On the Importance of User Role-Tailored Explanations in Industry 5.0

Inti Gabriel Mendoza[1][a], Vedran Sabol[1,2][b] and Johannes Georg Hoffer[3][c]

[1]*Know-Center GmbH, Sandgasse 36, Graz, Austria*

[2]*Graz University of Technology - Institute of Interactive Systems and Data Science, Sandgasse 36, Graz, Austria*

[3]*voestalpine BÖHLER Aerospace GmbH & Co KG, Mariazellerstraße 25, Kapfenberg, Austria*

Keywords: eXplainable AI, human-AI Interface Design, Explanations, Personalization, Process Engineering.

Abstract: Advanced Machine Learning models now see usage in sensitive fields where incorrect predictions have serious consequences. Unfortunately, as models increase in accuracy and complexity, humans cannot verify or validate their predictions. This ineffability foments distrust and reduces model usage. eXplainable AI (XAI) provides insights into AI models' predictions. Nevertheless, scholar opinion on XAI range from ”absolutely necessary” to ”useless, use white box models instead”. In modern Industry 5.0 environments, AI sees usage in production process engineering and optimisation. However, XAI currently targets the needs of AI experts, not the needs of domain experts or process operators. Our Position is: XAI tailored to user roles and following social science's guidelines on explanations is crucial in AI-supported production scenarios and for employee acceptance and trust. Our industry partners allow us to analyse user requirements for three identified user archetypes - the Machine Operator, Field Expert, and AI Expert - and experiment with actual use cases. We designed an (X)AI-based visual UI through multiple review cycles with industry partners to test our Position. Looking ahead, we can test and evaluate the impact of personalised XAI in Industry 5.0 scenarios, quantify its benefits, and identify research opportunities.

1 INTRODUCTION

Machine learning (ML) and artificial intelligence (AI) models continue to evolve and improve. AI has already achieved human superiority, as it may outperform humans in tasks like pattern recognition (He et al., 2016). Thus, AI models see usage in many areas, from video games, over criminal recidivism (Kennedy et al., 2022), to production engineering and optimization (Hoffer et al., 2022), and more. Some of these areas are of paramount importance to get right. For example,

failing to predict the need for maintenance of a machine can interrupt a production process for days, leading to huge losses. Nevertheless, usage of ML models in critical areas will continue, and it should - their efficacy in most cases is too valuable. However, as the Russian proverb goes, “trust but verify”. The need for explainable artificial intelligence (XAI) is thus also increasing.

XAI is a subset of AI which aims to explain a model's predictions. In essence, a model is explainable (or interpretable) if *the user* can understand why a model produced a specific output prediction. On the other hand, accurate but very extensive explanations do not mean helpful or useful explanations. Some scholar dispute the need for XAI suggesting use of white-box models, which often includes an accuracy loss. In any case, through useful explanations, a model made transparent or explainable, should be trusted by users. (Miller, 2019).

Few XAI research integrates social and behavioural sciences findings. This lack of transdisciplinarity is worrisome as researchers keep creating explanation quality metrics. Social and behavioural sciences have already defined what makes up a proper explanation (Miller et al., 2017). Furthermore, XAI methods seldom target the user. The assumption stands that if researchers understand the explanations, the target user will. Unfortunately, this is an example of “logic gymnastics” that does not follow. XAI methods are supposed to provide explanations. Explanations, as a whole, are a form of conversation. In conversations, humans tailor conversations to the recipient's level. Thus, XAI must also do this when

[a] https://orcid.org/0000-0001-6779-7948
[b] https://orcid.org/0000-0002-0599-445X
[c] https://orcid.org/0000-0002-2441-7279

Mendoza, I., Sabol, V. and Hoffer, J.
On the Importance of User Role-Tailored Explanations in Industry 5.0.
DOI: 10.5220/0011748300003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 2: HUCAPP*, pages 243-250
ISBN: 978-989-758-634-7; ISSN: 2184-4321
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

presenting explanations to end users.

We argue that XAI models' output explanations are not the ultimate output. These explanations must take the work environment and the users' abilities to understand the explanations into account when providing the ultimate output. Thus, we, the authors, believe that in the current XAI research landscape, there is a need for further handling of explanations and tailoring the communication to specific target groups.

