




# Towards Trustworthy AI in Demand Planning: Defining Explainability for Supply Chain Management

Ruiqi Zhu<sup>1</sup>, Cecilie Christensen<sup>1</sup>, Bahram Zarrin<sup>2</sup><sup>a</sup>, Per Bækgaard<sup>1</sup><sup>b</sup>  
and Tommy Sonne Alstrøm<sup>1</sup><sup>c</sup>

<sup>1</sup>Technical University of Denmark, Kongens Lyngby, Denmark

<sup>2</sup>Microsoft Reserach Hub, Kongens Lyngby, Denmark

**Keywords:** Explainable AI, Supply Chain Management, Demand Planning, User-Centric Explainability.

**Abstract:** Artificial intelligence is increasingly essential in supply chain management, where machine learning models improve demand forecasting accuracy. However, as AI usage expands, so does the complexity and opacity of predictive models. Given the significant impact on operations, it is crucial for demand planners to trust these forecasts and the decisions derived from them, highlighting the need for explainability. This paper reviews prominent definitions of explainability in AI and proposes a tailored definition of explainability for supply chain management. By using a user-centric approach, we address the practical needs of definitions of explainability for non-technical users. This domain-specific definition aims to support the future development of interpretable AI models that enhance user trust and usability in demand planning tools.

## 1 INTRODUCTION


Supply chain management (SCM) is a central process in businesses around the world. It contributes to fulfilling customer goals, gaining competitive advantage, and minimizing the loss of resources in the production cycle. As a result of the prominent benefits, there is a large market for SCM solutions designed for companies to manage their supply chain, some of the leading solutions being SAP Supply Chain Management, Blue Yonder, and Dynamics365 Supply Chain Management (D365 SCM).


A key process within SCM is demand planning (DP), which includes forecasting the future demand for products. DP enables companies to foresee an increase or decrease in the sales of their products, making it possible for them to plan downstream processes accordingly (IBM, 2017). An important element of demand planning is demand forecasting, which is an estimation of future demand, e.g. using time series data and mathematical computation to gain insights from the data. Traditionally, demand forecasting has been carried out using statistical methods such as ETS (Hyndman and Khandakar, 2008) and ARIMA (Box


and Jenkins, 1970). However, advances in machine learning (ML) and artificial intelligence (AI) are gradually replacing these statistical methods (altexsoft, 2022). ML forecasting models offer the potential of noticeably better predictions compared to statistical models, which is a big advantage in demand planning (GeeksforGeeks, 2024). However, with the increase in ML and AI models, we also see an increase in complexity, and the issue of the so-called "black-box" is claiming its space in the field of demand planning. Perceiving ML models as a black-box is currently a hot topic as their predictions get more difficult for humans to interpret, given their increased complexity. This is critical in fields, where the prediction is used to draw important conclusions (e.g. the medical field (Adadi and Berrada, 2018)) or make significant decisions (e.g. DP).

While not easy to solve, the black-box issue has set the foundation for a new field within ML and AI, namely *explainability*. Explainability in AI is a topic that has gained a lot of interest in recent years because of its ability to open up the black box of ML model (Adadi and Berrada, 2018).

The concept of explainability dates all the way back to the 1980s where it was first mentioned (Moore and Swartout, 1988). Later, in 2004, the term Explainable AI (XAI) was introduced (Van Lent et al., 2004). However, it is not until recent years that

<sup>a</sup> <https://orcid.org/0000-0001-8790-9396>

<sup>b</sup> <https://orcid.org/0000-0002-6720-1128>

<sup>c</sup> <https://orcid.org/0000-0003-0941-3146>

the concepts gained traction, naturally following the increase in AI complexity and reliance (Adadi and Berrada, 2018). Despite its rising popularity, a common definition of explainability has not found consensus. Instead, researchers of different fields give different meanings to the term, often taking either a computer-centered approach focusing on the correctness and completeness of the explanation or a human-centered approach focusing on how the explanation resonates with end users. Both approaches require the use of explainability methods, which are essential for providing the actual explanation of the ML model. Over time, a wide range of models for XAI have been developed to describe the decision-making process of ML and AI models. Explainable ML models generally exist in two forms: those that are interpretable by nature, e.g., decision trees, and those that become interpretable after adding an explainability method post-training (Naqvi et al., 2024; Lopes et al., 2022; Retzlaff et al., 2024). As the amount of research on explainability has increased, it has become more user- and context-specific. Research on explainability for demand planning is still sparse, and the need for analyzing the users and their needs remains.

