Noname manuscript No.
(will be inserted by the editor)

An Exploration of Combinatorial Testing-based Approaches to Fault Localization for Explainable AI

Ludwig Kampel · Dimitris E. Simos · D. Richard Kuhn · Raghu N. Kacker

Received: date / Accepted: date

Abstract We briefly review properties of explainable AI proposed by various researchers. We take a structural approach to the problem of explainable AI and examine the feasibility of these aspects by extending them where appropriate. Afterwards, we review combinatorial methods for explainable AI which are based on combinatorial testing-based approaches to fault localization. Last, we view the combinatorial methods for explainable AI through the lens provided by the properties of explainable AI that are elaborated in this work. We pose resulting research questions that need to be answered and point towards possible solutions, which involve a hypothesis about a potential parallel between software testing, human cognition and brain capacity.

Keywords AI · explainable AI · combinatorial testing · fault localization

1 Introduction

Artificial intelligence (AI) systems have improved rapidly, with their performance now surpassing human abilities in many or most domains, especially vision and image recognition applications, but also in more safety-critical tasks

SBA Research
Vienna A-1040, Austria
E-mail: lkampel@sba-research.org

SBA Research
Vienna A-1040, Austria
E-mail: dsimos@sba-research.org

National Institute of Standards & Technology
Gaithersburg, MD, USA
E-mail: kuhn@nist.gov

National Institute of Standards & Technology
Gaithersburg, MD, USA
E-mail: raghu.kacker@nist.gov

such as autonomous vehicles [13], [26]. The increase in numbers of AI applications, and their integration into everyday life, has created a public demand for understanding the behavior and decisions of AI systems. This demand has led to the research field of explainable AI (XAI) that has the goal of making AI systems or their decision-making humanly understandable.

Many researchers around the globe are actively working on bringing the "X" to the "AI". The approaches are as diverse as there are AI systems, reaching from self explaining systems, externally explained systems, global explainable AI algorithms (e.g. SHAP [14]) and per-decision explainable AI algorithms (e.g. LIME [20]). These and other developments on XAI can be found in a recent survey [1].

In this paper, we briefly review recently introduced properties for XAI in Section 2 and examine their feasibility in Section 3. Thereafter, we review combinatorial methods for XAI, which are based on combinatorial testing-based approaches to fault localization (CT-FLA), in Section 4. Last, we reflect on these in Section 5, where we hypothesize about potential parallels amongst computer science (software testing) and psychology (human cognition and brain capacity).

2 Properties of XAI

First, we want to have a look at AI and XAI. It is presently difficult to give a generally accepted and detailed definition of AI. There exists a plurality of approaches on how to define AI, for example being centred around human performance or rather around thought processes and reasoning, see [21] and references therein. On top of that the understanding of AI is controversial and may also change over time - just think of Deep Blue defeating Garry Kasparov [10].

However an AI system may appear, today we see various realizations of AI, the majority using Bayesian networks, deep learning or symbolic approaches.

We point out these different understandings and realizations of AI, because we believe that it has a heavy impact on the explanations that we can expect to get or produce for the respective AI. A full understanding of XAI will require extensive human factors research, but the topic of explanation has been studied in psychology for decades, and much of this work can be adapted to the problem of explainability in AI [5], [17], [24], [27]. One useful categorization of the human factors aspects essential for explanations, whether machine or human-generated, is provided in Ehsan et al. [6], who studied how psychological research on understanding can be applied to machine-generated explanations. They consider the following dimensions in order to rate the endorsement of explanations:

- Confidence: This rationale makes me confident in the character's ability to perform it's task.
- Human-likeness: This rationale looks like it was made by a human.
- Adequate justification: This rationale adequately justifies the action taken.

Ehsan et al. [6]	Proposed Properties
Understandability	Existence
Human-likeness	Clarity
Adequate justification	Adequate justification
Confidence	Trust

Table 1 Mapping between the four *dimensions* given in Ehsan et al. [6] and the properties considered in this work.

– Understandability: This rationale helped me understand why the agent behaved as it did

For the purpose of this work, we adopt these properties from [6] and modify them as follows:

1. *Existence*: For each output an explanation is provided that helps to understand why this output was generated.
2. *Clarity*: The explanations can be understood comprehensible by humans/users.
3. *Adequate justification*: The explanations adequately justifies the system's output or its process for generating the output.
4. *Trust*: In the system to accurately generate the output based on a description of events and its environment.