This position paper provides the steps we intend to follow to develop a proof-of-concept demo, with focus on tailoring XAI to user archetypes in production processes. Based on that, we intend to show that automatic output manipulations are possible, helpful, and, most importantly, *needed*.

The structure of this position paper is as follows: Section 2 gives related works on XAI and social science's research on explanations to support our Position that further handling of XAI output is needed. Section 3 describes our setting, clarifies our Position's context, and validates the need for user-targeted explanations. Section 4 discusses the validity of our Position in a more general case. Section 5 describes how we intend to prove our Position empirically.

2 RELATED WORK

This section provides an overview of the current XAI landscape and information obtained from a literature survey, including papers with research from the social sciences. The main focus of our literature survey was on explanations. However, this section will also show the need to treat XAI research interdisciplinarily.

2.1 Explanations According to XAI

People should be able to challenge AI models' predictions for critical tasks. The reasoning behind the prediction should not be "because the model said so". This case is one of many factors contributing to the increased need for Explainable AI (XAI) research - even being explicitly pushed by the European Commission (Carvalho et al., 2019).

XAI is currently considered another field in computer science where its models produce human-understandable explanations for interpreting AI models. However, due to the field's relative youth, the most basic XAI concepts must be clarified. For example, researchers use the terms "explainability" and "interpretability" interchangeably. Furthermore, there has yet to be an agreed-upon method to assess an XAI system's explanation's quality (Šimić et al., 2022).

Most notably, researchers have yet to reach a consensus on the nuclear goal of XAI. Most say that the nuclear goal is to establish trust in users (Sperrle et al., 2020). Besides model validation, another commonly stated goal is model debugging, i.e., identifying reasons for wrong predictions. Finally, our literature survey shows a reluctance to admit that XAI is not a pure computer-science discipline. The field of XAI shares the academic landscape with the social sciences. Regardless of these shortcomings, XAI methods have already seen the light of day in areas like medical diagnoses, autonomous vehicles, and process engineering, among many more.

2.1.1 XAI Used in Process Engineering

XAI already sees usage in process engineering. For example, the car manufacturer, Volvo, has already implemented XAI to optimize its process lines. Here we will describe the utilization of XAI in process engineering in more detail.

(Mehdiyev and Fettke, 2021) described Volvo's problem context and ultimate solution. Volvo maintains three repair lines in charge of repairing broken parts in their other process lines, one more specialized than the previous. However, lesser specialized lines kept misjudging task complexities and forwarding them to deeper lines, creating a bottleneck problem. Researchers, to combat this, developed a (black-box) AI model to predict which repair line should handle a given task. However, this model must work with humans; thus, its inability to explain its decisions fomented distrust and lack of usage. To circumvent this, they integrated an XAI layer and, in so doing, developed a framework to develop XAI methods to tackle this type of problem.

2.2 Explanations According to Social Sciences

(Puiutta and Veith, 2020) notes the need for behavioural sciences references in XAI papers. *Explanations* are an area already studied in the social sciences. AI researchers should refrain from attempting to reinvent the wheel when defining proper explanations. Furthermore, discoveries and knowledge of social sciences on explanations are abundant, available, and valuable. It is then unreasonable not to take advantage of this knowledge. Our literature survey identified a set of helpful social sciences knowledge that is paramount and useful for XAI research and usage.

2.2.1 Explanations Are Contrastive

Behavioural science teaches us that humans prefer contrastive explanations. When humans ask "*why*" questions, they implicitly ask for a contrastive case (Mannheim et al., 1990). Humans want to know why a unique explanation is correct instead of a different one. When humans turn to XAI for explanations, they ask these "*why*" questions, thus posing a unique challenge for XAI research. Proper implementation of this knowledge will ultimately lighten the load of providing a dense, complete, and thus *improper* explanation (more on this statement later). Furthermore, an explanation does not have to cover or provide all cases - it does not have to *find* all cases (Lipton, 1990).

2.2.2 Attribution Theory & Temporal Causality

Social sciences developed *attribution theory*. This theory tries to explain how people attribute causes to consequences. Generally, there are two types of attributions: *social attribution* and *causal attribution*. Social attribution describes how people attribute others' behaviours. People attribute causality based on the acting party's characteristics. Successful actions are different from unsuccessful actions. Unsuccessful actions often have an unmet precondition attributed to them. This theory is essential when designing XAI methods, particularly when an explanation comes from an intentional action viewed as a cause.