As research of user-specific needs for explainability is not traditionally covered by the tools available in the field of XAI, it is beneficial to draw on methods from UX design. Introducing UX design methods enables the possibility of analyzing the needs of the users of the DP applications and, consequently, design explanations that resonate with them specifically.

Based on the above, we pose two concrete research questions:

1. How can explainability be defined in the context of the demand planning domain?
2. Who are the users of the demand planning applications, and what are their explainability needs?

This paper is structured into the following sections: *Explainability in AI*, *Identifying Explainability in DP applications* and *Conclusion*.

In the **Explainability in AI** Section, we introduce some of the existing work that has been done in the field of XAI. In the section **Identifying Explainability in DP applications**, we converge towards the end of the problem space and use the background research and theory to define the problem of what explainability is for the users of DP applications. Lastly, we sum up the findings and refer back to the initial problem statement in the **Conclusion** Section.

## 2 EXPLAINABILITY IN AI

This section outlines the foundation of our research on defining explainability in the DP domain. We look into how explainability is currently defined across literature, how to evaluate an explanation on its explainability, and lastly, explore currently existing methods for explainability.

### 2.1 Explainability in ML

So far, no formal definition of explainability has been broadly accepted among researchers in the XAI field. One reason for this is that the need and benefits of explainability vary greatly between different fields and users, meaning that a good explanation for one group of people might not be relevant to another (Suresh et al., 2021; Mohseni et al., 2020). Despite this, there have been different attempts at designing frameworks for how to provide useful explanations. One example of this from Vilone et al., who, in their paper *Notions of explainability and evaluation approaches for explainable artificial intelligence* (Vilone and Longo, 2021), identify four main factors that constitute a good explanation. These include a consideration of who the end-user is, what their goals are, what information they should receive, and the language used to deliver it.

Following this idea of user-dependent explanations, we find that explainability is not binary and should be defined by the degree to which it satisfies a set of relevant metrics for specific targeted users (Pawlicka et al., 2023; Nauta et al., 2023; Liao and Varshney, 2022). A wide range of metrics for explainability have been described in the literature, making it difficult to navigate and select the relevant ones. In Table 1, we have collected some of the most frequently encountered terms during the research on XAI, with the purpose of showing how often they are used and whether they are mentioned as being human-centered or computer-centered.

Explainability is very much dependent on the receivers, and a lot of research is currently being done on how users have different needs for explainability, e.g. as described in the survey *A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems* (Mohseni et al., 2020). Here, the authors explain the distinct goals and needs of users by grouping them into AI Novices, Data Experts, and AI Experts. **AI Novices** are the end-users who will interact with the ML product. They are generally assumed to have very little or no knowledge of machine learning and have no need for it either. The level of explainability for this group is

Table 1: Summary of the literature study on explainability in ML, including the explainability criteria mentioned in each paper.

	Human-centered							Computer-centered	
	Fairness	Accountability	Understandability	Trustworthiness	Usefulness	Performance	Satisfaction	Fidelity	Interpretability
Lopes et al. (Lopes et al., 2022)			x	x	x	x	x	x	
Mohseni et al. (Mohseni et al., 2020)	x	x	x	x	x	x	x	x	x
Hoffman et al. (Hoffman et al., 2019)			x	x	x	x	x		
Lim et al. (Lim et al., 2009)			x	x	x	x			
Pawlicka et al. (Pawlicka et al., 2023)	x	x	x	x	x	x	x	x	x
Markus et al. (Markus et al., 2021)	x		x	x	x	x		x	
Nauta et al. (Nauta et al., 2023)			x	x	x	x	x		x
Zhou et al. (Zhou et al., 2021)	x		x	x	x	x	x		x
Adadi et al. (Adadi and Berrada, 2018)	x	x	x	x	x		x		x
Binns et al. (Binns et al., 2018)	x	x		x		x			
Bussone et al. (Bussone et al., 2015)			x	x					
Seong et al. (Seong and Bisantz, 2008)				x		x			

