we illustrate the idea by means of one complete, and fully tested example, which fulfils most of the criteria of a MOVEE, and therefore acts as a proof of concept. Furthermore, a few additional examples are given in the same section, on a more conceptual and tentative level. This is followed by a discussion in Section 5 and some conclusions in Section 6.

2 AI SYSTEMS: TYPES AND PROPERTIES

There are many kinds of systems (or models; the two terms are used interchangeably here) in the field of AI, including, for example, linear regression models, decision trees, support vector machines, Bayesian networks, systems based on fuzzy logic, as well as various versions of neural networks, including DNNs. In recent years, DNNs have found many uses, in a wide variety of applications, as exemplified above (many other examples exist as well). The DNNs involved in those applications share several features: They are all very large non-linear statistical approximators, with millions or billions of computational elements, and make decisions using a distributed form of computation, drawing upon huge data sets for their training. These so-called foundation models (Zhou et al., 2023) are then typically fine-tuned for use in specific applications, a process that generally requires much less data than the original training.

Typically, a DNN is fed with an input example, for example an image or a set of features pertaining to a classification task, and the network then outputs a probability distribution over the set of possible outputs (classes) available. However, what happens in between, that is, the concerted action of the many huge layers of the DNN, typically remains completely opaque to a human observer, who would be unable to follow the millions or billions of non-linear calculation steps carried out by the DNN.

Now, in many applications, all that matters is the accuracy of the output, rather than the possibility (or lack thereof) of interpreting *how* the DNN arrived at its decision. This is especially true in *low-stakes* applications, such as, for example, restaurant recommendations, movie reviews, automated selection of music tunes, the action of characters in a (casual) computer game, AI-generated art, and so on, where an occasional error has little or no serious impact on any user. Moreover, in conversational AI, the LLMs that were recently publicly released, such as ChatGPT, are excellent tools for generating, for example, a draft text that does not require exact factual correctness (see also below).

However, there are also *high-stakes* applications, where an error may have severe impact on the health or well-being (physical, mental, financial, and so on) of various stakeholders, particularly the users of the system, but also the developers. Such applications include, for example, credit scoring, automated driving, recidivism prediction, and many health-related applications (e.g., classification of MRI scans or other medical images).

Turning to conversational AI, one also finds many high-stakes applications: Whereas a black-box, LLM-based chatbot can perhaps be entrusted with a casual conversation with a patient, using it as a counsellor or in any other situation where it is supposed to give medical advice (unsupervised) would be fraught with danger (Daws, 2020). In fact, ChatGPT has already been extensively evaluated in a variety of contexts, such as medicine (Vaishya et al., 2023), law (Choi et al., 2023), scientific writing, and so on, many times with very impressive results, but often also with catastrophic failures (Borji, 2023), for example its propensity to cite non-existing papers in scientific writing (Tyson, 2023).

Thus, in addition to the advantages that DNNs bring, there are also several disadvantages, the most important being their black-box nature. This is manifested in various ways, one of them being what one could call a lack of common sense, where DNNs sometimes make completely unexpected catastrophic errors; see, e.g., (Eykholt et al., 2018). In these situations, the main problem is perhaps not the error itself: Any AI system (and indeed any human) makes errors from time to time. The problem is instead that the black-box nature of DNNs makes it difficult to ascertain that such errors will not occur again, in critical situations. Once identified, a specific error can perhaps be removed by further training, but any number of other, similar errors may still lurk in the opaque interior of the black box.

Moreover, the sheer size of the data sets involved in the training of many DNNs (e.g., the foundation models mentioned above) implies that it is nearly impossible to curate the data sets before they are used in DNN training, meaning that the training data sets may (and often do) contain unwanted biases, e.g., sexist, racial, or other biases, which can then be perpetuated by being incorporated in the vast interior of the DNN (Bender et al., 2021).

As mentioned in Section 1, there are two main approaches for dealing with the problems outlined above. The first approach is so-called *explainable*

AI (Angelov et al., 2021), where one attempts to provide post-hoc, human-understandable explanations for the decisions taken by black-box models, primarily DNN-based ones. A diverse set of methods has been defined within this framework, involving techniques such as saliency maps, LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and others. In many cases, explainability involves the use of a secondary, simpler model that is supposed to approximate the primary (black-box) model and to provide an explanation for its actions. While that is a laudable aim, the explanations thus obtained are typically only partial, sometimes contradictory (Krishna et al., 2022), and sometimes unreliable (Slack et al., 2020). In many cases it is unclear whether explainability really explains anything at all (Rudin, 2019).