Causal connection describes how people connect causes. They do so by seeing a prediction based on how it would have changed if it had stemmed from a different action (Kahneman and Tversky, 1981). If the "changed" action is too far from the past, the causal chain might be too complex to imagine or see its relevance. This theory focuses on how people choose which action to "undo". Actions that are proximal, abnormal, and more "controllable" are often chosen to create this causal chain (Miller et al., 2017).

2.2.3 Explanations Are Tailored to the Recipient's Knowledge

As field experts, it is not uncommon to think of specific field-specific knowledge as fundamental or common knowledge. However, even something as basic as terminology can elude "regular" (non-field experts) people. Moreover, it is not uncommon to think that learning this knowledge is a trivial task - it is not so. It is thus troublesome when XAI researchers, who develop explanations-producing XAI models, can understand these explanations and incorrectly assume that end-users will too. These researchers, therefore, assume that their explanations are *good* explanations. This flawed thinking is why (Miller et al., 2017) warns of "inmates running the asylum".

2.3 Supporting Our Position

The abovementioned Volvo example proves the applicability of XAI methods in process engineering contexts. Furthermore, it illuminated the problem of workers not trusting - and thus not using - AI models to help them and how an XAI layer implementation alleviated some of these problems. This insight supports our Position that XAI methods are helpful in process engineering contexts to increase workers' trust and usage of AI models. Furthermore, our literature survey on explanations supports our Position that following social sciences guidelines creates helpful explanations that foment user trust, and XAI methods should follow them (Miller et al., 2017).

3 USER-TAILORED XAI

This section elaborates on what "further handling of XAI output" means. First, we describe an example problem context where a black-box ML model performs valuable predictions. Moreover, we create an XAI model that explains the initial model's predictions. Finally, once we obtain candidate explanations, we handle and manipulate them in order for the resulting explanations to conform with social science's research on what constitutes a proper explanation.

3.1 Problem Context

For our problem context, we set ourselves in a process engineering setting. Specifically, we want to help a company manufacture complex parts efficiently. Input materials (materials as obtained by resource providers) must undergo work in multiple process chains - each consisting of process steps - to reach an acceptable output state. Inaccuracies, mistakes, and blunders are bound to occur and create different consequences. For example, an item with an inaccuracy is repairable, but an item with a mistake must be discarded and sold off at a loss. An item with a blunder might not even be worth selling. It can also happen that an item repair is attempted (incurring repairing costs) but ultimately reassessed as un-salvageable. The opportunity cost of working on an ideal or salvageable item instead is, of course, incurred as well.

Thus, it would be beneficial to create an AI model that can predict with sufficient accuracy whether some material, after processing, will be transformed into an acceptable output material. Since we are working in the process engineering area, the AI and the employees must work together. Teamwork involves user trust in the model, and, as Volvo's experiences show, the AI must explain its predictions to work with humans productively. This model will have to work with each identified user archetype and thus be able to help in material processing, process design, and data mining.

3.2 Identifying User Archetypes Requirements

Explanations must tailor to the recipient's knowledge and working environment; thus, we must identify our user demographic and their (explanation) requirements. Through numerous talks with our business partners, we identified three relevant user archetypes - the Machine Operator, Field Expert, and AI Expert - and their UI and explanation requirements. The Machine Operator only needs to identify obvious blunders in the AI model related to one process step, but has a little time to do so. The Field Expert is knowledgeable and needs to verify the AI model's knowledge and identify blunders, mistakes, and inaccuracies in the model predictions. Finally, the AI expert must identify reasons for wrong model predictions, in order to fix it. These three user archetypes are common in most AI-augmented process engineering/optimization use cases, even ones different from those in this scope. Talks with our business partners support this fact. In general, the AI expert creates the model, the field expert verifies the model's output, and an operator will only want to use the model that can be trusted. Thus, our planned framework is helpful in the general scope of process engineering.

3.3 Decision Trees

Decision Trees (DTs) are rule-based white-box models. Once induced, it is possible for humans to follow a prediction on reasonably sized trees. DTs group data by creating discrimination criteria (nodes) to extract and label the structure of the data (Dasgupta et al., 2020). These criteria are attribute thresholds. If an input attribute falls within a specific range, the prediction "goes" in a particular direction down the tree. The prediction branch consists of the traversed nodes and the final leaf. This leaf is the prediction.

