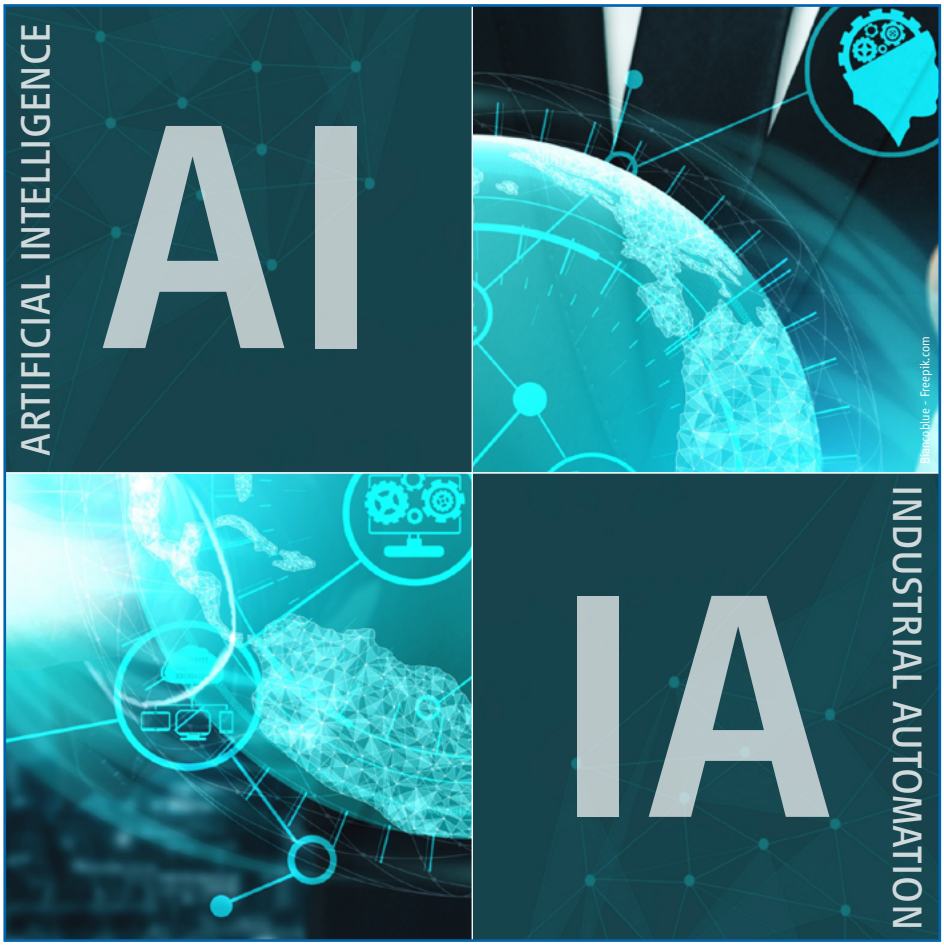
ZVEI:
Die Elektroindustrie

Discussion Paper

Surprisedness
A Human-Centric Metric of Trust in Industrial AI



German Electro and Digital Industry Association

Content

1 Human Centred AI branded in Europe

AI developments in Europe are facilitating a European brand of trustworthy, ethical AI that enhances human capabilities in decision making. With respect to core AI topics, fundamental gaps in knowledge and technology must be addressed in three closely related areas[1]:

- Learning, reasoning and planning with the human in the loop
- Multimodal perception of dynamic real-world environments and national values
- Human-friendly collaboration and co-creation in mixed human-AI settings

The human decision making process is quite error prone as we know from Nobel Memorial Prize in Economic Sciences laureate Daniel Kahneman[2] 2002. His main thesis is that of a dichotomy between two modes of thought resulting in decision making: „system 1" is fast, instinctive, emotional and biased; „system 2" is slower, rational and statistics based.
Humans consider themselves as rational and analytical. Thus, we assume, that most of our decisions are based on the "system 2"-mode. But actually we decide most of our time fast, instinctive, emotional and biased (system 1) avoiding subconsciously the effort of the quite exhaustive rational and statistics based approach (described by system 2). Only if we encounter something unexpected, or if we make conscious effort, we engage "system 2".

AI-based decisions rely on data and in most cases on trained models, verified by human experts. In this sense AI today acts as an automated support of "system 2". With growing complexity, AI may suggest solutions humans have not thought of.
In this case the AI system may have suggested a solution challenging the fast decision process of "system 1". Confronted with the unexpected the human will react with surprisedness, being one of the six basic and universal emotions that were found constant across all cultures (Paul Ekman). When AI-based systems are interacting or cooperating with humans, this state of surprisedness needs to be looked at in detail to qualify and initiate further resulting actions. Being surprised the human wants to resolve this state starting a dialog to get deeper insight into the relying on dataspace and system boundary to verify, reject or modify the AI-based solution.

For the interaction with the AI system, we therefore will need

- A metric of trust based on the level of surprisedness
- A human-AI system interface supporting interaction and dialogue

Establishing the metric and interaction method in mixed human-AI settings will increase the trust in AI-generated solution proposals. This will also support the development of AI-systems, being able to handle the human surprisedness and thus support human learning.