One may further argue that if indeed a simpler model can do the job, why does one even need the primary model? Either the simpler model *can* approximate the primary one with high degree of accuracy, in which case the primary model would not be needed, or else the simpler model *cannot* accurately represent the primary one, in which case the use of the secondary model is fraught with danger and the explanations that it provides are not likely to be very useful or accurate, in many cases.

An alternative approach would be to avoid black-box models altogether in applications involving high-stakes decisions, as advocated by (Rudin, 2019), and instead use *interpretable* methods. Here it is important to pause for a moment to clarify and contrast the two terms *explainability*, on the one hand, and *interpretability* on the other, since many authors use these terms more or less interchangeably.

In our view this is unfortunate, since each term has an important role to play and, at least by our definition below, the two terms pertain to different classes of systems. We (and others) define *explainable AI* as the set of methods and processes that aim to explain various aspects of black-box models, especially DNNs, mainly on a post-hoc basis. By contrast, we define *interpretable AI* the use of glass-box systems that consist of human-interpretable primitives (components) such as, for *example*, if-then-else rules. Examples of such systems include decision trees, linear regression models, (some) systems based on fuzzy logic, and so on, as well as modified and augmented versions of those systems (Wahde et al., 2023).

It should be noted that, in the specific case of image recognition, there are also systems that make use of interpretable prototypes (Chen et al., 2019; Angelov et al., 2021) that provide a sort of interpetability for the DNNs in question. However, it is not clear how such approaches would generalize to the case of language processing, for example.

We also remark that, while black box (DNN) models have improved performance strongly in many AI tasks, it is *not* so that such models always outperform interpretable models, as exemplified by (Rudin and Radin, 2019). Moreover, in cases where DNN-based models are compared with interpretable ones, the comparison often involves the most recent state-of-the-art DNN versus standard, off-the-shelf versions of interpretable models, a comparison that the DNN generally wins hands down; by contrast, as exemplified in (Wahde et al., 2023), if some effort is applied in order to improve and fine-tune also the interpretable models, the performance gap can be reduced significantly, and perhaps even eliminated, at least for some tasks. In comparing black-box models and interpretable ones, it is also not necessarily *only* performance (accuracy) that matters: Even if a black-box model slightly outperforms an interpretable model, the latter could still be the better choice, taking transparency and accountability into account, as is generally required in applications involving high-stakes decision-making.

However, even a supposedly interpretable model may not always be easy to understand (Angelov et al., 2021; Virgolin et al., 2021), given that the process of understanding a decision or a statement is a subjective one. Furthermore, even if a system consists of components that are easily interpretable in principle, the overall interpretability of the system as a whole may be significantly reduced if, say, the number of components is large or if there are many decision variables. Thus, one may argue, as indeed we do here, that an AI system should ideally be able to provide a clear verbal explanation of its reasoning, such that the explanation is correct *by construction*.

The latter condition is crucial for the concept to be meaningful: While an LLM-based chatbot will happily provide a sequence of words when asked for an explanation of an earlier statement, there is no guarantee that the explanation (or the original statement, for that matter) is *correct*, as such systems are prone to embarking on incoherent rants, referred to as hallucinations (Zhang et al., 2023), with little factual correctness, a problem that can be alleviated by means of so-called *retrieval augmented generation* (Lewis et al., 2020), but probably not eliminated altogether. By contrast, the MOVEEs proposed here will, by construction, provide only factually correct verbal explanations of their reasoning, albeit perhaps less eloquently than an LLM-based chatbot.

3 THE MOVEE APPROACH

Here we propose a novel approach, involving models augmented with the ability to provide (if prompted) a clear, *verbal* explanation of their decision-making. These models, for which the acronym MOVEE is used (as introduced above), consist of distinct components, each of which has the capability of generating its own verbal explanation that, moreover, is correct by construction: It simply describes what the component does, without approximation. A simple example of such a component is one that sorts a list of elements, in which case the explanation involves a static part (*I sorted the list of* ...) and a dynamic, context-dependent part, involving a specification of the kind of elements contained in the list; see also Figure 1 and the discussion in Section 5.

By presenting, in sequence, the partial explanations obtained from each component, an overall explanation for the entire system (or, rather, its decision-making) can be generated. The definition of the MOVEE concept is relevant for systems that mostly process information sequentially, rather than in a parallel, distributed, and non-linear fashion as in DNNs, and where each component is sufficiently high-level so that an explanation makes sense to a human user.