determined by how useful and satisfying the explanation is to them, as well as how much they trust it. **Data Experts** are data scientists or similar who use the ML product to conduct analyses or research. They are assumed to be working directly with the ML model while not necessarily having a deep technical understanding of how the specific model works. Explainability to them is determined by how much it helps them to perform their tasks, in addition to how well the model itself is performing. **AI Experts** are the developers or engineers working on the ML model. Their explainability needs are described as different from the other two groups, as their focus is more on debugging and understanding the model itself. In particular, there is a difference in the needs and goals of explainability for the different user groups. The AI Novices generally have human-centered needs, as opposed to the AI Experts, whose needs are more computer-centered. This differentiation between human-centered and computer-centered metrics is common among researchers and is described, among others, by Lopes et al. (Lopes et al., 2022) and Mohseni et al. (Mohseni et al., 2020). They define human-centered explainability as the extent to which an ML system is understandable to humans, as well as how it affects them when interacting with it. On the other hand, computer-centered explainability is about how well the ML system is explaining the ML model itself, including how accurate the explanation

Table 2: Overview of key explainability terms in ML literature and their definitions, with all relevant references listed.

Term	References	Definition
<b>Fairness</b>	(Deck et al., 2024), (Mohseni et al., 2020), (Pawlicka et al., 2023), (Bussone et al., 2015)	Assessing the fairness of an ML model, particularly in sensitive domains like loan applications.
<b>Accountability</b>	(Lepri et al., 2018), (Binns et al., 2018)	The ability to attribute responsibility for decisions made by the model.
<b>Understandability</b>	(Lopes et al., 2022), (Mohseni et al., 2020), (Butz et al., 2022)	The extent to which the XAI system is understandable to users, facilitating the prediction of its outputs.
<b>Trustworthiness</b>	(Lim et al., 2009), (Pawlicka et al., 2023), (Vilone and Longo, 2021), (Adadi and Berrada, 2018)	Reflects the user's confidence in the system's reliability and alignment with their expectations.
<b>Usefulness</b>	(Seong and Bisantz, 2008), (Lopes et al., 2022), (Mohseni et al., 2020)	Evaluates the practical value of the explanations in assisting user decision-making.
<b>Performance</b>	(Lopes et al., 2022), (Mohseni et al., 2020), (Lount and Lauzon, 2012), (Markus et al., 2021)	Concerns user task performance when interacting with the XAI system.
<b>Satisfaction</b>	(Gedikli et al., 2014), (Vilone and Longo, 2021)	Represents user satisfaction with the provided explanations.
<b>Fidelity</b>	(Lopes et al., 2022), (Markus et al., 2021)	Reflects the accuracy of the explanation in representing the model's actual behavior.
<b>Interpretability</b>	(Bussone et al., 2015), (Lopes et al., 2022)	Describes the ease with which explanations can be understood by human users.

is to the truth. As seen in table 1, human-centered and computer-centered explainability are both umbrella terms, covering a number of other principles within explainability. In order to gain a deeper understanding of what meaning these terms carry across literature, table 2 describes each of the terms.

A human-centered perspective on what explainability involves is the assumption that an explanation is an answer to a question the user might have when interacting with the system (Liao et al., 2020; Preece, 2018; Vilone and Longo, 2021). The most typically asked questions relating to explainability are *why* and *how* (Vilone and Longo, 2021), answering questions such as *why* a certain prediction was made, or *how* a certain feature impacts the prediction. According to Liao et al., answering the right questions can constitute a good explanation, but what is considered right again depends on the person asking it (Liao et al., 2020). They did a study on the explainability needs of different user groups by conducting semi-structured interviews with 20 people. Notably, the user group labeled 'Business Decision Support' showed strong interest in explanations that enhance their decision confidence by showing the importance of attributes as well as explanations that are made in natural language.

## 2.2 Explainability Methods

In this section, we will go through different types of explainability methods and how they relate to the needs of end-users.