Together, the first three properties aim for ensuring that generated explanations are *plausible* to humans. We want to mention that there is a fundamental difference between *the output* that a system generates and *how* (the) output is generated. We therefore need to be precise with regards to our demands to XAI: do we want an explanation for the output of an AI-based system, or an explanation of the underlying process? We will elaborate on this and similar questions in the rest of the paper.

3 Remarks related to the Development of XAI

We first want to propose a classification of XAI as an adaptation and addendum to the works cited above. We believe it is worthwhile to explicitly mention distinctive features of XAI as these will help to reason about it, especially with regard to what we can expect and demand from explanations.

3.1 Classifications of XAI

For example, the general characteristics of plausible explanations (clarity and adequate justification) must take into account that explanations may need to be varied for different users, who have different levels of knowledge and expectation. However, as Hilton [8] writes: "*The verb 'to explain' is a three-place predicate: Someone explains something to someone*". Thus, not only the receiver of the explanation is crucial in this differentiation, but also the matter that is being explained, i.e. the AI system, its input and output. In the

following, we give some dimensions and criteria along which we can differentiate XAI. We do not claim completeness or even correctness and leave it to future investigations to revise or improve them. However, we believe it is an important step in understanding where specific explanations can be applied (to which AI) and for whom they are produced (the user or receiver of the explanation). In the following discussions, we refer to these two entities as the *human* and the *AI*.

Who receives the explanation? It is generally accepted that we need to distinguish explanations according to who asks for them, with the expert versus non-expert example being the most prominent one. One possible way is to differentiate a number of groups that differ in the quality and quantity of information they demand or expect.

- *Non-experts*: They want to know the key reasons why a specific output is produced - details are not needed, or even desired.
- *Experts*: They want to have detailed reasons why a specific output is produced, however these need not or should not be dependent on the AIs implementation.
- *Developers*: They want many details that can or should be implementation dependent, in order for the explanation to guide his or her debugging or development process.
- *Algorithms*: They require details in a machine readable format, where requirements can be formally specified.

We can see in the last group, that the *human* (i.e. the explanation receiving part) can also be an AI, or an algorithm more general, e.g. an algorithm that is rating the quality of explanations.

Who gives the explanation? We can differentiate XAI systems according to where the explanation comes from:

- *Self-explainable models/systems*: These are AI systems that provide the needed explanation themselves; these can be systems where the underlying algorithm itself represents the explanation, e.g. AI systems based on decision trees or ones that provide explanations without giving algorithmic details, such as class activation mappings [28].
- *External explanation models/systems*: In this case the explanation of the AI's output is produced by a separate algorithm.

What is being explained? The subject of explanation can be differentiated in manifold ways:

- *Decision vs decision process*: Is the output explained or the process that leads to the AI system's output?
- *Global explanations vs per-decision explanations*: Is a single output explained or a set of outputs?
- *Kind of AI system*: For example a classifier/decider, an AI system performing tasks like driving a car or an automated theorem prover.

- *What is the input to the AI*: The explanation for an output has to relate to the input (black-box case). An AI algorithm also starts with the input, hence an explanation for the AI relates to the input.
- *Black-box vs white-box model*: Is the internal mechanism of the AI system accessible or not?

Again, we do not claim that this list of categories is complete. Further, we consider also that the above categories are not necessarily excluding each other and may very well be mutually influenced as there exist some causalities between them. For example, when we ask for an explanation of the decision process of an AI system, then the system is generally a white-box model, as we need to have access to its internal working mechanism in order to explain it. Furthermore, self-explaining systems already provide such insights. Another example is that a global explanation could be used to generate a per-decision explanation or any black-box approach can be also applied to a white-box AI system.

3.2 Solution Processes to NP-Complete Problems may be too Difficult to Explain

In this section we focus on the comparison of explaining AI output generation processes versus explaining outputs of AI systems.

The "adequate justification" component of explanations can encompass *how* the system came to its conclusion, and the system's "output" itself. Of course, an explanation for an output process can yield an explanation for the resulting output itself, but it can be significantly more difficult to explain an output process compared to an output, as set out below.

From basic computational complexity notions we know that *finding a solution* to a problem and *verifying* a solution as such can result in significant difference of computational effort. The well known P versus NP problem, includes the question whether the solution to an NP-complete problem can be found as easily as it can be verified. Let us assume that P $\neq$ NP, how does this influence the explainability of AI? For NP-complete problems the length of the solution derivation would not be bound by any polynomial function in the length of the input, while for the solution verification there would be such a bound.

This analogy to computational complexity is not too far off compared to the explainability of AI: Assume we have developed an AI system that (optimally) solves routing problems, such as TSP. Asking for a *meaningful* and *accurate explanation* of the *decision process*, means asking for an understandable (somewhat short) and correct explanation of the lengthy solution process to an NP-complete problem [19]. To give another example, lets consider a constraint satisfaction problem (CSP) solver as an instance of an AI system. The decision process to a query itself can be extremely lengthy, but once a decision is made, it can be verified fairly easy in some cases. For example, provided the correct formulation, we can query a CSP solver whether a map can be colored

with only three different colors, which is an NP-complete problem [23]. The derivation process itself can be extremely lengthy and difficult to follow, but when a solution is found and the answer is 'yes', then this can be easily explained by providing the three-coloring of the map, something very accessible to human beings. These remarks beg the question whether there is an analogue notion to NP-completeness in explainable AI, i.e. *a solution process that requires significantly more effort to be explained compared to the explanation of the solution.*

Clarifying if solution processes and solutions to NP-complete problems are an example of this is one way to address this question. Solution processes can appear in the form of a decision or a search algorithm. A related research question is: Does the *length* of a solution process, here we mean the formal length of the derivation, make an explanation more difficult to generate?

3.3 Explaining AI Systems and Distinguishers between Computers and Humans

The authors of [22] consider computers and humans together for solving problems, and as a result they propose a theory towards AI-completeness. In particular, they define *human-assisted Turing machines* in order to examine computations that can be split between humans and computers. Formally this is done by extending Turing machines with *human oracles*, see [22] for details. By putting computers and humans in the same context the computational complexity is generalized via capturing how often a human oracle is called. By defining an appropriate measure for algorithmic complexity, this work presents the means to formally reason about questions such as "How much human interaction is required to solve a specific problem (more or less efficient)?". As the authors mention in [22], this investigation can also lead to a set of problems that one can use to distinguish computers from humans. There already exist work on such problems that are easy to solve for humans, but difficult for computers, e.g. CAPTCHA problems given in [2]. Such CAPTCHA problems provide a distinguisher to tell humans and computers apart, which can be used practically as a Turing test.

Based on the above, it is intriguing to ask the following question: Is the problem of producing humanly understandable explanations for the output (generation processes) of AI systems solvable without human input? We are aware that this question most likely cannot be answered generally but needs to be examined on a case-by-case basis, especially considering the problem that is solved by an AI system.

4 CT-FLA for XAI

The work in [12] presents combinatorial methods that are inspired by ideas and methods from CT-FLA for explaining classifications and decisions made

by AI systems. The justification of the assignment of an object to a specific class is given by the identification of feature combinations that are present in the object and in members of the assigned class, while being absent (or rare) in objects of other classes. A related black-box approach for per-decision explanations of AI systems is presented in [20].

We briefly outline the connection between CT-FLA and classification systems. The idea of CT is to test a system under test (SUT) against misbehaviour caused by *interactions* of its input parameters based on optimized test sets that *cover* all demanded interactions. Applying CT requires an input parameter model (IPM) of the SUT, which consists of parameters with respective values. There is empirical evidence [11] that software faults are triggered by input parameter-value interactions up to a certain *strength*. The fundamental idea of CT-FLA is to automatically recover the fault causing interactions provided only the test set and a pass/fail assignment to them. There exist statistical and deterministic methods [4], [7] for CT-FLA. For more details on CT-FLA see [9], [11].

To apply CT-FLA to explain classifications generated by AI systems, we need to correspond the notions of AI classification systems with the respective ones related to CT:

- The *input* to a classification system as the equivalent to a *test vector* in CT,
- The *assigned class* to an object as the equivalent of the *resulting pass/fail-assignment* of the test vector execution,
- The *identified feature combination* as the equivalent of the *failure inducing parameter-value interaction*.

Provided this mapping of notions, in order to search for an explanation why an AI system classifies a specific object o to a class c, we simply map class c to *fail* and all members of c to the failing tests. Once a failure inducing interaction of the comprised test set is identified, we have found a feature combination that is present in the members of class c while not present in any other class.

We visualize this mapping in Table 2 where we present a part of the example given in [12], featuring a database of animals with attributes. On the right hand side of Table 2 we see a snippet of a database with animal records; due to space limitations only five (of originally 16) attributes are shown. The classification of Testudo as a reptile is explained by the feature combination triplet (non-aquatic,toothless,four-legged) that is unique to reptiles and present in Testudo. This triplet represents a counterfactual explanation: if Testudo had 6 legs, it would be an insect. This concludes the review of [12] which shows how methods for CT-FLA can be used to produce counterfactual explanations for AI classification systems.

As an additional remark, we want to mention the notion of *minimal failure inducing t-way interactions* [18], which are parameter-value combinations that when being reduced or deviated, do not necessarily cause tests to fail any more. Translated to the field of XAI, these allow the derivation of *counterfac-*

test	p_1	p_2	p_3	p_4	p_5	result		class	hair	aquatic	egg-laying	toothed	nlegs	object
$t_1 =$	0	0	1	0	3	pass	$\longleftrightarrow$	insect	no	no	yes	no	6	Mantis
$t_2 =$	0	0	1	0	2	fail	$\longleftrightarrow$	reptile	no	*no*	yes	*no*	4	Testudo (Turtle)
$t_3 =$	0	1	1	0	3	pass	$\longleftrightarrow$	insect	no	yes	yes	no	6	Water scorpion
$t_4 =$	0	1	1	1	1	pass	$\longleftrightarrow$	bird	no	yes	yes	yes	2	Penguin
$t_5 =$	0	0	1	1	2	fail	$\longleftrightarrow$	reptile	no	*no*	*yes*	yes	4	Crocodile
$t_6 =$	0	0	1	0	3	pass	$\longleftrightarrow$	insect	no	no	yes	no	6	dung beetle

Table 2 Analogy between CT (left) and explanations for an AI system which produces classifications (right).

tual cores, which when being modified yield different classifications of the AI algorithm and hence serve as source for counterfactual explanations.

5 Reflection of CT-FLA methods for XAI

Having revisited XAI through the four principles and CT-FLA methods, we want to examine *where* the latter can be applied and to which degree we can apply the combinatorial lever.

Research Question: How can CT-FLA methods for XAI be categorized?

Answer: Processing the described categories proposed in Section 3.1 bottom up, we can categorize combinatorial methods for XAI as:

- *Black-box*: not relying on AI system internals
- *Input*: systems where the input is modelled via an *IPM*
- *Kind of AI system*: *classifier and decision systems*
- *Per-decision explanations*: primarily per-decision
- *Decisions*: are being explained, not decision processes
- *External explanation*: explanation is independent of the AI system's internals and provided by an external source
- *Explanation receiver*: explanations produced are suited for *non-experts*, *experts* and potentially *other algorithms*.

Note that, CT-FLA methods are primarily suited to produce per-decision explanations, but they can also characterize whole classes and thus not only explain an individual object. Thus, in how far counterfactual cores give a global explanation for an AI system is debatable. Devising the required IPM can be straightforward, e.g. when the input is already given as a list of attributes; or can require to additionally model the input space to the AI system.

Now, we consider again the four properties from [6] in conjunction with CT-FLA:

Existence. Translated for CT-FLA methods applied to XAI, especially classifiers, this property requires that for each object that is classified as a member of a specific class, there must be at least one characteristic feature-combination that can be identified for this object and members of its assigned class. Otherwise, the description is clearly inaccurate, and a user cannot trust it, as it has failed to identify any characteristic features.

Research Question: Can decisions of systems that classify input that is modelled via an IPM always be explained via feature-combinations of the input?

Answer: It is up to further investigation whether this question can be answered in the same style as for software faults [11].

Clarity. We outlined how CT-FLA methods provide explanations via feature-combinations.

Research Question: Are explanations generated from CT-FLA for XAI humanly understandable?

Answer: The feature-combinations generated by CT-FLA methods serve as *counterfactual explanations*. We outlined previously (Section 4) how minimal failure inducing t-way interactions can yield counterfactual cores. There exist several studies that suggest that counterfactual explanations suit the human way of casual explanations, see e.g. the work of Hilton [8] and references therein, [15]. Further, some works investigate the role of counterfactual explanations in the realm of XAI [3], [17], [25]. This leads us to consider the following:

Research Question: How complex or lengthy can counterfactual explanations become and still be humanly understandable? Further, is the length independent from the classification process?

Answer. A potential answer to this question can be found in the well known observation, by the psychologist Miller [16], states that the capacity of the human brain in terms of short-term memory is limited to about 7 ± 2 *chunks*, i.e. information units. Such or similar insights might translate to an *upper bound* on the strength of feature-combinations that need to be identified as class characteristic by combinatorial methods, as any feature-combination beyond this upper bound is not easily processable by the human brain. This would be a natural bound for the applicability of CT-FLA methods for XAI, and thus present a psychological equivalent of the empirical study conducted by NIST [11], that suggests that it is (largely) sufficient to consider parameter-value combinations of up to six parameters for combinatorial software testing.

Adequate Justification. One cannot expect that "short" feature-combinations can explain all AI decisions, e.g., automated theorem proving or SAT solving, where the results likely depend on the entire input.

Research Question: Are the explanations produced by CT-FLA methods adequate to explain AI decisions?

Answer: This question could be answered (partly) by a case study, comparing explanations from self-explaining classification system with those generated by CT-FLA methods in order to evaluate the explanations. Such a comparison may reveal cases where CT-FLA methods are suited, and others where they are not suited for generation explanations to AI systems.

Trust. This aspect primarily concerns AI systems, rather than (external) explanation systems, however it raises the following question.

Research Question: Does the absence of a characteristic feature-combination imply an inaccurate action in the given situation?

Answer. This can be reasonably addressed, only once the previous research questions have been addressed, especially we need to know whether decisions within the knowledge limits of the system can lead to characteristic feature-combinations.

6 Conclusion

A number of researchers (see Section 2) have considered the application of psychological research on explanation quality to the problem of XAI. We investigated the applicability of combinatorial methods to XAI considering these general characteristics of explanation quality and formulated open research questions, providing answers where possible. We can only hope that fully answering them, can lead to a further improvement of combinatorial methods and advance XAI.

Acknowledgements

SBA Research (SBA-K1) is a COMET Centre within the framework of COMET - Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.
Disclaimer. Any mention of commercial products in this paper is for information only; it does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST).

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: CAPTCHA: Using hard AI problems for security. In: E. Biham (ed.) Advances in Cryptology — EUROCRYPT 2003, pp. 294–311. Springer Berlin Heidelberg, Berlin, Heidelberg (2003)
3. Artelt, A., Hammer, B.: On the computation of counterfactual explanations–a survey. arXiv preprint arXiv:1911.07749 (2019)
4. Colbourn, C.J., McClary, D.W.: Locating and detecting arrays for interaction faults. Journal of Combinatorial Optimization **15**(1), 17–48 (2008)
5. Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: A survey. In: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), pp. 0210–0215. IEEE (2018)

6. Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., Riedl, M.O.: Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 263–274 (2019)
7. Ghandehari, L.S., Chandrasekaran, J., Lei, Y., Kacker, R., Kuhn, D.R.: BEN: A combinatorial testing-based fault localization tool. In: 2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW), pp. 1–4 (2015)
8. Hilton, D.J.: Conversational processes and causal explanation. Psychological Bulletin **107**(1), 65 (1990)
9. Jayaram, R., Krishnan, R.: Approaches to fault localization in combinatorial testing: A survey. In: S.C. Satapathy, V. Bhateja, S. Das (eds.) Smart Computing and Informatics, pp. 533–540. Springer Singapore, Singapore (2018)
10. Kasparov, G.: Deep thinking: where machine intelligence ends and human creativity begins. Hachette UK (2017)
11. Kuhn, D., Kacker, R., Lei, Y.: Practical combinatorial testing. NIST Special Publication 800-142 (2010)
12. Kuhn, D.R., Kacker, R.N., Lei, Y., Simos, D.E.: Combinatorial methods for explainable AI. In: 2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), pp. 167–170 (2020)
13. Lugano, G.: Virtual assistants and self-driving cars. In: 2017 15th International Conference on ITS Telecommunications (ITST), pp. 1–5 (2017)
14. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
15. Mandel, D.R., Hilton, D.J., Catellani, P.E.: The psychology of counterfactual thinking. Routledge (2005)
16. Miller, G.A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review **63**(2), 81 (1956)
17. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547 (2017)
18. Nie, C., Leung, H.: The minimal failure-causing schema of combinatorial testing. ACM Trans. Softw. Eng. Methodol. **20**(4) (2011)
19. Papadimitriou, C.H.: The euclidean travelling salesman problem is NP-complete. Theoretical Computer Science **4**(3), 237 – 244 (1977)
20. Ribeiro, M.T., Singh, S., Guestrin, C.: "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, KDD '16, p. 11351144. Association for Computing Machinery, New York, NY, USA (2016)
21. Russel, S., Norvig, P.: Artificial intelligence: a modern approach. Pearson Education Limited (2013)
22. Shahaf, D., Amir, E.: Towards a theory of AI completeness. In: AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, pp. 150–155 (2007)
23. Stockmeyer, L.: Planar 3-colorability is polynomial complete. ACM Sigact News **5**(3), 19–25 (1973)
24. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems (2020)
25. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology **31 (2)** (2018)
26. Wotawa, F.: On the importance of system testing for assuring safety of AI systems. In: AISafety@IJCAI (2019)
27. Zhang, Y., Chen, X.: Explainable recommendation: A survey and new perspectives. arXiv preprint arXiv:1804.11192 (2018)
28. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)