Assessing stakeholder perspectives on the explainability of AI solutions for smart production planning with just-in-time logistics

Michael Heider, Marcus Albrecht, Johannes Schilp, Jörg Hähner

Angaben zur Veröffentlichung / Publication details:

Heider, Michael, Marcus Albrecht, Johannes Schilp, and Jörg Hähner. 2026. "Assessing stakeholder perspectives on the explainability of AI solutions for smart production planning with just-in-time logistics." *Expert Systems with Applications* 314: 131352. https://doi.org/10.1016/j.eswa.2026.131352.

Expert Systems With Applications 314 (2026) 131352

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Assessing stakeholder perspectives on the explainability of AI solutions for smart production planning with just-in-time logistics

Michael Heider [a,*], Marcus Albrecht [b], Johannes Schilp [b], Jörg Hähner [a]

[a] Organic Computing Group, Universität Augsburg, Am Technologiezentrum 8, Augsburg, Germany
[b] Digital Manufacturing Group, Universität Augsburg, Am Technologiezentrum 8, Augsburg, Germany

ARTICLE INFO

Keywords:
XAI
Explainability
Case study
Smart production
Production planning

ABSTRACT

In recent years, intelligent systems have increased their capabilities greatly increasing their practical applicability. However, for the foreseeable future, such AI-powered agents will not act autonomously but assist a human that will ultimately responsible. Here, the explainability of agents' suggestions becomes paramount to provide trust and acceptance by their human co-workers. For the field of stochastic/evolutionary optimization, it has not yet been investigated what levels of explainability real human stakeholders without deep technical knowledge of these systems actually request. In this article, we report an exploratory case study where we questioned a group of production planners ($n = 11$) about their needs for AI assistance and what types of explanations they would require to integrate AI into their day-to-day work-flow and still feel comfortable with the cooperation. While five participants expect their individual position to be threatened by these systems in the mid-term, all participants agree that AI is beneficial for safeguarding the location against competitors or migration. We find that AI-based assistance is requested to a large degree across all age groups and that stakeholders greatly request explainability of the agent's recommendations. From this real-world empirical evidence it becomes evident that implementing explainable optimization in production planning is a crucial next step towards Industry 5.0.

1. Introduction

As artificial intelligence (AI) systems continue to become more sophisticated, there is a growing interest in deploying them in domains where their application was previously limited or considered infeasible. Despite their rapid evolution, AI-based agents are generally not yet capable of autonomously executing complex tasks without human oversight. Consequently, AI typically acts as a supportive tool, augmenting human decision-making processes while the ultimate responsibility for decisions remains with human operators. This raises a number of design questions that are not easily answered and heavily dependent on the specific circumstances. Whereas human-machine interfaces have been discussed for decades (Amershi et al., 2019), the need for explainability has often been an afterthought when research into AI systems was mostly concerned with advancing the technologies to a point where they could solve relevant real-world tasks (Rai, 2020).

More fundamentally, technology adoption has been a subject of research since the large-scale introduction of computers to private businesses. Technology adoption is viewed as a key factor for staying competitive in the market; however, the adoption of new technologies is always linked to costs, making usefulness and actual usage within a company critical metrics. Davis developed the now widely accepted Technology Acceptance Model (TAM), linking ease of use and perceived usefulness to current and self-predicted future usage (Davis, 1989). Subsequent research has identified additional predictors and discerned differences in impact across various technologies and applied fields. One factor often recognized as critical by the reviewed literature, especially for AI-based technologies, is trust in the new technology (Glikson and Woolley, 2020).

Human stakeholders indicate tendency of scepticism or resistance to AI recommendations, particularly in high-stakes contexts where potential errors could result in significant harm or financial losses (Glikson and Woolley, 2020; Kostopoulos, Davrazos, and Kotsiantis, 2024). Prior research identifies trust as a key factor underlying such hesitancy, alongside related concerns such as transparency, accountability, and perceived control. Contributing to this scepticism is the inherent opacity of many AI models, often described as "black-box" systems, which obfuscate the underlying reasoning behind their outputs (Guidotti et al., 2018). This lack of transparency can erode confidence in AI-driven decisions and hinder their broader adoption (Rai, 2020). For these reasons,

* Corresponding author.
E-mail addresses: michael.heider@uni-a.de (M. Heider), marcus1.albrecht@uni-a.de (M. Albrecht), johannes.schilp@uni-a.de (J. Schilp), joerg.haehner@uni-a.de (J. Hähner).

https://doi.org/10.1016/j.eswa.2026.131352
Received 29 August 2025; Received in revised form 16 January 2026; Accepted 24 January 2026
Available online 14 February 2026

the adoption of AI technologies has increasingly been studied within the framework of TAM. Baroni et al. expanded the model by introducing AI Output Trust and AI Output Quality alongside XAI-related latent variables. Their empirical evaluation demonstrates that explainability-related factors are positively associated with perceived ease of use, and that trust in AI outputs significantly influences perceived usefulness, underscoring the central role of XAI-related perceptions in AI acceptance (Baroni, Calegari, Scandolari, and Celino, 2022).

To mitigate these concerns, there has been a substantial focus on the development and implementation of explainable AI (XAI) techniques. XAI seeks to enhance the interpretability of AI models and provide human users with clear, understandable explanations for system outputs (Doshi-Velez and Kim, 2017; Zhou et al., 2024). By bridging the gap between complex algorithmic reasoning and human comprehension, XAI fosters greater trust, improves decision-making collaboration, and facilitates safer and more responsible AI deployment in high-impact domains (Barredo Arrieta et al., 2020).[1]

In this article, we present a case study about the requirements for an implementation of (X)AI for which we prompted a group of production planners ($n = 11$)[2] from a large multinational manufacturing company with multiple plants throughout Europe and world-wide, although all interviewees were from Germany and German speakers. These will soon be assisted by more sophisticated algorithmic approaches to turn the currently heavily manually assisted production planning into an intelligent setup of smart production where AI, in the form of stochastic optimization techniques like evolutionary algorithms, will perform the optimization of a variety of different challenges, from the domain of planning and scheduling, currently solved or finalized in time-consuming manual processes (Wittmeir et al., 2023).[3]

In early talks (pre-study), we found that there are some additional levels of information necessary beyond presenting them with just the solution of their current task, i.e. a new or fixed production plan. From this, we derived a questionnaire to find what specifically they want (or do not want or need). Our hypothesis (which is described in detail in Section 4 as the goal of our study) was that they do not want to understand the algorithmic details yet have not only a broad general interest but rather need a level of detail usually unavailable about the proposed solution(s) and the path to get there, e.g. what parts of the search space were explored.

