



PANEL #4

NICE
June 2023

ComputationWorld 2023 & DataSys 2023

Theme:

Challenges on Explainability of AI-based Decisions



Chair Position

NICE
June 2023

As AI systems become more sophisticated, we face more challenges:

- Increasingly important to understand how these systems arrive at their decisions or predictions
- Achieving explainability is a complex task
 - **Black box** nature of AI models:
 - Lack of transparency poses a challenge when attempting to explain how decisions are reached
 - **Lack of interpretability:**
 - While some parts of the model's decision-making process can be understood, other aspects remain obscure
 - **High dimensionality** and **feature interactions:**
 - Models can operate on high-dimensional data, often with numerous features.
 - Understanding how features interact and contribute to the model's decision can be intricate.
 - Identifying which features are most influential becomes increasingly difficult as the dimensionality of the data increases.



Panos Nasiopoulos



Chair Position

NICE
June 2023

- **Data availability and quality:**
 - Explainability often relies on access to **comprehensive** and **accurate data**
 - **Limited data availability** or **poor data quality** can hinder efforts to understand the reasoning behind AI decisions.
 - **Biases present in the training data** can propagate through the model, leading to biased or unfair decisions that are challenging to explain.
- **Trade-off between performance and explainability**



Panos Nasiopoulos

Addressing these challenges requires interdisciplinary efforts from researchers, policymakers, and industry practitioners.

Developing new techniques for model interpretability, promoting transparency in AI systems, and establishing ethical guidelines for AI deployment are some steps toward addressing the challenges surrounding explainability in AI-based decisions.



CONTRIBUTORS

NICE
June 2023

Moderator

Prof. Dr. Panos Nasiopoulos, University of British Columbia, Canada

Panelists

Prof. Dr. Przemyslaw Pochec, University of New Brunswick, Canada

Prof. Dr. Anders Fongen, Norwegian Defence University College, Norway

Prof. Dr. Subhasish Mazumdar, New Mexico Tech, USA

Dr. Hans-Werner Sehring, Tallence AG, Hamburg, Germany

Prof. Dr. Petre Dini, IARIA, USA/Europe



Panelist Position

Nice
June 2023

Establish standards for explainability in AI based decision making

AI based decision making: applications

- Automation of high-volume decision making

Explainability: motivation

- Establish trust
- Support audit process
- Satisfy policy criteria
- Monitor and improve decision making

Explainability process: understanding

- Understand and interpret predictions
- Actionable explanations, human interpretable explanations

Trade-offs in explainability

- Explainability vs efficiency vs accuracy
- Is explainability possible at all in some cases?

Algorithms for AI decision making and their explainability potential

- Rule based, theorem proving
- Statistical classifiers
- Neural network based
- other



Przemyslaw Pochec



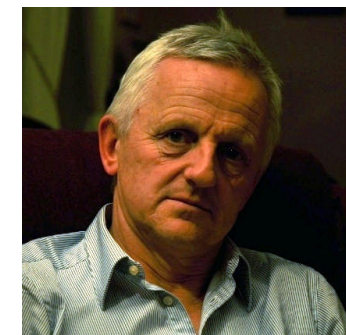
Panelist Position

Nice
June 2023

Scalability and explainability in AI based info services

For AI systems to improve **trust and transparency**, their explainability will also be related to their **scalability**. Why?

- * Provide consistent performance for a large group of users and a high number of queries (*timeliness*)
- * Groups of users should be able to share their experiences and results for consistency (*predictability*)
- * Training of AI systems should be efficient to keep the results updated with new information and trends (*relevancy*)