We intend to use DTs as model-agnostic surrogate models (metamodels) of an accuracy-focused (black-box) AI model. The AI model becomes an oracle to create training data for the DT-inducing process. This way, we extract what the AI model learned, including the structures leading to erroneous predictions, which are helpful for model debugging (Vilone et al., 2020).

Shorter trees use fewer attributes, thus only employing essential ones and sacrificing accuracy. On the other hand, the longer the tree is, the more accurate and specific it will be, as branches can accommodate more discrimination points (nodes). Since DTs can be parametrizable, we propose that they can be used to imitate proper explanations. Selection of the tree depth for the DT-inducing process offers the choice between generating accurate but complex (deep DT) or approximate but simple (shallow DT) explanations. Further, selection of the truly informative nodes provides an opportunity for personalization, as overwhelming a user with obvious or irrelevant facts would be counterproductive.

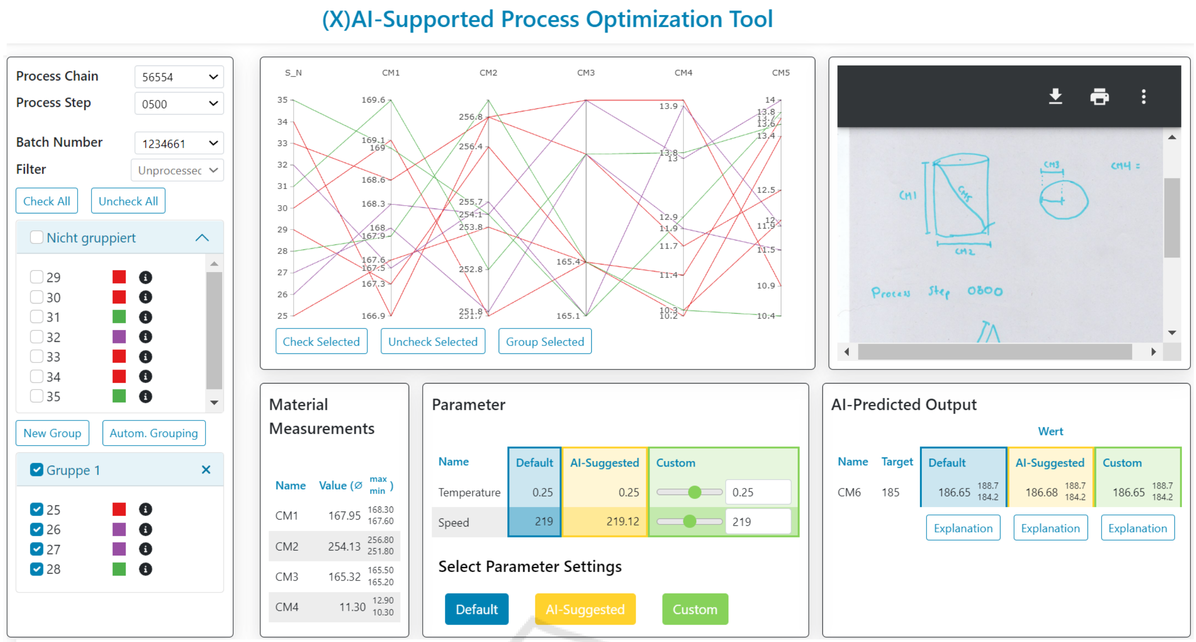
We note that explanations are contrastive. Any branch different from DTs's prediction branch can be considered a contrastive branch. Thus, these are the "what if" branches we can use to provide contrastive elements in explanations. Next, branches can mimic attribution theory because the nodes in the prediction branches are akin to chain links: prediction branches' nodes depend on each other - a node depends on the previous node. Furthermore, the location of a node in the prediction branch mimics a temporal value. Nodes at the start of the prediction branch are "farther" preconditions than nodes closer to the prediction leaf.

3.3.1 Main UI

In our problem context, all three user archetypes use the same base UI. As seen in Figure 1, we identified three main areas: left, top, and bottom. The left area allows the user to select the process step they want to work in and the materials they want to process. The user can then select and group the available materials to process them - which have been automatically colour-coded by an (out-of-scope) clustering algorithm to identify similar materials helpfully.