We found (cf. Section 5) that the stakeholders require some form of XAI / explainable optimization and this case study demonstrates their exact requests based on our comprehensive questionnaire. In essence, they expect a user-centric approach (in line with general industry 5.0 trends) that explains the assumptions made about the task, the general optimization method, the resulting search space coverage (including the reasoning for a specific search distribution), the validity of proposed solutions, and key performance indicators that individuals from a set of proposed solutions can achieve. Crucially, due to the nature of this agent (not utilizing a machine learning (ML) model) many existing methods and terminologies are not applicable. There are also no notable studies that include non-expert users and focus on their needs with regards to explainable optimization. We hope that our insights raise interest in closing this gap as we strongly think that the problem at hand is relevant for real-world applications and solvable with reasonable combined efforts.

The remainder of this article presents the context, methodology, and key insights from our current case study where we inquired with the stakeholder group most affected by an AI-based assistance system about whether they need or want this type of assistance and what they wish the interactions with it to look like. Section 2 presents some similar studies and context from the field of evolutionary computation (which is where the implemented AI approaches will most likely be from). The general setting in which we aim to employ the new assistance system is described in Section 3. Section 4 introduces the methodology of our case study and the general design of our questionnaire which can be found in Appendix A. In Section 5, we lay out the key findings especially regarding the clearly expressed desire for explainable AI and some details regarding what this could or should look like for the two different applications of an AI-based smart production planning approach highlighted above. While these results are not necessarily generalizable they may generate valuable hypotheses and design insights for subsequent research. Finally, Section 6 draws some conclusions from our data and gives a look ahead into potential next steps.

2. Related work

Explainability of evolutionary computation (EC) has gained more interest in the last few years (cf. Niki van Stein, 2025; Zhou et al., 2024) but is not entirely new either. Especially for making explanations to researchers and other technical experts, there has been a long history of research, e.g. regarding the fitness landscape these algorithms traverse (Stadler, 2002) or analysing their search behaviour (Fyvie et al., 2023; Ochoa et al., 2021; Stegherr et al., 2023). XAI can help in optimizing more effectively, e.g. by guiding the exploration based on variable importance (Hunter et al., 2025). EC has also successfully been used to create explainable ML models, e.g. in the fields of Learning Classifier Systems (Heider et al., 2023a; Urbanowicz and Moore, 2009) or Genetic Programming (Mei et al., 2023).

When stakeholders from non-technical backgrounds become involved that are not familiar with these algorithms, explainability of EC faces new challenges (Rodemann and Attig, 2025). Here, it becomes critical to assess whether the users specifically care about the optimization process of these algorithms. In some cases, e.g. (Heider et al., 2023b), stakeholders might not care about explanations of the optimization process. However, it is important to assess if they do and to what extent. Vincalek, Walton, and Evans (Vincalek et al., 2021) have also used a questionnaire to determine why GAs are not commonly used in industry and what design requirements different stakeholder groups might have for them to be considered.

Established theoretical frameworks for technology adoption, most notably TAM (Davis, 1989), have long emphasized that perceived usefulness and ease of use are primary drivers of system integration. In the context of industrial, high-stakes decision-making, these factors are inextricably linked to trust. As Glikson and Woolley (Glikson and Woolley, 2020) argue, trust in AI is a multi-faceted construct that evolves through interaction. Recent adaptations of TAM, such as the AI-TAM model (Baroni, Calegari, Scandolari, and Celino, 2022), specifically highlight that the quality and transparency of AI outputs (i.e., explainability) are critical antecedents to both perceived ease of use and trust, which in turn dictate the actual adoption of AI tools in professional environments.

Explainability in production planning has increasingly been recognized as an important factor for the successful introduction of AI-based

[1] For specific definitions of key terms from the field of XAI, such as explainability and trust, we direct the reader to Doshi-Velez and Kim (2017), Barredo Arrieta et al. (2020), and Zhou et al. (2024) that provide an excellent overview over the most critical terms, issues, and developments.

[2] While $n = 11$ is of course a number of respondents insufficient for general claims and quantitative statistical analysis, this corresponds to the full team of planners and thus all perspectives that can be gained without different companies which in-turn would blur any analysis due to differences in their methods and setups related to different fields and specializations. Importantly, we assume that focussing on the affected directly (rather than across multiple companies) greatly improves the acceptance and trust that can be gained by XAI. We encourage other researchers to implement similar studies with their domain experts and to share these insights so that the community can construct a more broad perspective and derive more general systems from that.

[3] Evolutionary algorithms are a key method in this setting for two reasons: 1. They do not have to finish computations to provide usable solutions, i.e. even with a fixed computation budget they can still optimize complex tasks even if not to an optimal solution. 2. Many of the optimization tasks are actually multi-objective and cannot easily be expressed in scalarized functions and evolutionary algorithms are by far the state-of-the-art for approximating a Pareto-front.

decision-support systems in domains where such technologies were previously limited. Baum, Baum, Gros, and Wolf (2023) discuss the potential benefits of explainability for fostering trust, supporting knowledge transfer, and strengthening decision-making processes, but also explicitly note the need for further studies involving industrial decision makers and stakeholders. This highlights that, although explainability is widely regarded as relevant for production planning, empirical insights from real-world industrial settings remain scarce.

Across industrial contexts more broadly, various application areas for AI and XAI have been identified. Ahmed, Jeon, and Piccialli (2022) provide an overview of AI and XAI within Industry 4.0, outlining several decision-support use cases in manufacturing environments. Their analysis, however, is primarily conceptual and technological, and does not focus on the explainability needs of practitioners in day-to-day operational planning. Similarly, Kostopoulos, Davrazos, and Kotsiantis (2024) present a structured overview of XAI-enabled decision-support systems, including examples in manufacturing. Their work illustrates the breadth of potential applications and the perceived benefits of explainability. A detailed analysis on the effects of adding XAI to these applications is, however, not given.

Complementing these broader perspectives, more technical contributions propose explanation methods tailored to specific model classes. Theumer, Edenhofner, Zimmermann, and Zipfel (2022), for example, demonstrate the use of feature-based force plots for explaining neural-network outputs in an industrial context. Their work shows how explanations can support system developers in understanding model behavior and improving configuration choices. While expected benefits regarding user acceptance are mentioned, these aspects are not the primary focus and are not examined empirically.

Studies involving a larger number of participants and explicitly evaluating the impact of XAI on user behaviour have so far primarily been conducted in other domains, non-adjacent to industrial production. For example, Papenmeier, Kern, Englebienne, and Seifert (2022) examine explanation effects in a social-media monitoring scenario and report dependencies between classifier accuracy and explanation usefulness. Such findings can inform general XAI research, but the decision context, stakes, and accountability structures differ substantially from those in production planning where planning decisions directly affect material flows, deadlines, and operational robustness.