One of the main distinctions in explainability methods is *ante-hoc* and *post-hoc* approaches. When an explanation is retrieved directly from the model itself, e.g. from decision trees, it is said to be ante-hoc. Conversely, if an explanation is generated after model training, it is called post-hoc. Post-hoc explanations require the addition of explainability methods, which are applied separately from the model itself (Naqvi et al., 2024; Retzlaff et al., 2024), and can be either *model-specific* or *model-agnostic*. The distinction between the two lies in whether the method is specifically applicable to a given model or is generally applicable to a range of different ML models. Explainability methods can be further broken down into *local* and *global* explanations, where local explanations are used to describe the reasons for a single prediction, while global explanations are used to describe the overall model (Liao and Varshney, 2022). Examples of global, post-hoc explainability methods include Accumulated Local Effects (ALE) plots (Apley and Zhu, 2020) and Partial Dependence Plots (PDP) (Friedman, 2001), while examples of local post-hoc methods include Local Interpretable Model-agnostic Explanations (LIME) (Singh and Guestrin, 2016) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017).

Each of these explainability methods provides users with different types of explanations and is suitable for different purposes. Liao et al. closed the gap between algorithmic explainability methods and the needs of end-users by developing a *question bank*, which is a collection of questions that users might ask in relation to explainability, along with explainability methods that can be applied to answer these questions (Liao et al., 2020).

Liao et al. argued that there is no *one-fits-all* solution to a good explanation and suggested a collaborative approach, where UX designers and data scientists work together to identify relevant explainability methods. For this purpose, they used the question bank to develop a *mapping guidance* between user questions and explainability methods (Liao and Varshney, 2022).

### 2.3 Evaluating Explainability

There are several methods for evaluating explainability. Nauta et al. (Nauta et al., 2023) described how evaluating explainability is about measuring the degree to which an explanation satisfies a set of defined metrics and that each aspect of the explanation should be evaluated separately. Pawlicka et al. (Pawlicka et al., 2023) presented a similar approach by arguing that an explanation should be evaluated by **1)** check-

ing whether explainability is achieved by how well it fulfills the defined objectives, and **2)** comparing explanation methods to identify the most preferred one.

As for the definition of explainability, methods for evaluating explainability are also often divided into human-centered and computer-centered approaches. The human-centered evaluation methods include humans in the evaluation process and apply user testing of domain experts or lay people as a way to measure an explanation (Molnar et al., 2020). Meanwhile, computer-centered approaches use quantitative metrics to evaluate the explanation, e.g. in terms of fidelity (see Table 1).

Doshi-Velez et al. distinguished between functionally grounded, application grounded, and human grounded evaluations (Doshi-Velez and Kim, 2017). The functionally grounded evaluation corresponds to the computer-centered evaluation and does not include humans. The human-centered evaluation is divided into the application-grounded evaluation and the human-grounded evaluation, where the application-grounded evaluation is based specifically on target users, while the human-grounded evaluation is based on lay users, meaning humans that are not necessarily domain experts or targeted users. Human-centered and computer-centered evaluations each have their merits. Human-centered evaluations are perceived to be more accurate in determining the level of explainability (Zhou et al., 2021). However, they are also time-consuming and can have issues such as bias and inefficiency. At the same time, computer-centered evaluations are "objective" and require fewer resources in terms of time but does not include the user perspective (Pawlicka et al., 2023).

Hoffman et al. (Hoffman et al., 2019) developed a conceptual model to map out the process of evaluating explainability in a ML context. A slightly modified version of the conceptual model is shown in Figure 1. A user of an XAI system initially feels trust or mistrust in the ML model, and immediately forms a mental model about how the system works. An explanation is then provided to give a greater understanding of the system, which affects the user's mental model and builds trust. Hoffman further argued that a user's perceived trust/mistrust in the system greatly affects how they interact with the system, which in turn affects the performance. Each of these stages of the conceptual model can be assessed to evaluate the explanation, as a good explanation will provide the user with trust and understanding and, hence, better performance. Different methods for evaluating explainability have been suggested. Generally, we distinguish between quantitative and qualitative evaluation methods, in addition to subjective and objective evaluation