The top area has two elements: a Parallel Coordinates visualization and an area containing the Design Document for the current process step. Parallel Coordinates shows batch materials' control measurements to help with material selection and grouping. The Design Document allows the user to obtain the information they need about the current process step.

The bottom area has three elements: the Material Measurements table, the Parameter Settings Vector (PSV) table (including "Go" buttons), and the AI-Predicted Output measurement table. The first column of the PSV-table displays the default parameter settings used in the current process step. The second column contains an AI-suggested PSV that considers the selected materials with their exact measurements. The user can opt to receive an explanation for this prediction. The user can also use their custom PSV



(X)AI-Supported Process Optimization Tool

Process Chain: 56554
Process Step: 0500
Batch Number: 1234661
Filter: Unprocessec

□ Nicht gruppiert

□ 29	🟥	ℹ
□ 30	🟥	ℹ
□ 31	🟩	ℹ
□ 32	🟪	ℹ
□ 33	🟥	ℹ
□ 34	🟥	ℹ
□ 35	🟩	ℹ

☑ Gruppe 1

☑ 25	🟥	ℹ
☑ 26	🟪	ℹ
☑ 27	🟪	ℹ
☑ 28	🟩	ℹ

Material Measurements

Name	Value (Ø $^{max}_{min}$)	
CM1	167.95	168.30 167.60
CM2	254.13	256.80 251.80
CM3	165.32	165.50 165.20
CM4	11.30	12.90 10.30

Parameter

Name	Default	AI-Suggested	Custom
Temperature	0.25	0.25	0.25
Speed	219	219.12	219

Select Parameter Settings

AI-Predicted Output

		Wert		
Name	Target	Default	AI-Suggested	Custom
CM6	185	186.65 188.7 184.2	186.68 188.7 184.2	186.65 188.7 184.2

Figure 1: Main (base) UI.

within some allowed threshold (third column). The rightmost table allows the user to see the target output measurements, as detailed by the design document, and the AI-predicted target outcomes measurements. In this table, an AI model predicts the (average) target output if the selected materials underwent the process using each parameter vector. The user can obtain explanations for each prediction at the bottom of the table. Clicking to obtain explanations opens the Explanation UI view, which tailors the explanations to each user archetype's needs and requirements. Finally, in the "Go" buttons, the user can select which PSV to process the selected materials.

3.3.2 Explanation Differences

According to social science research, each user archetype needs a different explanation. Figure 2 shows how we intend to display explanations. The top part will always be visible and contain information on the selected materials, PSVs, and AI-predicted output target measurements. Users can change explanations to other output targets or use different PSVs. The bottom part is scrollable and contains tailored explanations as contributions of each control measurement and process parameter to the predicted output measurements in a visual format (detailed design to be decided upon). Tailored explanations are essential, and each user archetype has (already described) unique requirements. Therefore, explanations for Machine Operators must be as concise as possible, as their time is minimal. Field Experts have more time, so explanations can be more precise and tailored to their domain knowledge, enabling them to validate the model's knowledge. Explaining irrelevant or obvious facts is not only useless but also counterproductive. Finally, AI experts should be free to choose seeing as many different explanations as needed to understand the model's behaviour.

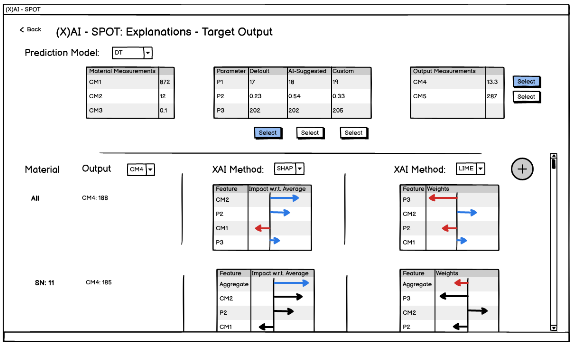
SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

3.4 Supporting Our Position

Large DTs can provide prediction branches, which are the source of explanations, that are big enough to overwhelm users. Therefore, DT models (and white-box models that do not consider user overwhelming) are not XAI models. It takes more than transparency to make an explanation understandable and useful. We have described how we propose DTs can produce explanations that follow social science's guidelines for good explanations. The ability to parametrize DTs and their mimicry of social sciences theory of explanations mean that explanation tailoring is possible. We could tailor them to each of our user archetype's knowledge and serve their unique requirements.