We hasten to add that defining such systems may be difficult or even impossible in many cases. The primary aim of this position paper is instead to propose the idea conceptually and to describe, by means of the examples below, the advantages of the MOVEE concept in those cases where such systems can reasonably be implemented and applied.

Another important issue concerns *learning* in MOVEEs. In the first example below, the agent was generated by hand-coding. However, in current work, an automated learning approach is being implemented, using a form of symbolic regression combined with evolutionary algorithms, making it possible to apply a data-driven approach, while maintaining all the relevant aspects of the MOVEE.

4 EXAMPLES

This section exemplifies the MOVEE concept. First, an existing implementation is described. Next, some potential future applications are described.

4.1 An Implemented Example

In (Wahde and Virgolin, 2023), a system was implemented that exhibits most of the features defining a MOVEE, even though the MOVEE concept itself was not introduced at the time. The system in question is a dialogue manager for task-oriented conversational agents, i.e., systems that, unlike the currently popular LLM-based chatbots, are intended for conversations with high precision, over a limited set of tasks. Thus, such systems can be applied in high-stakes interactions where the factual correctness of the agent's answers is more important both than its human likeness and its ability to conduct a conversation over almost any topic (as is instead possible with systems such as ChatGPT). In these cases, it is likely that, from time to time, the user will want to have a clear, step-by-step, verbal explanation of a statement, response, or suggestion offered by the agent.

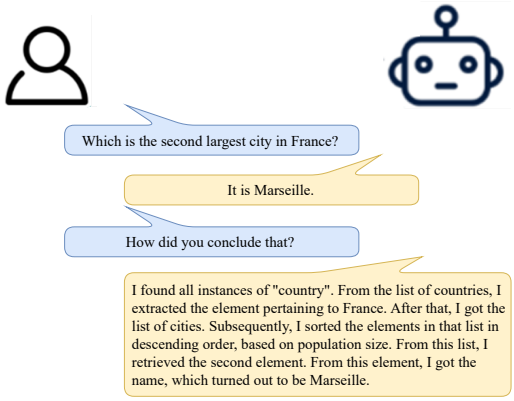

Figure 1: A simple example of an automatically generated explanation by a task-oriented agent based on DAISY (Wahde and Virgolin, 2023).

The conversational system (called DAISY) described by (Wahde and Virgolin, 2023) is capable of providing such an explanation, if prompted by the user. The so-called cognitive processing in DAISY, i.e., the part where the agent determines what to say (usually in response to user input), is structured as a sequence of generic elementary operations (referred to as *cognitive actions*), each associated with a dynamically formulated verbal explanation, which takes into account the variables associated with the user's input and the agent's own output, as exemplified in Figure 1. The explanation for each operation is correct by construction, as it simply amounts to a verbal enunciation of what the component actually does, without any approximation. Thus, whenever the agent formulates its output, a full explanation of the process is automatically generated as a by-product, ready to be presented to the user upon request.

However, the implementation in (Wahde and Virgolin, 2023) was a preliminary one, and it was tested and illustrated as a proof-of-concept in rather basic conversations on, for example, hotel reservations or geography; see Figure 1.

4.2 Examples of Potential Applications

Here, two potential applications are described, in some detail but also in a tentative and preliminary way, given that those applications have yet to be implemented. There are, of course, also other potential applications, which are briefly discussed in Section 5.

4.2.1 Automated Driving

The automotive industry is undergoing a transformation worldwide, involving two main trends, namely electrification and automated driving (Parekh et al., 2022). Already now, automated driving occurs in controlled environments (such as work yards and mines) and also, to a lesser degree, in normal traffic.

In this major transformation, safety is a paramount concern, especially in the transition phase where human driving is gradually phased out. Yet, much work in this field is centered on the use of black box models, such as DNNs. Some even envision end-to-end approaches, in which a DNN handles every step from perception (via onboard sensors, such as cameras and lidars) to action (acceleration and steering), the rationale being that, even though such systems are black boxes and sometimes fail in unexpected and unpredictable ways, they may still reduce the number of road accidents, bearing in mind that human drivers also fail in similar ways, from time to time. However, the predicted safety improvement is far from certain. For example, a DNN that recognizes road signs (and then acts accordingly) may achieve near-perfect performance over its test set, yet may fail spectacularly when encountering road signs with rather small perturbations (such as an added sticker) that would not fool a human driver (Eykholt et al., 2018).