The existing literature shows growing interest in (X)AI for industrial decision-support, conceptual analyses of its benefits, and technical demonstrations of explanation methods. However, systematic studies involving industrial production planners and other stakeholders responsible for day-to-day planning decisions remain limited. Similarly, studies that involve explanations of stochastic optimization are missing. With decision support systems that can rely on optimizers that guarantee optimal solutions but might exceed the available computational budgets in cases similar to ours (cf. Section 3), our questions are relevant than in the case of fixed-budget optimization wear near-optimal solutions are the only result that can reliably be expected. However, with increasing complexity in the tasks expected to optimized, these methods will increasingly be needed and therefore we must find ways to make them deployment-ready by being trustworthy and comprehensive. The present study addresses this gap by directly eliciting explainability requirements from practitioners within the production-planning domain of a large multinational manufacturing company. By focusing specifically on their needs, expectations, and preferences across two concrete planning scenarios, this work contributes empirical insights that complement the conceptual and technical perspectives provided in the literature.

3. Industrial setting for an AI-based assistance system

As mentioned previously, a full team of production planners (a total of 11) for German production sites of a large multinational production company participated in our study. They are currently assisted by a simplistic metaheuristic-based planning algorithm that is neither close to state-of-the-art nor tailored to the use case at hand. Due to its simplicity, this agent has only limited capabilities which leads to ineffective solutions for the tasks it can be used for but also to many manual planning problems, as the agent can not be applied to all day-to-day tasks. The main optimization problem it is currently used for is planning the job order for the (next day to enter the) frozen zone (varies from three to ten days depending on plant). However, manual adjustments of its solutions are needed every day. Due to this, a replacement of this agent is planned for the near future. The replacement will utilize advanced techniques from the field of evolutionary computation (EC) as this is a hard black-box optimization problem for which we can not expect to find optimal solutions in reasonable time but where techniques from EC, e.g. modern genetic algorithms or potentially state-of-the-art ant-based optimization methods, can find good solutions within the available computation budget.

For the past years, we have already assisted this company to solve another task that planners are regularly faced with and have to solve manually as fast as possible (in the following text also called Scenario A): When a disturbance occurs in logistics, i.e. a scheduled delivery will not arrive on-time even though it is necessary to complete the planned jobs (just-in-time logistics), the production plan (jobs to manufacture) has to be readjusted. Here, a large number of soft and hard constraints as well as different performance metrics have to be taken into account: available parts, product inventory balance, personnel, setup costs, schedule nervousness, intra-factory logistics, and many other aspects. A detailed description of this use case and one of our proposed agents, that can fully solve this task, can be found in Pleier et al. (2025), Wittmeir et al. (2023). Besides the specific agent introduced in Pleier et al. (2025), experiments with different types of evolutionary algorithms have also shown good results for this task but are ultimately not needed as the true optimum can be found in acceptable time with a CP-SAT solver.

Importantly, from our pilot studies and day-to-day interactions with affected stakeholders, we realized that due to the sheer number of hard and soft constraints and the often complex interactions between them, it is not immediately obvious to users why a particular replanned schedule is proposed and why it might be optimal under the taken assumptions. Critically, due to incomplete knowledge within existing data bases, these assumptions might not be correct in some cases even if they are applicable in a majority of past situations. To support planners' intuition and understanding of the solution towards making an informed decision about usage or modification of a proposed schedule, explanations have to, for example, highlight which hard constraints excluded certain candidate jobs or sequences; domain experts can typically verify such constraints quickly. Soft constraints and preference-oriented objectives are much harder to assess. Feature importance attribution methods such as SHAP (SHapley Additive exPlanations) have proven effective (cf. for example Hunter et al., 2025) in assigning local importance scores (based on a trained surrogate regression model) to features the optimizer directly operates on, offering an interpretable view on why the solver or optimizer favoured exploring certain search space dimensions. However, in the scheduling scenarios we are currently examining, a surrogate model is not easily available. While it might be possible to bootstrap a sufficiently accurate model with future gathered data, it is not available as of now or the near future.

Additionally, explanations such as SHAP and most other techniques currently used within the XAI community are not easy to interpret even for experts in these methods. In many cases, it might be impossible for non-experts, especially those without a formal education in statistics or data science, to use such metrics effectively. From limited testing in pilot studies with our partners, we can assume that our planners would require extensive further education to benefit from such methods which is unlikely to be feasible within budgetary requirements and their time constraints of having to perform the tasks at hand. An additional factor that exacerbates issues of usability of many techniques from the field, is the focus on ML applications rather than optimization. While there are some works on the explainability of optimization specifically (Niki van Stein, 2025; Zhou et al., 2024), the vast majority is concerned with ML models.

When implementing schedule repair and optimizers with more complex tasks, e.g. for full schedule construction, later Scenario B, we must adhere to users' backgrounds, individual requests, and possibly even hard requirements stated by them about explanations to see the adoption of this system. Especially with hard requirements, we also need to consider other stakeholder levels, e.g. time allotments allowed by management for further education/training and later on for interacting with the explanations. Obviously, the explanations should be at a brevity where checking them takes substantially less time than completing the optimization task manually. Without explanations stakeholders where quite clear that they would recheck the validity of the proposed solution(s) in each instance or, even worse, ignore them altogether. In the presently implemented system, only very limited explanations about key KPIs were available. This lead to scepticism among planners which, in-turn, motivated us to apply a more formal approach at gathering insights into how we can reduce this scepticism, the results of which are presented in this article.

4. Study goals and design

With our exploratory user study, the primary goal was to assess the attitudes of the most affected stakeholders towards a smart(er) production planning system.[4] This can be split into multiple objectives:

1. Do stakeholders believe in the capability of AI to assist them in the real-world?
2. Would stakeholders participate willingly and take an AI's advice?
3. Would stakeholders trust the AI solutions blindly or do they request explanations?
4. If explanations are needed, what form would production planners prefer and in which cases are they least/most important?
5. Do we "only" have to explain the result the AI returns or also the way in which it arrived there?
6. Is there a difference between different tasks / complexity levels in these questions?

Additionally, we wanted to learn about experiences with smart technologies and AI beyond what the stakeholders know from their work, whether they feel like AI implementation may threaten their job security, and whether they assume that smarter systems are a competitive advantage that might secure their plant's location for the foreseeable future.

These questions arose from previous works in similar fields (especially with regards to ML-based decision support systems Heider et al., 2021, 2023b) and years of cooperative work regarding an improvement to the optimization algorithms that are currently used within the company. We validated the general objectives beforehand in discussions with relevant stakeholders from within and outside the company in question. Most importantly, we discussed the objectives with management as well as a senior team member. We also ran a small pilot round of the questionnaire with data scientists involved in the development project that were asked to answer in a way they expect their production planner colleagues to respond. We did not run a full pilot study with our production planners due to their small number where we did not want to bias the results by questioning the same person in both studies.