4 EVALUATION CYCLES

UI and XAI explanations design occurred with industry partners to aid one of their process engineering divisions. This collaboration ensured that the UI design decisions were always consistent with their process



(X)AI - SPOT: Explanations - Target Output

Prediction Model: DT

Material Measurements	
CM1	872
CM2	12
CM3	0.1

Parameter	Default	AI-Suggested	Custom
P1	17	18	19
P2	0.23	0.54	0.33
P3	202	202	205

Output Measurements	
CM4	13.3
CM5	287

Material Output CM4 XAI Method: SHAP XAI Method: LIME

All CM4: 188

Feature	Impact w.r.t. Average
CM2	
P2	
CM1	
P3	

Feature	Weights
P3	
CM2	
P2	
CM1	

SN: 11 CM4: 185

Feature	Impact w.r.t. Average
Aggregate	
CM2	
P2	
CM1	

Feature	Weights
Aggregate	
P3	
CM2	
P2	

Figure 2: Mock-up design of Explanations UI.

engineering and operation needs. The constant evaluations, revisions, and feedback resulted in the ultimate state of the UI and user archetype requirements. We elaborate on this process in this section.

4.1 Continous Evaluation

A continuous evaluation has continued for almost one year, with evaluation iterations occurring every two weeks for six months and fewer periodic evaluations afterwards. The initial three months consisted of mockup designs of the base UI (Figure 1, and the next three consisted in implementing these mockups into a clickable UI. The last six months have consisted of minor UI improvements and the integration of synthetic data and the initial AI and XAI methods. In the evaluation sessions, we presented the current progress and industry partners would then evaluate the included functionalities and ensure that the included functionality was valuable and necessary. This functionality assurance occurred in consideration of each user archetype. In addition, another partner conducted grassroots interviews with real users to get their perspectives concerning their acceptability of a tool to help them during their tasks. These interviews were included in our iterative evaluations, further contributing to our UI design decisions.

4.1.1 Following Guidelines for (X)AI UIs

Presenting differently tailored explanations for one prediction or algorithm produces different effects on users. Therefore, differently tailored explanations will have their advantages and disadvantages (Cai et al., 2019). Also, the medium and techniques employed to present these explanations are important. To address this, Microsoft (Amershi et al., 2019), Google (Google, 2022), and Apple (Apple, 2022) published guidelines on presenting explanations to users. In addition, Microsoft emphasizes meeting user expectations, Apple emphasizes smooth user experience, and Google emphasizes concepts that the *developer* must consider rather than the user.

While designing our (X)AI-assisted process engineering tool, we followed the guidelines published by these three companies. Following Microsoft's initial guidelines, we present to the user exactly what they can do with the tool. For our use case, the user must be able to do two things: see the materials about to be processed and process them, either using default, AI-predicted or self-defined parameters. Our UI allows and helps them to do just that and only that. When AI predictions appear implausible, or more information is needed, the UI allows easy access to explanations or the design document. Over time we aim to provide future functionality that would allow user flagging of suspicious AI results and allow the tool to adapt as user's knowledge increases.

Following Apple guidelines, we have identified the role of AI within our (X)AI tool concept. We make it clear, through colour highlighting, which UI areas are relying on the AI results. We initially hide the explanations for simplicity, but make them accessible on-demand through a single button. As Apple suggests, obtaining the explanations is a voluntary action.

Following Google guidelines, we aim to present trustworthy explanations. We will achieve this by following social sciences guidelines of what constitutes a *proper explanation* and providing explanations that follow these guidelines. Ideally, AI-suggested parameters should be more successful and desirable than the other non-AI-enhanced parameter vectors, so it makes sense for users to accept them. Furthermore, to use this model, our tool will nudge users to use these suggestions more often by explaining the predicted output when using these suggestions. Finally, in order not to lure users into false security, we will need reliable quality metrics to estimate and confirm the faithfulness of the explanations.