In any case, this is a field where especially the *developers* of automated functionality would benefit from an approach centered on a (yet-to-be-developed) MOVEE. This approach would not exclude the possibility of using DNNs as components, for example in image recognition. However, in a MOVEE-based approach the system would not operate in an end-to-end fashion, but would instead be divided into modules. As a minimum, there would be one perception module, one planning module, and one module for taking action, each with the ability to provide explanations.

As a specific example, consider a case where a DNN-based image recognition system mistakenly interprets a stop sign covered with a sticker (or defaced in a similar manner) as something else, such as a speed limit sign (Eykholt et al., 2018). Assume also that the system is arranged as a MOVEE, as described above. In this case, when the vehicle is tested in a simulated environment, such as a high-fidelity simulator of the kind typically used in the vehicle industry, the following conversation might ensue, either in written or spoken form:

Developer: *You missed the stop sign!*
Vehicle: *I did not see a stop sign.*
Developer: *How did you interpret the most recent road sign that you passed?*
Vehicle: *It was a speed limit: 50 km/h*

At this point, the developer can pinpoint the error, stop the simulation, and take corrective action, either improving the DNN by further training or in some other way altogether. Without a MOVEE, the developer would have to sift through the program code and its output logs, to find the reason for the error[2].

In addition to developers, the *users* (passengers) of automated vehicles could also benefit from a MOVEE-based approach. For example, if the vehicle does something unexpected, the user may wish to obtain a reassuring explanation. As a specific example, consider the case of fuel-consumption minimization, where a vehicle, driving over a hilly road, modulates its speed in order to minimize fuel usage, an application that has been considered for heavy-duty trucks (Torabi and Wahde, 2017) but which could also be generalized to cover passenger vehicles. In some cases, the acceleration or deceleration may not always make immediate sense to the occupants of the vehicle. Thus, with a MOVEE-based approach, the following conversation might take place:

User: *Why did you just accelerate?*
Vehicle: *I accelerated in order to gain some speed before the uphill climb that we will encounter in 2 km. This will save some fuel.*

At this stage, being implemented in a production vehicle, the system should already operate as intended; it should not be the job of the passenger to debug its functionality, but she or he may nevertheless want an explanation for the actions taken by the vehicle.

4.2.2 Safety at Sea

In parallel with the trend towards automated driving on roads, a similar transformation is taking place in the maritime environment (Veitch and Alsos, 2022). This development has not yet gone as far as in the case of road vehicles, but it is likely to follow a similar trajectory in the years ahead. One may argue that

[2]Of course, many computer programs can provide error messages, but they are not always easy to interpret, unlike the verbal explanations exemplified above.