Comparing the results we found that the data scientists had a good idea about their colleagues needs even if they expressed that they did not agree with all questions (they would prefer statistical assurances rather than instance-based measures whereas planners have no experience interpreting such statistical models).

From this set of overarching questions, we constructed a questionnaire of 31 questions with a mix of single-choice, multiple-choice, and Likert-item answers, although we provided explicit write-in field, where we deemed them appropriate. A translated version of these questions and the answer options is shown in Appendix A. Within the questionnaire we distinguished between

- general questions,
- questions regarding Scenario A (a job scheduled within the frozen zone can not be executed due to a logistic disturbance, cf. Pleier et al., 2025),
- the identical questions for Scenario B (daily run to plan future job orders and especially the next day to enter the frozen zone),
- questions on the importance and frequency of explainability,
- questions about the human-machine interface of an XAI approach, and
- two concluding questions about individual job security and plant location security.

The questionnaire was answered in a live setting, with researchers from within the company and the university present and capable of answering any upcoming ambiguities in the questions. Before it was handed out, both groups gave a quick summary of the purpose of this study and a detailed description of Scenarios A and B to clarify what elements of the stakeholders' day-to-day tasks were identified. Additionally, we stressed that participation is voluntary and that none of the filled-out questionnaires (or any information that could lead to the identification of individual participants) would be provided to any company employees, especially not any managers of the participating production planners. While this cannot eliminate bias within responses, we assume, based on prior interactions, that at least no respondents felt that being pro or contra explainability would in any way be expected of them. A stronger bias control would have been desired but was not attainable within the sample size.

Due to the specific relatively wide-reaching interests that should be covered by the study and to insure participation within the timeframe allotted by management (at most 45 min including all questions and introductions made in advance), we had to keep the number of questions within a reasonable limit. Additionally, we knew the number of possible participants would be limited. In the end, we were able to collect answers from all production planners within the company that are currently active. As this number cannot yield quantitative results or results that are fit for many (or even most) statistical methods, we designed our questions in a way that should yield the most impactful overview. These two reasons led us to mix different types of answer scales. For example, rather than using Likert-item answers for all questions, including individually asking about specific categorical options, we used 5-point Likert-items where we really needed differentiated answers but only single-/multiple-choice categoricals where we did expect users to have binary rather than scaling preferences. One example is question 25 where we asked users to check their most preferred way of receiving information rather than individually asking the users how much (from 1 to 5) they would like to receive "Bullet points", "Graphs", and "Tables". This makes a statistical analysis a bit more difficult but allowed us to probe the production planners with generally more and, critically, a more diverse set of questions inquiring about multiple aspects that are relevant. Such condensations of questions and the lack of "redundant" but reformulated (to check real sentiment rather than preception based on wording), which we also had to omit to keep within a manageable timeframe, are clear limitations of our study. However, we think that definitive answers to many of the raised points can only be given with specific implementations of each option and a deeper measurement

[4] For the abstraction level the non-technical stakeholders should interact with such a system, we simply convey it as "using AI techniques" whereas, internally, most algorithmic parts of this agent are optimization algorithms rather than LLMs which are currently almost synonymous to the term AI. To simplify the reading of this text, we will also just use "AI" or "optimizer" in place of "an agent that uses techniques from the field of AI especially stochastic optimization algorithms".

of the real-world effects of explanations provided in one of the different ways (e.g. similar to A/B-testing) and could not be provided with a questionnaire. Thus, while our study does not provide these final answers, it does provide researchers that face similar questions or are interested in developing the fundamentals to solve explainability of stochastic optimization with real-world non-expert users with a good and informed intuition on where a good starting point is. It helps with eliminating some of the less promising options and, crucially, shows that there even is a need for such research within this subfield.

5. Key insights

In total, eleven stakeholders that are currently involved in logistics and production planning answered the questionnaire fully. With these, we covered a wide range of age and experience levels, and the majority of the respective department. In this section, we present the general results and take some central insights especially regarding the diverse goals as laid out in Section 4.

Interestingly, on the Likert-item for self-assessment of job experience / competency (scale with five elements; inexperienced to expert) none of the participants went for the highest rating even though at least two participants had over 20 years of experience with one between 10 and 20 and 4 between 3 and 10 years in this position.

Eight of the eleven participants privately use the currently popular AI systems, e.g. ChatGPT, DALL-E, etc. without assistance while one respondent has used such systems with the help of family and/or friends. While the internal mechanisms governing these systems are vastly dissimilar from what would be used in the professional setting of smarter production, we assume that this experience still shapes a more positive or at least open-minded view among non-technical stakeholders than we encountered just a few years ago. At that time, AI was not yet widely recognized as a realistic option for real-world applications, and existing uses were often branded under different names in corporate marketing.

All of the participants expressed that AI will be able to assist effectively and that this is necessary in the future, although two assumed it would only be helpful in limited tasks while the other nine expected it to be helpful for many. On the question whether they could envision using the recommendations from an AI (multiple answers possible) only one expressed to rather ask a co-worker, two would especially use it when pressed for time, five would use it when stuck, and five would at least consider it in all circumstances.

Regarding Scenario A (fixing a job schedule after a disturbance in the logistics occurred), all expressed that their tasks could be made at least somewhat easier by an AI, although six expressed that they would not trust the new schedule if the AI would not provide additional information. The participants were unanimous that this additional information would not need to be a full white box model/approach that gives thorough explanations but that a good rating in the relevant objectives (metrics, KPIs)—including a description of the solution and an explanation of why this solution is rated that way—is sufficient. We also asked whether the participants would trust the AI to perform the task well if it had done so in the past, where no participant answered “yes”, one straight out rejected, three requested such solution descriptions in the future and seven requested additional explanations on the process of finding the solutions on top of the individual descriptions.

For Scenario B (planning the frozen zone and beyond), participants were even slightly more positive that this will ease their day-to-day tasks but this time eight expressed that they would not trust a solution without additional insights. Two participants expressed that they needed more than good ratings to accept the proposed solution but this time all participants would expect the AI to be trustworthy to perform well in the future if it did in the past, provided it would describe the solutions and the process.

To more specifically assess what type of information one should provide based on the optimization process, we asked a series of more specific questions: First, if the algorithm should express at which time the selected solution was found compared to the total computational budget, i.e. if the solution was discovered late during the process or rather early. This is essentially a question about convergence without having to explain the concept to a group of skilled workers that however do not know about metaheuristic or other optimization algorithms. Here, two wanted to have this information in all cases, eight would at least want to be notified if the solution was found close to the end and then potentially decide to add more computational budget. In a Likert-item on the importance of explanations why no new solutions were found (which is not easy to do with current XAI techniques from optimization), two expressed moderate, three high, and six very high interest (scale with five elements; labelled “irrelevant” to “integral”). Interest about the search space coverage was mixed with slightly higher interest for an explanation why only parts of the search space could be visited. Eight of eleven requested a detailed description of the internally used algorithms. For a step-by-step explanation of the process how the algorithms find the solution, the majority requested an example based on illustrations and abstract examples over very detailed overviews or even central milestones of each specific process for a joint task.