4.2 Supporting Our Position

After multiple evaluation cycles, our (X)AI UI concept is ready for further steps.Initial internal reviews of the UI's usability hint at positive user experiences. Satisfied industry partners also vouch for the value our UI will provide to their process engineering division. Finally, our attempts to prevent hyperspecialization in our industry partner's process engineering division allow us to present our UI as a possible general solution. This solution includes the generalization of user archetype explanation needs, further supporting the value of our Position could have.

5 EXPERIMENT PROPOSAL

Our literature survey outlines three challenges of XAI within the scope of this paper. These are:

1. *Involve end users when developing (X)AI methods.*
2. *Marry XAI research with social sciences and human behaviour studies.*
3. *Standardize studies that consider user traits like personality and education.*

To tackle these challenges, we aim to test the following hypothesis: we propose that the current XAI models' outputs must undergo further transformations to be proper explanations. Performing no transformation should result in (generally) users not understanding a prediction as much or as quickly as possible.

5.1 Experiment Proposal

To test our hypothesis, we will set up an A/B experiment for each user archetype using real user archetype-belonging subjects. Once we allocate a reasonable number of test subjects, we will divide them into two groups. One will receive the explanation in its untailored form, and the other will receive the transformed, tailored version. Next, the subjects will attempt to comprehend their explanation in a constant allotted timeframe. This sequence will occur several times on the same data set for multiple explanations. Afterwards, we will assess the knowledge they retained about the process and how they rate the explanations received. Finally, we will compare the scores of each group against each other, as well as the time and the effort it took for subjects to understand the explanations. Compiling all results will determine if the experiment supports or undermines our hypothesis.

5.2 Expected Outcome

Given proper handling of XAI outputs, a shorter, tailored explanation should be more comprehensible and thus easier to remember and learn. Moreover, presenting something easily digestible should also yield more satisfied users. In conclusion, we expect tailored explanations to be rated as more understandable, more useful and more trustworthy. We expect our hypothesis to be significantly supported, providing an empirical support for our Position.

ACKNOWLEDGEMENTS

This work was supported by the BrAIN project. BrAIN – Brownfield Artificial Intelligence Network for Forging of High Quality Aerospace Components (FFG Grant No. 881039) is funded in the framework of the program 'TAKE OFF', which is a research and technology program of the Austrian Federal Ministry of Transport, Innovation and Technology.

Know-Center is funded within the Austrian COMET Program, under the auspices of the Austrian Federal Ministry of Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Austrian Federal Ministry of Labour and Economy (BMAW) and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

We also thank colleagues from the voestalpine BÖHLER Aerospace GmbH & Co KG and Virtual Vehicle Research GmbH for their input and fruitful discussions concerning user requirements and user acceptance.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13.

Apple (2022). Human interface guidelines.

Cai, C. J., Jongejan, J., and Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 258–262.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

Dasgupta, S., Moshkovitz, M., Rashtchian, C., and Frost, N. (2020). Explainable k-means and k-medians clustering. In *International Conference on Machine Learning*, pages 7055–7065. PMLR.

Google (2022). People + ai guidebook.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE.

Hoffer, J. G., Geiger, B. C., and Kern, R. (2022). Gaussian process surrogates for modeling uncertainties in a use case of forging superalloys. *Applied Sciences*, 12(3):1089.

Kahneman, D. and Tversky, A. (1981). The simulation heuristic. Technical report, Stanford Univ Ca Dept Of Psychology.

Kennedy, R. P., Waggoner, P. D., and Ward, M. M. (2022). Trust in public policy algorithms. *The Journal of Politics*, 84(2):000–000.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266.

Mannheim, U. et al. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65–81.

Mehdiyev, N. and Fettke, P. (2021). Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring. *Interpretable Artificial Intelligence: A Perspective of Granular Computing*, pages 1–28.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv e-prints*, pages arXiv–1712.

Puiutta, E. and Veith, E. (2020). Explainable reinforcement learning: A survey. In *4th International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)*, pages 77–95. Springer International Publishing.

Šimić, I., Sabol, V., and Veas, E. (2022). Perturbation effect: A metric to counter misleading validation of feature attribution. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1798–1807.

Sperrle, F., El-Assady, M., Guo, G., Chau, D. H., Endert, A., and Keim, D. (2020). Should we trust (x) ai? design dimensions for structured experimental evaluations. *arXiv e-prints*, pages arXiv–2009.

Vilone, G., Rizzo, L., and Longo, L. (2020). A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence. pages 85–96.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS