IARIA

PANEL #2

VALENCIA
November 2023

SocSys 2023 & SoftNet 2023

Theme

Explainability and Self-Control on AI-driven Narratives and Systems

IARIA

CONTRIBUTORS

VALENCIA
November 2023

Moderator

Prof. Dr. Stephan Böhm, Hochschule RheinMain, Germany

Panelists

Prof. Dr. Martin Zimmermann, Offenburg University, Germany
Prof. Dr. Radek Koci, Brno University of Technology, Czech Republic
Prof. Dr. Matthias Harter, Hochschule RheinMain, Germany
Prof. Dr. Anders Fongen, Norwegian Defence University College, Norge

Prof. Dr. Mo Mansouri, Stevens Institute of Technology, USA

IARIA

Chair position

Valencia
November 2023

- **The existing lack of explainability and self-control can raise concerns about the outcomes of AI-based decisions**
  - No native self-control or explanatory power in AI systems
  - There may be an undetected bias in training data
  - Unwanted or unforeseen behavior may lead to a loss of trust
- **Self-control in AI systems is crucial to prevent unintended consequences and ensure ethical use**
  - Mechanisms for control and bias detection are required
  - Misuse and unethical behavior of systems has to be prevented
  - Regulation bodies to establish self-control are to be defined

Stephan
Böhm
RheinMain
University

Chair position

Valencia
November 2023

- **Self-control and explainability have domain-specific different meanings and relevance**
  - Self-control is of particular importance when AI (system) behavior is to be governed and restricted
  - Explainability is of special importance in domains where (human) decisions are automated and need to be understood
- **New skills and awareness are needed for humans using AI**
  - Approaches and “the drive” to question AI results of an (in principle) trustworthy technology ("pocket calculator effect")
  - Ability to recognize limitations of AI systems as well as maintain the ability to continue manual intervention in irregular situations ("autopilot problem")

Stephan
Böhm
RheinMain
University

IARIA

Chair position

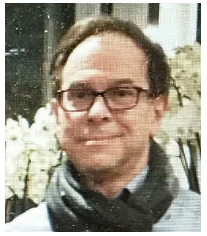
Valencia
November 2023

- **A balance between human control and AI autonomy is required**
  - Human intervention can negatively influence the performance and efficiency of AI
  - Intransparent control interventions might compromise the perceived “objectivity” of AI decisions.
- **There are significant challenges – especially in democratic societies**
  - High sensitivity for data protection and ethical dimensions
  - There are no simple solutions in sight – due to high complexity and a very dynamic environment
  - Potential risk of overregulation and loss of competitiveness

Stephan
Böhm
RheinMain
University

Panelist position

Valencia
November 2023



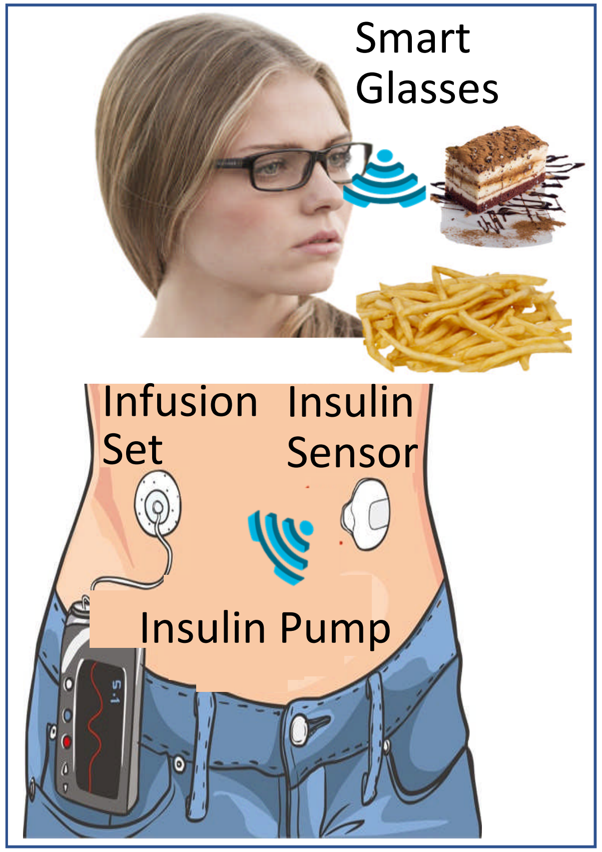
Martin Zimmermann
Offenburg University

- **Explainability is important**
  ... where users or stakeholders need to understand why AI systems made a particular choice / decision

- **Key points**
  - Transparency
  - Trust and Accountability (e.g. a future AI-based insulin control system)
  - Regulatory Compliance

- **However, explainability is less important in scenarios like**
  - Entertainment and Creativity
  - Non-critical Recommendations
  - Rapid Prototyping and Experimentation

→ **The importance of explainability depends on the specific use case, the impact of AI decisions, … and user expectations**



IARIA

Panelist position

Valencia
November 2023

- Systems that learn our preferences based on the behavior we provide (e.g., recommendations, report summarization).
- Systems that learn based on large amounts of known data (e.g., insurance companies, technological diagnostics, code design).

- AI does not discriminate facts but works with large amounts of known data.
  - It is possible to get correct conclusions from the data.
  - Can be used as basic suggestions or proposals for human decisions.
  - Cannot retrospectively infer (explain) causality of reasoning leading to solutions.
  - Will we have enough (legal) data? What about privacy?

- Mutual learning AI algorithms (e.g., deepfakes vs. their detection) - algorithms improve, but could they be beneficial (for humans)?
- With the use of AI (even as guidance), won't the ability of humans to discriminate and make independent decisions decrease?

Radek Kočí,
Brno University of Technology

IARIA

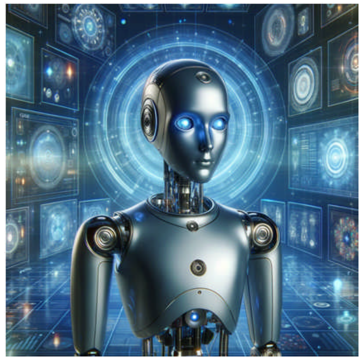
Panelist position

Valencia
November 2023

generated by DALL-E 3

- **Product development assisted by AI: the near future**
  - **Not only for SW, co-pilot for HW (domain-specific) imaginable!**
  - **AI assistance: Just another step in automation? Just a tool?**
  - **AI even for safety critical applications? Is the black box a problem?**
  - **Human-in-the-loop policy (human-machine alignment) critical!**
  - **Our job: what do we (humans) want AI to (not) do?**
  - **Engineers link society and technology (i.e. AI), act as interpreters**

Matthias Harter
Hochschule
RheinMain



generated by DALL-E 3

- **AGI acting as stand-alone engineer: the far end**
  - **Eventually no more design / implementation done by humans, only guidance without deeper understanding**
  - **Human-in-the-loop principle hard to maintain (due to cost, efficiency)**

IARIA

Panelist position

Valencia
November 2023

Anders Fongen,
Norwegian Defence
University College

- Does *anyone* know the structure of the "black box"?
  - Explainability – "*You may not like it, but now you know*"
  - Self control – Ability to tailor the service to individual users, not only the output, but the selection of training material as well
  - Fear – from not knowing the consequence of your own decisions
  - Mistakes – Are we giving AI "a slack", like we give humans? Who do we appeal to, who are we holding responsible?
  - Fairness – Will self-control lead to unfair and discriminating decisions?

IARIA

Panelist position

Valencia
November 2023

Importance of Explainability and Self-Control

- Trust and Transparency
- Ethics
- Ensuring Accountability
- Enabling Human Oversight and Intervention

Explainability Techniques and Methods

- Explainable Model: Decision Trees, Linear Models, Rule-based Systems…
- Post-hoc Explanation Methods: Black Box Model…
- Sensitivity Analysis: Feature and Counter Factual Explanations…
- LIME (Local Interpretable Model-Agnostic Explanations): Simplification…

Mo Mansouri
Stevens Institute of Technology, USA

Where Self-Control is imperative!

- **Autonomous vehicles:** They need to have self-control to follow the traffic rules, avoid collisions, and adapt to changing road conditions. They also need to balance the trade-offs between safety, efficiency, and comfort for the passengers and other road.
- **Autonomous weapons:** They need to have self-control to comply with the laws of war, respect human dignity and rights, and minimize collateral damage and civilian casualties. They also need to be able to distinguish between combatants and non-combatants, and to abort or deactivate themselves in case of malfunction or loss of control.
- **Artificial agents:** These are AI systems that can interact with humans or other AI systems in various domains, such as games, social media, e-commerce, or education. They need to have self-control to achieve their objectives, cooperate or compete with others, and follow the norms and expectations of the domain. They also need to be able to explain their actions and intentions, and to respect the preferences and privacy of the users.