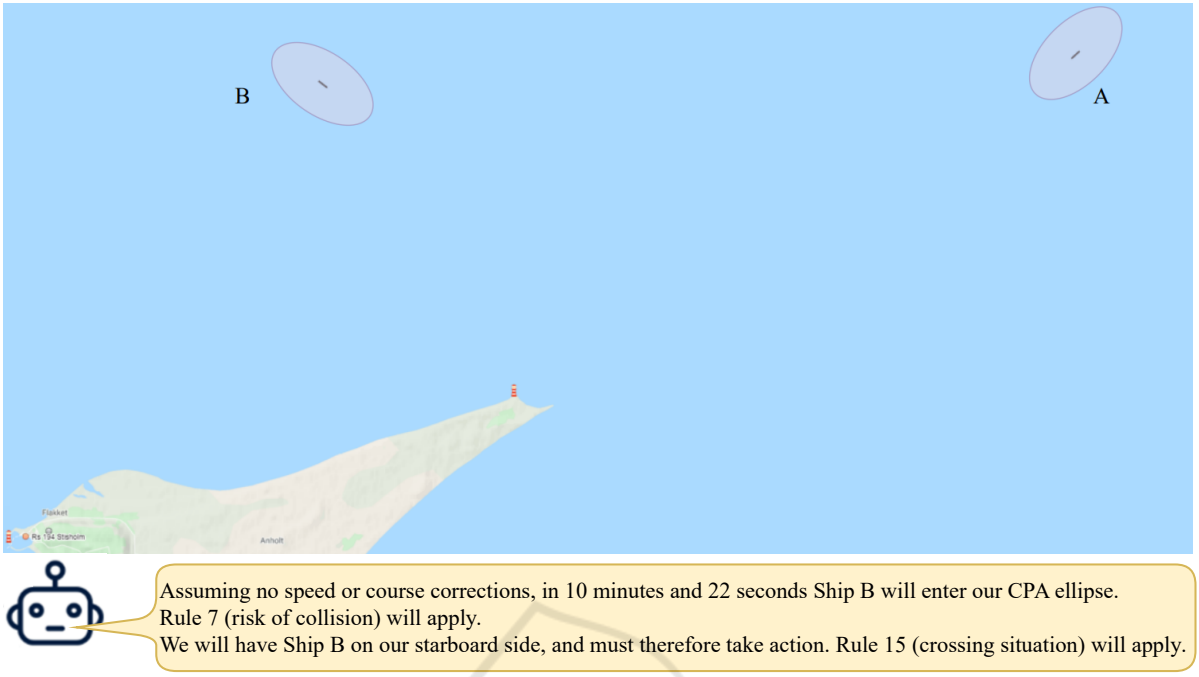

Figure 2: An example of an explanation given by a MOVEE in response to a request (not shown) by the captain of ship A.

the case of maritime applications is every bit as important as that of road vehicles, especially since (i) at sea, vessels may approach each other from any angle, making it a very complex environment, especially in narrow passages in the vicinity of large cities, where larger vessels may share the environment with many smaller vessels, such as ferries and recreational boats; (ii) the potential effect of collisions can be even more severe than for road vehicles, for example in the case of a collision between two oil tankers.

In the maritime environment, vessels are required to follow the Convention on the International Regulations for Preventing Collisions at Sea (COLREGs) [3] that determine, among other things, the actions (if any) that a vessel should undertake when encountering other vessels. Applying these rules is not always easy, especially in cases involving three or more vessels in close vicinity of each other. This is thus another case where a MOVEE may be useful, perhaps more as a decision-support system for the captain of a vessel than for automated decision-making, even though a transition towards the latter may become a reality eventually. A basic example is shown in Figure 2, where the captain of ship A has asked a MOVEE to explain what needs to be done as ship A encounters another ship (B). In current work, we are implementing a MOVEE for handling situations such as the one shown in Figure 2.

[3]https://www.imo.org/en/About/Conventions/Pages/COLREG.aspx

5 DISCUSSION

In addition to the examples given above, there are many other applications where systems that provide correct-by-construction, verbal explanations can be useful. For example, one may argue that interactive systems in healthcare and elderly care (MacDonald et al., 2022), as well as risk-management systems that make decisions on whether or not to grant a loan (John-Mathews, 2022), should be able to provide a clear verbal explanation of their decision-making.

However, implementing such systems may be challenging in many cases. First of all, a requirement for applying a MOVEE, as defined here, is that the decision-making should be divisible into a sequence of elementary operations, something that requires quite a different approach than the end-to-end style processing that occurs in (some) DNN-based applications, and may not always be feasible.

Such a division was natural in the implemented example given in Section 4.1, where the required steps (cognitive actions) involved sequences of simple operations, like finding elements (in memory) that fulfilled certain criteria, sorting lists of elements, extracting attributes from elements, as well as mathematical set operations (unions, intersections, and so

on). While many of those operations are likely to be useful in other cases as well, extensions will be required when considering other applications. For example, in an automated driving context (see Section 4.2.1), the set of operations will also have to include those that are relevant for driving, such as accelerating, steering, braking, processing traffic sign information, and so on, and the system as a whole must consist of sequences of such operations. Writing these operations may not always be easy, especially since their associated verbal explanations are not static but depend on variables pertaining to the situation at hand; see also the conversation in Figure 1, where the explanation involves information both from the user's question and the agent's response.

Second, even if each elementary operation is capable, by construction, of providing a correct explanation of its actions, in order to be useful a MOVEE must also make sure that the complete explanation (presented to the user) is brief enough to be clear. For example, some decision-making may involve loops, where a given operation is repeated a number of times. In such cases, it would not make sense to present every iteration in the loop step-by-step (as in: *I did action A, then incremented the counter by one, then did A again, then incremented the counter by one* and so on) but rather to summarize the explanation (as in: *I iterated action A ten times.*).