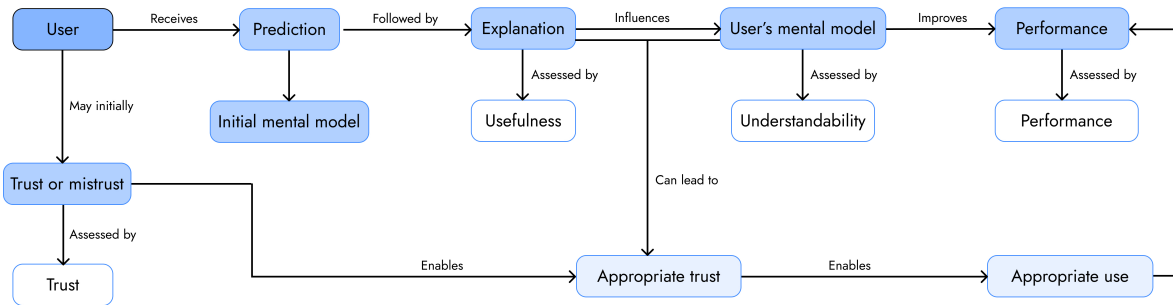


Figure 1: Conceptual model of explainability in ML. Adapted from (Hoffman et al., 2019; Lopes et al., 2022).

Table 3: Summary of most frequently encountered human-centered evaluation methods for explainability in ML.

Method	Type	Mentioned In
Likert Scale	Subjective/ Quantitative	(Bussone et al., 2015), (Berkovsky et al., 2017), (Nourani et al., 2019), (Lim et al., 2009), (Binns et al., 2018), (Gedikli et al., 2014)
Interview	Subjective/ Qualitative	(Gedikli et al., 2014), (Binns et al., 2018), (Lim et al., 2009), (Lount and Lauzon, 2012)
Think Aloud	Subjective/ Qualitative	(Bussone et al., 2015), (Binns et al., 2018)
Task Performance	Objective/ Quantitative	(Lim et al., 2009), (Huysmans et al., 2011), (Kulesza et al., 2010)
Self-explanation	Subjective/ Qualitative	(Bussone et al., 2015), (Cahour and Forzy, 2009)

methods. Subjective evaluations captures the participants’ own opinions or perceptions, while objective evaluations measure some defined objectives independent of the users’ opinions. Human-centered evaluations generally cover all types of evaluations, while computer-centered evaluations usually apply quantitative and objective methods. An overview of the evaluation methods is presented in Table 3.

A popular method for evaluating explainability is the of the Likert scale. This method is used to capture the subjective opinions of participants and get insights on how they perceive different metrics. Bussone et al. investigated how explainability affects the trust and reliability in users of Clinical Decision Support Systems (CDSS), and use a 7-point Likert scale to evaluate the users’ trust in the system before and after receiving an explanation (Bussone et al., 2015).

The research by Gedikli et al. is about improving satisfaction in recommender systems by helping the user to understand why certain predictions are given (Gedikli et al., 2014). They follow a similar approach

to evaluation, by asking participants to evaluate transparency and satisfaction on a 7-point Likert scale after receiving an explanation, and comparing transparency to satisfaction. Both of these papers use a similar approach by applying a qualitative evaluation method to support the quantitative methods. Bussone et al. applied the "think aloud" method for evaluation, by asking users to share their thoughts during the performance of a given task, and then used post-task interviews to gather additional information. Gedikli et al. also perform post-task interviews with the purpose of validating the results of the quantitative approach. The papers by Lim et al. (Lim et al., 2009) and Huysmans et al. (Huysmans et al., 2011) both objectively evaluate their explanations using task performance. Lim et al. researched the effectiveness of different types of explanations (*why* and *why not*) in context-aware intelligent systems. They use task performance as one of the measures for evaluating the explanations, and measure it in terms of completion time, Fill-in-the-Blanks test answers and answer correctness. The answer is rated into one of four groups, depending on the actual correctness and the level of detail the participant was able to convey.

Huysmans et al. evaluated the explainability (which they refer to as 'comprehensibility') of decision tables, decision trees and rule based predictive models. Participants were asked to answer a list of yes/no questions, and rate their confidence on a Likert scale. The authors then evaluated the explanations based on the perceived confidence of the participants and their task performance, which is determined by the accuracy (percentage of correct answers) and the task completion time.