Anders Fongen




Panelist Position

NICE
June 2023

Human decisions

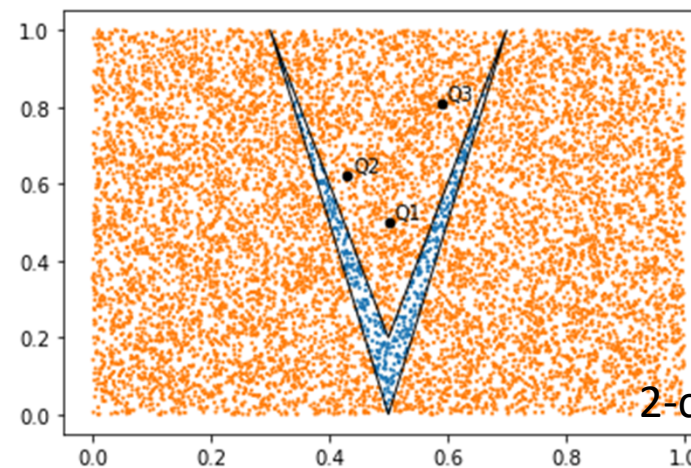
- Random choice, e.g., coin toss
- Arbitrary rule by fiat
- Based on defined Process
 - no obvious fraud / cheating
- Explanations needed (process + results)
 - legal disputes
 - selection from applicant pool
 - diagnosis of disease
 - launch of nuclear weapons
 - ...
- Forms of explanation:
 - text
 - with citations
 - with arguments
 - with emotions
 - Visual
- Audience for explanations: humans

AI-based decision systems

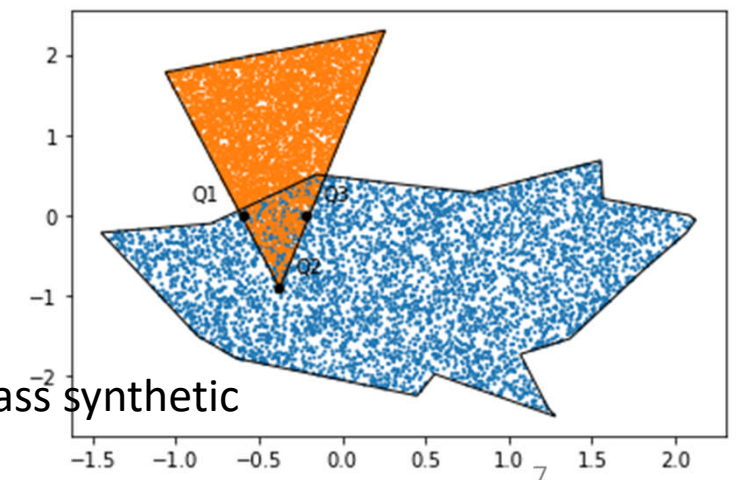
- Explanations
 - 19,000-dimensional hyperplane
 - Nearest training sample
 - Decision tree: simple predicates
 - can be deep
 - Tools: *LIME*
 - *Process (Bias)? Audience? Text?*
-  Tradeoff: accuracy vs communicable explainability



Subhasish Mazumdar



(a) LIME succeeds



(b) LIME fails



Panelist Position

NICE
June 2023

New challenges with “2nd spring of AI”, e.g., generative AI

Explainability of a results requires insights into ...

... the **deduction process**

- which reasoning steps led to an answer?
- heuristics applied?
⇒ local optimum
- manual intervention in the process (e.g., for chat bot)?
⇒ personalized result

... the **training data / fact / document base**

- local set of documents
⇒ bias
- content crawled from Internet
⇒ incomplete, inconsistent, etc. information
- authorship of and changes to the data (data provenance)
⇒ reliability

... of the **training**

- pretraining
- reinforcement, user input
⇒ bias
- underfitting and overfitting
⇒ accuracy, relevance



Hans-Werner
Sehring



Panelist Position (i)

Nice
June 2023

Guidance for an understandable explainability

- Low level (explain technical terms, acronyms, or abbreviations); striving to make the explanation as clear and accessible as possible is the main target. Adaptation of the level of details to the audience
 - Explicit the objectives in trust-able and measurable terms
 - Feedback (an iterative) loop for explanation is suitable
 - Explicit input data (procedure, tools) pre-processing (e.g., in cleansing for erroneous or incomplete data, etc.)
 - Refer and compare to concrete real-world situations/examples.
-
- **Danger:** Simulating AI-based potential actions is a destructive science fiction, as generalizes the perception that nobody can be hurt or lose something.