For the primary form of the interface, we gave multiple-choice options (including a write-in that was not utilized) between bullet points (5x), graphs (2x), or tables (8x). As part of the standard interface for explanations of solutions (multiple-choice), one should include assumptions about the problems made during the development phase (4x), (hard/soft) constraints (5x), or reasoning why a solution might be violating a constraint but should still be chosen (8x). With the same question (and answering options) about the path to the solution the answers were yes for 6, 3, and 6 participants, respectively. If there were multiple valid solutions one should display (multiple-choice): the best (2x), all (2x), the top three (7x), a diverse set regarding constraint handling (0x), or a diverse set regarding product types (3x). For the length of all information about the search process, nine participants expressed that it should be limited to a tooltip while two preferred half a page. For this question, one participant remarked that for documentation (and potential future debugging) purposes a full description should be logged but not shown.

Three participants see their individual job security acutely threatened while two expect this in the coming 10–15 years. For why participants do not expect this to be problematic, we offered a write-in field that was heavily utilized. The general picture assumes that humans are still better at the task, have knowledge inaccessible to their current computer systems, are needed for supervision, or simply that the number of available workers will drop due to demographic changes so saving a few individuals would even benecessary.

For the location security of the plant, we see a clearer picture with seven participants expecting that the introduction of such systems would grant competitive advantages for at least 10–15 years. One participant expressed that other companies will also introduce such system so that this essentially nullifies the effect but that everyone that would not integrate AI further would probably loose out.

While the answers of participants were generally not too divergent, we still wanted to analyse them in context of previous responses rather than solely on an individual question basis. First, we performed hierarchical clustering with different metrics visualized using dendrograms. Given the mixed types of data, we find Gower distance (Gower, 1971) to be the most expressive and appropriate metric over other common metrics like euclidean (which can be mislead as the scales might be ordinal but not equidistant) or Manhattan (which is not designed for categorical or ordinal data). In Fig. 1, we show one of these dendrograms and label the individual participants (branches) with the self-assessed job experience / competency level. Given the high levels of the response “4” interpretations should be made with a bit of caution but we can see that at least those that are very new at the job, fall closer together. The two “4” responders on the right side can also not be clustered together with the other four participants on that side by any of our other demographic markers or by the responses to AI experience.

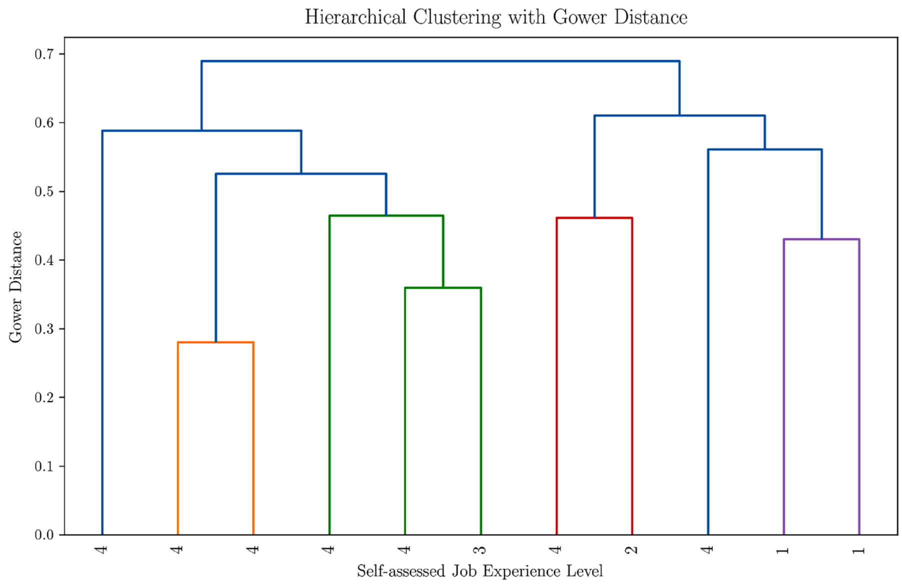

Fig. 1. Dendrogram plot of hierarchical clustering based on Gower's distance over all numerically assessed questionnaire responses. The y-axis shows Gower's distance at which clusters merge. Lower joints symbolize higher similarity between two individual sets of answers. Therefore, high joints express great dissimilarity. x-axis labels are self-assessed job experience ratings (scale 1–5). We can see that the most inexperienced respondents form another group while within the most experienced other factors are critical. This insight is crucial for the design as we can assume that any XAI approach will not become obsolete just because of a more experienced work force.

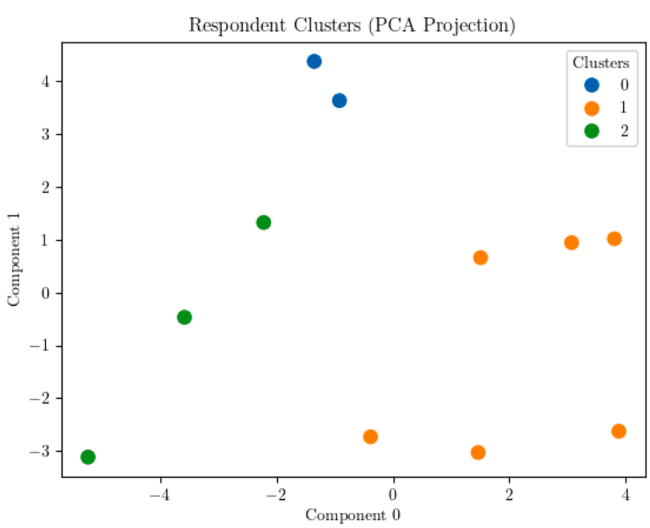

Fig. 2. Scatterplot of the individual respondents using a projection from the high dimensional answering space into a 2D representation using a Principle Component Analysis. Highlighted in colour are the results of clustering the original responses with the k-Means algorithm with $k = 3$. The PCA allows an indication of general respondent distance and is thus giving an additional perspective to the one in Fig. 1. Similarly, the k-Means clustering indicates that some moderate grouping of the answers is possible.

In Fig. 2, we show another perspective on the responses. First, we cluster the responses using k-Means with $k = 3$.[5] Then, we apply a Principal Component Analysis (PCA) with two components (two used for better visualisation options). We find that three clusters emerge, which somewhat corroborates Fig. 1, but that generally the distances between responses are relatively even for many.

To assess correlations between specific questions / answers to these, we compute the pair-wise Kendall's τ correlation between each option, cf. Fig. 3.[6] Multiple-choice questions are encoded into bits while all other questions use ordinal encodings. Generally, the correlations follow the expected similarities, e.g.: age correlates with job experience and the time in the company; importance for explanations about convergence correlates with the importance for search space coverage explanations and explanations for the search process (20, 21, and 22). We can also see that users that responded that they would prompt the AI for help if they are stuck themselves (encoded as 8_2 in Fig. 3) would not imagine using an AI recommendation (8_1) every time. There are also some very likely non-causal correlations visible, e.g. responding "bullet points" for additional information or explanations (25_1) inversely correlates to the answer on whether one feels threatened by AI or assumes these to be beneficial for the plant location security (30 and 31).

Generally, clear overarching design decisions for the user interface cannot yet be made. However, we have a much clearer picture on what we should test and what to exclude and what should be implemented first (see the previous paragraphs for specifics). Ideally, after testing a solution in application, we would find that it is already sufficient and would not need to test for example graphical representations of additional information such as soft constraint violations. Most likely, showing information in condensed tabular form and short expression bullet points in tooltip length covers the majority of knowledge our stakeholders need about the solutions. This result-based perspective should be aided by details about convergence and feature space coverage of the search process. The AI / optimizer would thus need to track enough data to provide comprehensive be-

[5] $k = 2$ fails to cluster results at all in a meaningful way while $k = 4$ does make an irrelevant vertical split in Cluster 1 (bottom right; yellow).

[6] We use Kendall's τ over Spearman or other common correlation coefficients as our scales are not equal-distanced / non uniformly-spaced, i.e. when answering "Do you believe that AI could support problem-solving in production planning" (Q7) the assumption that the distance between "For many tasks" and "Only for a few narrowly defined tasks" is the same as the distance between "Only for a few narrowly defined tasks" and "It [AI] is not necessary" is not obviously correct. On the contrary, we should assume that the distance is substantially different. This is the case with many of our questions / answering options.

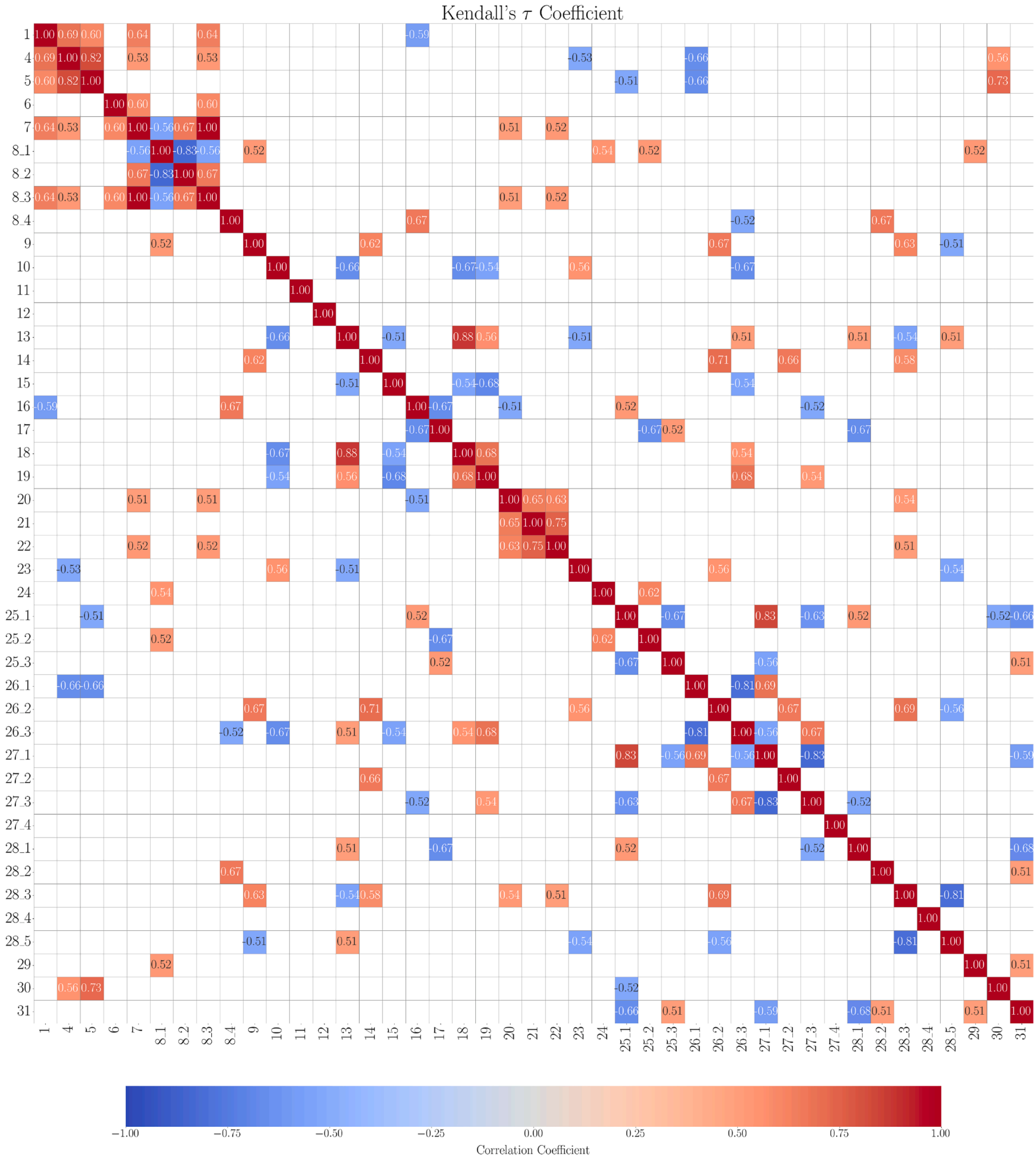

Fig. 3. Matrix-Heatmap of Kendal-Tau-correlation for the bitwise encoded answers to the questionnaire. To improve the readability, questions with check-boxes are encoded with “true” for checked and false for unchecked. For multiple choice questions, we create bit-wise encodings, where checking one response is translated to True, i.e. checking the first and second option for question 8 would result in *8_1* and *8_2* to be True. We only show moderate to strong correlations (greater 0.5) and use a heatmap colouring for these.

haviour analysis measures and compute and convey these metrics in a form that can be easily understood by non-experts, e.g. it should be easy to judge whether a run went well or not without much ambiguity. Current options for behaviour analysis of evolutionary algorithms, especially well-developed and advanced options favoured by the research community such as for example (Cleghorn, 2021; Korošec, 2020; Ochoa et al., 2021; Stegherr, Heider, and Hähner, 2025), are difficult to interact with for standard users of other backgrounds and will only serve as explanations or visualisation tools for data scientists involved. Explainable design decisions for both the interface as well as the optimizer in cases like ours are still a fundamental research gap.