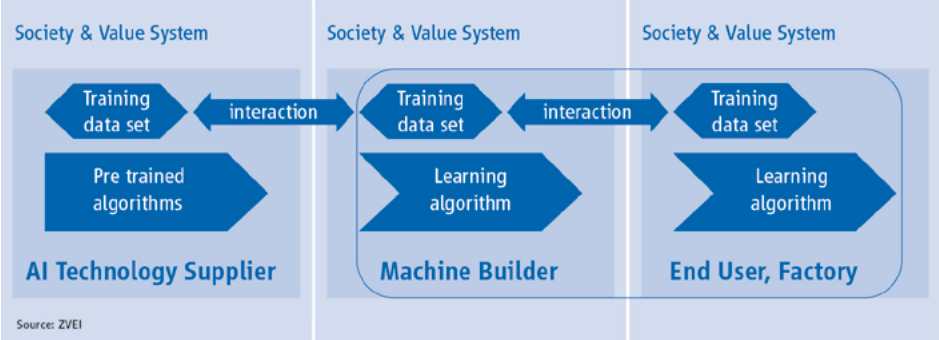
[1] https://www.humane-ai.eu/research-roadmap/

[2] Daniel Kahneman: Thinking, Fast and Slow. Farrar, Straus and Giroux, 2011. -ISBN 978-0374275631

2 Proposal of a Metric of Trust in mixed human-AI settings

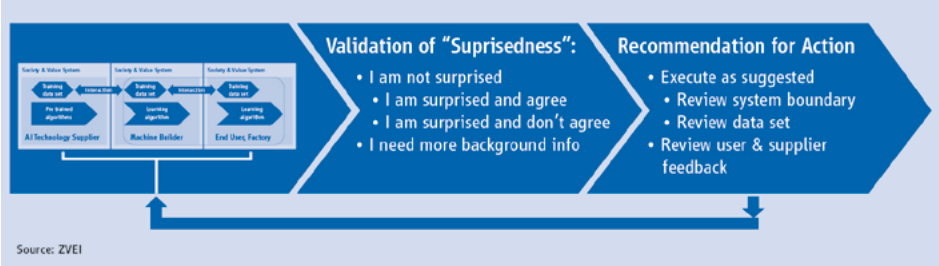
Our thesis is that evaluating the "level of surprise" encountered in decision making processes at mixed human-AI settings will significantly improve the interactions between the humans and the AI-system leading to more trust and a better understanding of AI-system responses. Overall, it will increase the acceptance by all kind of users within the value chains, which is based on the trust that develops from the matching of expectations and impact and to a certain extent through the traceability of technical system responses.

We therefore support the proposal by the working group 2 of the Plattform Industrie 4.0 to use "surprisedness" as a new metric in human-centred AI[2]. Deploying this metric will help to reduce errors in decision making processes and may be scaled across different stakeholders of the value chain, e.g. AI service suppliers, machine builders and end users or factory owners.



System Boundary of stakeholders in value chain and interaction across the value chain

Integrating the human in the loop by applying the new metric of "surprisedness" in above scheme we can differentiate the type of interactions between each system boundary encompassed in the value chain from supplier to integrator to factory. Further we have the information which will allow to develop an AI-system to resolve the human state of surprise.



Feedback loops triggered by the Human in the loop when validating levels of surprisedness

[2] https://zukunftsmonitor.de/2021/05/11/kuenstliche-intelligenz-und-akzeptanz-mit-ueberraschungen-umgehen

3 AAS as interface for humans in the loop

We consider the concept of the Asset Administration Shell (AAS) as a promising approach and interface for the future human-AI system dialogue. A sub-model may include machine readable data reflecting the human feedback and level of surprise.

First developments of such interfaces have already been proved in an AI-based demonstrator for analysing monitoring data and providing predictive maintenance, where human service is automatically initiated. Aside for describing the assets, the AAS is also used for describing the human skills, functional abilities, roles, communication channels and methods as well as the availability of human resources. This allows the system finding an appropriate support in time. (Dr. Thomas Kuhn, Fraunhofer IESE, Presentation at ZVEI, FK Industrie 4.0, Dec. 7th, 2021))

In our more advanced scenario, the AAS will also need to provide information about the personal, social and cultural background for trust and surprisedness. Any appropriate sub-model for the AAS allowing this dialogue will need to be dynamic as the ongoing interaction with the AI-System will lead to human learnings increasing level of experience and trust.

4 Scaling fully automated use cases will involve the human interaction

We strongly belief that even high scaling and fully automated uses cases (e.g. collaborative condition monitoring, CCM) will need to integrate the human in the loop when growing to larger, more complex ecosystems. When the system boundaries are extended on a global level including several cultures, we need to integrate the human feedback and re-action into the processes, to ensure the overall acceptance in the value chain encompassing suppliers, integrators, factory owners and customers.

A couple of questions will arise, like how Artificial Intelligences can be controlled, how traceable their applications are, and which system limits should be set. Answers to these questions are determined by the expectations and value systems of societies.
Including the human in the loop, data-based, rational decisions will be made transparent for the human in the dialogue.

This discussion paper results from ongoing discussions with representatives from different platforms (PI4.0-AG1, 2) and standardization activities (SCI4.0)

List auf Authors

Johannes Diemer (Diemer Consulting 4.0 e.K.)

Juergen Grotepass (Huawei Technologies Duesseldorf GmbH)

Johannes Kalhoff (Phoenix Contact GmbH & Co. KG)

Christoph Legat (HEKUMA GmbH)

Dieter Wegener (Siemens AG)

ZVEI:
Die Elektroindustrie

Surprisedness: A human-centric Metric of Trust in industrial AI

Publisher:
ZVEI e.V.
Automation Division

Lyoner Straße 9
60528 Frankfurt am Main

Isabelle Kuhn
Telefon: +49 69 6302-429
E-Mail: isabelle.kuhn@zvei.org
www.zvei.org

March 2022