Third, one may also wish to make the explanations as natural as possible. For example, while the explanation in Figure 1 is abundantly clear and, due to recent improvements, less robotic than the original explanation presented in (Wahde and Virgolin, 2023), it is still not completely natural, compared to the explanation that a human may give in the same situation. Thus, a MOVEE should also strive to make its explanations as condensed as possible, without loss of clarity. However, it should be noted that the naturalness of the explanations is (in our view) less important than their correctness. Moreover the interactions between a MOVEE and its users is typically rather elementary and entirely focused on providing explanations of the decision-making; unlike a chatbot, a MOVEE is not intended for general discussions on any topic.

Thus, while the implementation of a MOVEE for a given application may encounter plenty of difficulties, there are also many benefits associated with the possibility of obtaining clear, verbal, correct-by-construction explanations for the actions planned, suggested, or taken by an AI system, perhaps especially for the developers of such systems. We also remark that, even though a MOVEE will, per definition, consist of a sequence of well-defined and separate elementary operations, the use of black boxes *within* such a system is not excluded, as exemplified in connection with the first conversation in Section 4.2.1.

Finally, we also note that MOVEEs may have many benefits regarding legal requirements on AI systems. For example, in cases where a MOVEE controls an automated vehicle, one may add a requirement that the system should log all the explanations (of its decision-making) so that, in case of an incident or accident, the log can be made available to various stakeholders, such as the police, insurance companies, the vehicle manufacturer, and so on.

6 CONCLUSIONS

We have proposed an approach involving AI systems that consist of sequences of elementary operations, such that each operation is associated with a verbally enunciated explanation, which is correct by construction, paving the way for safe and accountable uses of AI. We have discussed a proof-of-concept implementation of such a system, and also proposed additional applications while, at the same time, acknowledging that such systems may not be suited for all applications. We conclude, however, that the benefits of being able to obtain a clear, verbal explanation for the decisions taken by an AI system should, in many cases, easily offset the difficulties associated with defining and implementing such a system, not least bearing in mind current and upcoming legal requirements.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Marco Virgolin for many discussions on the topic of interpretability.

REFERENCES

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bibal, A., Lognoul, M., De Streel, A., and Frénay, B. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29:149–169.

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494*.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Choi, J. H., Hickman, K. E., Monahan, A., and Schwarcz, D. (2023). ChatGPT goes to law school. *Available at SSRN*.

Daws, R. (2020). Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. AI News, https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/. Accessed May 2021.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634.

Gupta, A., Anpalagan, A., Guan, L., and Khwaja, A. S. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10:100057.

John-Mathews, J.-M. (2022). Some critical and ethical perspectives on the empirical turn of ai interpretability. *Technological Forecasting and Social Change*, 174:121209.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Li, J. et al. (2022). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

MacDonald, S., Steven, K., and Trzaskowski, M. (2022). Interpretable ai in healthcare: Enhancing fairness, safety, and trust. In *Artificial Intelligence in Medicine: Applications, Limitations and Future Directions*, pages 241–258. Springer.

OpenAI (2023). GPT-4 technical report (arxiv: 2303.08774).

Parekh, D., Poddar, N., Rajpurkar, A., Chahal, M., Kumar, N., Joshi, G. P., and Cho, W. (2022). A review on autonomous vehicles: Progress, methods and challenges. *Electronics*, 11(14):2162.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Rudin, C. and Radin, J. (2019). Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2):1–9.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Torabi, S. and Wahde, M. (2017). Fuel consumption optimization of heavy-duty vehicles using genetic algorithms. In *2017 IEEE Congress on Evolutionary Computation (CEC)*, pages 29–36. IEEE.

Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education*, 100(8):3098–3101.

Vaishya, R., Misra, A., and Vaish, A. (2023). ChatGPT: Is this version good for healthcare and research? *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 17(4):102744.

Veitch, E. and Alsos, O. A. (2022). A systematic review of human-ai interaction in autonomous ship systems. *Safety Science*, 152:105778.

Virgolin, M., De Lorenzo, A., Randone, F., Medvet, E., and Wahde, M. (2021). Model learning with personalized interpretability estimation (ML-PIE). In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1355–1364.

Wahde, M., Della Vedova, M. L., Virgolin, M., and Suvanto, M. (2023). An interpretable method for automated classification of spoken transcripts and written text. *Evolutionary Intelligence*, pages 1–13.

Wahde, M. and Virgolin, M. (2023). Daisy: An implementation of five core principles for transparent and accountable conversational ai. *International Journal of Human–Computer Interaction*, 39(9):1856–1873.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. (2023). Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.