In total, our results present a picture that our stakeholders are interested in moderate levels of explainability in their AI-based assistance system but need a user-centric design approach to trust the system. They do not require or even request full explanations of individual steps the optimizer takes (and would not have the time to analyse these on a daily basis) but they are interested in the selected solution(s) and why these were selected. This is a very critical insight for those designing optimizers for the real-world application as it more clearly defined the limits for how to apply XAI within optimization (Bacardit et al., 2022; Zhou et al., 2024). They also expressed that information about convergence and if more computation budget might yield better results is important to them. This is especially critical because we might not be able to generate such explanations using the CP-SAT solver used for Scenario A in Pleier et al. (2025). While we can guarantee an optimal solution in Scenario A now, it might be relevant to investigate EC options here as well, should the optimization task become more difficult in the future, e.g. because additional aspects that are currently solved by hand in other departments or currently deemed to rigid to change on the fly (e.g. pre-assembly in-house production) might become adjustable. That stakeholders do not request full information about the search process is quite advantageous for both EC-based and CP-SAT solvers. Especially within EC, where a lot of decisions are based on stochastic processes and in parallel, this could be quite complex to explain to non-technical staff. While there were some difference between individuals, the general trend is clear. Interestingly, the notion that older employees will be substantially more sceptical and the youngest employees will almost trust the AI-based recommendation system blindly, which is a common assumption made by management, could not be verified based on the answers. We can also see that users have not only interest in using AI more but also see it as advantageous for their companies stability and mid-term prospects without being afraid of a more widespread implementation.

6. Conclusion and future work

In this article, we presented a real-world case study into what the users at a large multinational company need in order to trust "AI" used to make the production planning process smarter. This is motivated by the clear insight that working towards smart(er) production planning, necessarily involves the inclusion of more and more sophisticated AI techniques. However, in many cases, humans will still ultimately be responsible for the efficient flow of production. Therefore, humans will have to use the advice from AI and make decisions accordingly. Human stakeholders only follow this advice when they deem it trustworthy. While this trust in an intelligent but black-box system could potentially be build over time, it is probably much more efficient if humans actually understand why they are giving a certain recommendation.

Based on our case study, we presented two different scenarios where such intelligent agents will be used in the immediate future based on optimization of schedules that are easily disturbed due to just-in-time delivery setups. Both of these can be solved with techniques from the field of evolutionary computation (EC), although one of our scenarios is—in its current slightly limited configuration—also optimizable in sufficient time using a SAT solver. After a short briefing about the two scenarios where the planners will be assisted by AI in the future, we had different production planners from all age and experience groups ($n = 11$) fill out an extensive questionnaire regarding their wishes for the information an AI system should provide for its solutions to be trustable.

We found that the participants can envision cooperative work with an AI-based agent that advises them and generally expect it to be useful in many circumstances (no participant deemed this unnecessary entirely). The participants would also not trust the AI blindly and use its recommendations without scrutiny. They would at least want an extensive explanation of why this new production plan was the result of the optimization process but generally do not require full information about the search process. Depending on the impact of the scenario, we also see slight differences in the level of explanations requested, with the more complex but less time-critical scenario being the one that generally leads to stronger requirements of explainability. Users are interested in general notions about the search space coverage and whether convergence occurred but do not wish to have detailed information. For the presentation of explanations, no clear style could be found (which means that this comes down to user preference) beyond that it should not take up too much space in any human machine interface and be used as additional information to confirm the solution rather than the main output of the AI-based agent. The participants also stated that they are generally not expecting AI to threaten their individual job security but strongly suspect that it will increase the location security of their plants by giving them a competitive advantage.

This creates a number of avenues for the designer of such a system where they do not need to provide full information but have to use a targeted approach. Crucially, existing methods for explainable EC are not clearly applicable here identifying a potential gap in research when it comes to "easier" explanations and visualizations. The integration of explanations is, however, crucial for the acceptance and therefore required if managers want their workers to comfortably rely on these systems. This thus warrants more implementation and practical studies as well as further theoretical works. Interdisciplinary approaches with psychologists and UX experts should be especially valuable here.

As our immediate next steps we will roll out the already implemented optimizers and test the daily interaction with the stakeholders. For this, we will integrate domain knowledge about the optimization task to explain the results of the optimization process based on constraint handling and a variety of defined metrics to assess the optimality of the solution. Here, we will integrate user feedback in quick iteration cycles to determine the level of requested insights more precisely and assess if interacting with the agents every day changes the perception. Most likely, this step will spark some more research into the explainability of the approaches and their recommendations as well and we hope to be able to demonstrate the effectiveness of some of the many approaches from the field of EC (Zhou et al., 2024) in the real-world. We want to specifically call for other researchers working on real-world optimization problems to consider whether the adoption of their methods would be more likely and whether the user experience would be improved when explainability was integrated into their optimizers.

Overall, we found that for implementing smart production with the help of modern AI (or more specifically, EC) explainability of the solution is without doubt a hard requirement and any successful system would therefore have to employ a user-driven interpretable optimization design rather than being fully black-box.

CRediT authorship contribution statement

Michael Heider: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Visualization; **Marcus Albrecht:** Investigation, Project administration; **Johannes Schilp:** Project administration, Funding acquisition; **Jörg Hähner:** Project administration, Funding acquisition.

Data availability

The data that has been used is confidential.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Michael Heider reports financial support was provided by Bavarian Ministry of Economic Affairs Regional Development and Energy. Marcus Albrecht reports financial support was provided by Bavarian Ministry of Economic Affairs Regional Development and Energy. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been funded by the Bavarian Ministry of Economic Affairs, Regional Development, and Energy.

Appendix A. Questionnaire for the case study

The supplemental material shows the questionnaire that was handed out and filled out on paper after an introductory briefing as described in Section 4. For the publication, we translated the individual questions into English to make it more accessible to a wider audience but will make the original German version available on request. Also, we removed the name of the company as well as the respective plant locations to keep in line with our NDAs and publication agreements.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.eswa.2026.131352.