Petre Dini
IARIA



Panelist Position (ii)

Nice
June 2023

Difficulties

- Plain language versus formalisms; A deep explanation might require an explainability system more difficult to be understood than the target-system itself.
- Visualization provide confidence: yet, complex information via charts, flowcharts, diagrams are not for everybody.
- In AI-based decision systems, I see that exposing the thinking decision flow might be acceptable and convincing.
- **Vicious circle:** not everybody will understand it; a mediator is needed; a new intermediate link will damage the trust and increase the susceptibility of biased/distorted output.



Petre Dini
IARIA

Explanations are for fairness, accountability, and trust

- **Trustfulness** requires regulations, design-by-regulations, audit, compliance verification and validation, certification; repeatable accuracy output is also an expectation
- **My position:** A government agency is useless for explainability
- **Solution:** A Neutral Association of all Parties with renewal of the members in charge every x months (Developers, Citizens, Governments)
Mandatory Explainability task force in any team/corporation that develops AI-based tools
Mandatory audit and enforced heavy penalties for wrong-doers



Panelist Position (iii)

Nice
June 2023

Complexity, Skills, Costs

- **System documentation** is not a strong activity, except safety (avionics, automotive) and security (transaction systems).
- **Explainability is a mandatory** process regardless whether a system is AI-based or not (explanation of the process, justification of the output)
- In traditional process, '*garbage in, garbage out*' was always a generally accepted motivation of any output
- Now, with **data pre-processing** (cleansing, provenance, etc.) and AI-based data processing (deep-learning), the above statement doesn't longer hold.
- The Dataset accuracy validation (**bias**) depends on the threshold between training and testing data
- Explainability must cover the **entire Life Cycle** of a product from specifications to routine maintenance, when deployed/in use (see: airbags, AI-sensing)
- Explainability **must allow an audit** on details of the data processing (see: asking for a loan and credit history; data or process)
- **Explainability should be personalized** per class and instance of an entity (see: Boeing 740 -MAX); this allows a bottom-up validation, leading to corrections to the entire design process.
- Practically, **explainability is used as a trigger for verification and validation processes** either due by an unsatisfactory output or for boosting the end-user trustfulness.



Petre Dini
IARIA



Panelist Position (iv)

Nice
June 2023

Position: Each AI-based project must have an associated team/expert in charge with Explainability process, as a condition for certification

YET, Explainability is not a panacea to all dangers.

- Explainability helps building trust! It does not replace **bias** and (dis)**honesty**.
- The airbag history (1952). Nowadays, there are still hundreds of thousands recalls/year due to airbags failures.
- The best way for improving the trust for explainability is testing validated use cases; **test-then-trust!**
- Raise the level of awareness and trust in an honest way, to reinforce perception and belief
 - *Belief is difficult to be enforced because of \$3.999 paradigm; see well-known "only \$3.99+9/10/gallon" instead of '\$4/gallon' that triggers suspicion onto honest users*
 - *By extrapolating, an AI-based system might tacitly exploit the human weaknesses and/or distractions, offering even a wrong 'yet, being credible' explanation.*
 - *When buying gas, you might move to another gas station; but, when a social decision (or, health, or security, or ...) uses the same approach, this is misleading the entire society.*



Petre Dini
IARIA

Explainability is (very) costly!

Then, a diligent ROI process is advisable!

Selection of critical systems under Explainability scrutiny is advised!



OUTPUT

Nice
June 2023

1. The purpose of explainability is to convince the users (and the authorities) about the validity of the conclusions derived by the AI based system. This should be a standard Software Engineering task of verifying that the responses of any system are correct.
2. Theoretically speaking, the task of explainability is parametrized differently than the operation of the target system: the AI system uses inputs to derive a conclusion. The explainability task has all the inputs and has the conclusion already derived, Then it has a new goal of SHOWING how that the conclusion was derived from the inputs. As a consequence, an explainability system has MORE inputs and potentially more complex.
3. Theoretical proof of correctness can be applied only to very simple systems.
4. In rule-based systems, the explanation is the list of rules used to derive the conclusion.
5. In other scenarios, a statistical argument can be made about the validity of the answer (this was disputed as insufficient for legal reasons by other panelists).



OPEN DISCUSSION

NICE
June 2023

Stage for the Audience