References

Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where. *18*(8), 5031–5042. https://doi.org/10.1109/TII.2022.3146552

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Cakmak, M., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233

Bacardit, J., Brownlee, A. E. I., Cagnoni, S., Iacca, G., McCall, J., & Walker, D. (2022). The intersection of evolutionary computation and explainable AI. In *Proceedings of the genetic and evolutionary computation conference companion*. ACM. https://doi.org/10.1145/3520304.3533974

Baroni, I., Calegari, G. R., Scandolari, D., & Celino, I., et al. (2022). AI-TAM: A model to investigate user acceptance and collaborative intention in human-in-the-loop AI applications . *9*(1), 1–21. https://doi.org/10.15346/hc.v9i1.134

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Baum, D., Baum, K., Gros, T. P., & Wolf, V. (2023). XAI requirements in smart production processes: A case study. In L. Longo (Ed.), *Explainable artificial intelligence* (pp. 3–24). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44064-9_1

Cleghorn, C. W., & Ochoa, G. (2021). Understanding parameter spaces using local optima networks: A case study on particle swarm optimization. In *Proceedings of the genetic and evolutionary computation conference companion* GECCO '21. ACM. https://doi.org/10.1145/3449726.3463145

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *13*(3), 319–340. Publisher: Management Information Systems Research Center, University of Minnesota. https://doi.org/10.2307/249008

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608.

Fyvie, M., Mccall, J., Christie, L., & Brownlee, A. (2023). Explaining a staff rostering genetic algorithm using sensitivity analysis and trajectory analysis. In *Proceedings of the companion conference on genetic and evolutionary computation* GECCO '23 Companion (pp. 1648–1656). ACM. https://doi.org/10.1145/3583133.3596353

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), (pp. 857–871).

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Heider, M., Nordsieck, R., & Hähner, J. (2021). Learning classifier systems for self-explaining socio-technical-systems. In A. Stein, S. Tomforde, J. Botev, & P. Lewis (Eds.), *Proceedings of LIFELIKE 2021 co-located with 2021 conference on artificial life (ALIFE 2021)*. http://ceur-ws.org/Vol-3007/.

Heider, M., Pätzel, D., Stegherr, H., & Hähner, J. (2023a). A metaheuristic perspective on learning classifier systems. In M. Eddaly, B. Jarboui, & P. Siarry (Eds.), *Metaheuristics for machine learning: New advances and tools* (pp. 73–98). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-3888-7_3

Heider, M., Stegherr, H., Nordsieck, R., & Hähner, J. (2023b). Assessing model requirements for explainable AI: A template and exemplary case study. *Artificial Life*, *29*(4), 468–486. https://doi.org/10.1162/artl_a_00414

Hunter, K., Thomson, S. L., & Hart, E. (2025). Into the black box: Mining variable importance with XAI. In P. García-Sánchez, E. Hart, & S. L. Thomson (Eds.), *Applications of evolutionary computation - 28th European conference, evoapplications 2025, held as part of evostar 2025, Trieste, Italy, April 23–25, 2025, proceedings, Part II* (pp. 20–35). Springer (*vol. 15613*). Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-90065-5_2

Korošec, P., & Eftimov, T. (2020). Insights into exploration and exploitation power of optimization algorithm using DSCTool. *8*(9), 1474. https://doi.org/10.3390/math8091474

Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable artificial intelligence-based decision support systems: A recent review. *13*(14), 2842. Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/electronics13142842

Mei, Y., Chen, Q., Lensen, A., Xue, B., & Zhang, M. (2023). Explainable artificial intelligence by genetic programming: A survey. *IEEE Transactions on Evolutionary Computation*, *27*(3), 621–641. https://doi.org/10.1109/TEVC.2022.3225509

Ochoa, G., Malan, K. M., & Blum, C. (2021). Search trajectory networks: A tool for analysing and visualising the behaviour of metaheuristics. *Applied Soft Computing*, *109*, 107492. https://doi.org/10.1016/j.asoc.2021.107492

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *29*(4), 35:1–35:33. https://doi.org/10.1145/3495013

Pleier, F., Wurth, J., Gerling, V., Albrecht, M., Heider, M., Hähner, J., & Schilp, J. (2025). Towards automated rescheduling: A framework for extracting domain knowledge in production planning and control. In *19th CIRP conference on intelligent computation in manufacturing engineering*. Gulf of Naples, Italy: Elsevier B.V. in press. https://doi.org/10.1016/j.promfg.2025.05.033.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rodemann, T., & Attig, C. (2025). Real application challenges in evolutionary optimization? People! In *EvoApplications 2025* , pp. 469–481). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-90062-4_29

Stadler, P. F. (2002). Fitness landscapes. In *Biological evolution and statistical physics* , pp. 183–204). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45692-9_10

Stegherr, H., Heider, M., & Hähner, J. (2023). Assisting convergence behaviour characterisation with unsupervised clustering. In N. van Stein, F. Marcelloni, H. K. Lam, M. Cottrell, & J. Filipe (Eds.), *Proceedings of the 15th international joint conference on computational intelligence, November 13–15, 2023, in Rome, Italy* (pp. 108 – 118). https://doi.org/10.5220/0012202100003595

Stegherr, H., Heider, M., & Hähner, J. (2025). A comparison of clustering approaches for metaheuristic behaviour data. In *Computational intelligence* , pp. 305–322). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-85252-7_17

Niki van Stein, A. V. K. (Ed.) (2025). Explainable AI for evolutionary computation. Springer Nature Singapore. https://doi.org/10.1007/978-981-96-2540-6

Theumer, P., Edenhofner, F., Zimmermann, R., & Zipfel, A. (2022). Explainable deep reinforcement learning for production control. (pp. 809–818). Book Title: Proceedings of the Conference on Production Systems and Logistics: CPSL 2022 Publisher: Hannover: publish-Ing. https://doi.org/10.15488/12158

Urbanowicz, R. J., & Moore, J. H. (2009). Learning Classifier Systems: A complete introduction, review, and roadmap. *Journal of Artificial Evolution and Applications*, *2009*(1), 25.

Vincalek, J., Walton, S., & Evans, B. (2021). It's the journey not the destination: Building genetic algorithms practitioners can trust. In *Proceedings of the genetic and evolutionary computation conference companion* GECCO '21 (pp. 231–232). ACM. https://doi.org/10.1145/3449726.3459483

Wittmeir, T., Heider, M., Schweiger, A., Krä, M., Hähner, J., Schilp, J., & Berlak, J. (2023). Towards robustness of production planning and control against supply chain disruptions. In D. Herberger, M. Hübner, & V. Stich (Eds.), *Proceedings of the conference on production systems and logistics: CPSL 2023* (pp. 65–75). Hannover: publish-Ing. https://doi.org/10.15488/13425

Zhou, R., Bacardit, J., Brownlee, A. E. I., Cagnoni, S., Fyvie, M., Iacca, G., McCall, J., van Stein, N., Walker, D. J., & Hu, T. (2024). Evolutionary computation and explainable AI: A roadmap to understandable intelligent systems. *IEEE Transactions on Evolutionary Computation*. https://doi.org/10.1109/TEVC.2024.3476443