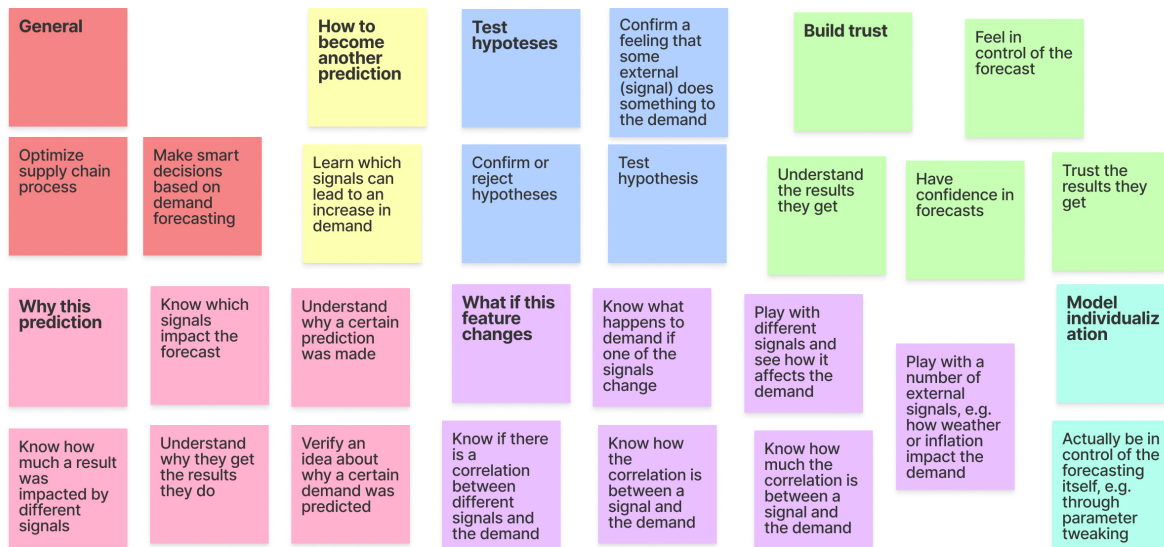


Figure 2: Final affinity diagram used to structure the SME Interview into concrete user needs.

### 3 IDENTIFYING EXPLAINABILITY IN DP APPLICATIONS

Explainability is highly dependent on the user and their specific needs, and for this reason, it is important for a good implementation of explainability, to research the specific needs of users of DP applications. The research combines different methods within UX Design, the purpose being to gather information about the users, and structure it into concrete needs and requirements for explainability. By the end of this section, we will have a clear definition of the explainability needs of users, what explainability is in context of DP applications, and which explainability methods can be applied to provide that.

#### 3.1 Identifying User Needs

##### 3.1.1 SME Interview

To gather information about the users of DP applications, we decided to do interviews with demand planners. The purpose of the interview is to gather information about the users, including how they use DP applications, what they use it for, and which requirements they have for current and future use. We did *unstructured* interviews since this interview type allowed the conversation to be more dynamic, and for emerging questions to be asked.

##### 3.1.2 Affinity Diagramming

The interviews provided a large amount of unstructured data, which needed to be organized to extract relevant information. For this purpose, we chose to apply affinity diagramming, which is used to organize the data by translating the raw qualitative data into a concrete mapping of the users and their needs. First, the information from the raw interview data was mapped out on post-its, disregarding their perceived relevance. Next, the post-its were grouped into subjects, each given a unique color. Lastly, labels were assigned to the groups in the diagram in order to define each of them more specifically. The result is seen in Figure 2, and provides a clear overview of the subjects and needs that were discussed during the interview with demand planners. We see the seven subgroups of user information identified through the interviews, and that the overall goal of users is to optimize the SC process and make smarter decisions based on the demand forecast. The users want to have confidence in the decisions they are making, and feel in control of the forecasting. They also want to trust the model predictions, the overall forecasting system, get a better understanding of the predictions, and some users even want the option to control the forecasting model itself.

There is a need for users to have an understanding of *why* certain predictions are made, as well as what happens if some of their features change. This includes e.g. the option to see correlation between different features and the demand, and being able to experiment with the feature values to learn how they

drive demand. Lastly, there is also a wish to learn how demand can be increased in the context of the available features.

It is clear that users have some ideas in mind about drivers of the demand of their products. They might have a feeling about something having an effect, without being able to check the relevance of that feeling. So, they use these ideas and feelings to make up hypotheses which they seek to confirm or reject.

Based on the findings from the affinity diagram, we have chosen to categorize the explainability needs of the users of DP applications into the following objectives:

- O1 Make better decisions
- O2 Trust the predictions
- O3 Understand why certain predictions were made
- O4 Understand what happens to demand if the features change values
- O5 Increase demand using available features
- O6 Test hypotheses
- O7 Be in control of forecasting model

The topics found in the affinity diagram, are usually prioritized to select the ones to move forward with. This means, that not all the findings above will necessarily be fulfilled within the scope of this paper.

### 3.2 User Story Mapping

After grouping the findings from the interviews in the affinity diagram, we have a list of objectives. As mentioned, we do not include all of these objectives going forward, and choose a subset. For this purpose, we apply User Story Mapping (USM) to prioritize selected objectives in terms of goals, activities, and tasks. The tasks are outlined as the necessary steps for the user to complete an activity, and are subject to change later in the design process. The USM is shown in the Appendix, where each of the objectives from section 3.1.2 are presented as either a goal or an activity.

From the USM, we found that the overall goals users are trying to achieve through explainability are O1 and O2. Additionally, we found that O6, can be obtained through O3, O4 and O5.

Based on a prioritization of the USM and in collaboration with demand planners, we decided to move forward with O3, O4 and O5. The overall goals of building trust and making better, more confident decisions, are natural derivations from good explanations as also described in section 2.3, and the users will be able to test their hypotheses on which features drive demand and why certain predictions were made.

### 3.3 Defining Explainability in DP Applications

After getting an understanding of the needs and goals of the end-users, we moved on to defining explainability in DP applications. Currently, we know who the end-users are, their goals, which questions they want answers to and their level of technical expertise. This means that we now have all the essential components to structure the explanations. However, we still need to establish a clear definition of what explainability is in the context of these components.

As we found in section 2.3, explainability in an ML system can be measured by how well it satisfies a set of user-dependent and measurable objectives. We have chosen to rely on this definition, and use the findings from section 3.1 to define a set of requirements that should be satisfied in order for the user to feel accomplished in their goals of gaining more trust in the forecasting model and making more confident decisions, as well as optimizing the SC process. Following this approach will ensure, that explainability becomes a measurable term, allowing us to evaluate and compare different explanations. Based on the objectives that were identified through the affinity diagram and USM, along with the research on explainability, more specifically table 1, we choose a set of relevant objectives to define explainability in DP applications. The objectives are chosen from existing literature based on the extent to which they can fulfill the goals and needs of users, and are listed below:

- **Usefulness:** The explanation should be useful and satisfying to the user.
- **Trustworthiness:** The explanation should meet the user's expectations and provide them with confidence in their decisions.
- **Understandability:** The explanation should be understandable and meet the user's expectations in terms of what information it provides i.e. the questions it answers.
- **Performance:** The explanation should help the user to perform their intended tasks more efficiently.

## 4 CONCLUSIONS

The aim of this paper was to explore and define explainability within the supply chain management domain, specifically focusing on the demand planning. In order to perform this investigation, we adopted an approach inspired by the double-diamond framework,

involving stages of discovery to deeply understand the problem space.

During the discover phase, we found that explainability is not a binary term, and that something can be explainable to one group of users while not necessarily being explainable to another. As a result, adding good explanations requires a study of the target users in terms of their needs and goals when interacting with the entire XAI system. In the define phase of the problem space, we found that the main goals of users of DP Applications is to 1) Make better decisions and 2) Trust the predictions they get from the system. More specifically, they want to know *why* certain predictions are made, and *what* happens to a prediction if certain features change.

In conclusion, this paper contributes to the evolving field of explainable AI in supply chain management by providing a user-focused description of explainability and identifying the specific needs of demand planning application users. This discovery built a foundation for implementing explainable AI solutions that can enhance user trust, satisfaction, and decision-making in demand planning processes.

## REFERENCES

- Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.
- altexsoft (2022). Demand forecasting methods: Using machine learning to see the future of sales.
- Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Berkovsky, S., Taib, R., and Conway, D. (2017). How to recommend?: User trust factors in movie recommender systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 287–300, United States. Association for Computing Machinery (ACM).
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018). “it’s reducing a human being to a percentage”: Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, page 1–14. ACM.
- Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.
- Bussone, A., Stumpf, S., and O’Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics*.
- Butz, R., Schulz, R., Hommersom, A., and van Eekelen, M. (2022). Investigating the understandability of xai methods for enhanced user experience: When bayesian network users became detectives. *Artificial Intelligence in Medicine*, 134:102438.
- Cahour, B. and Forzy, J.-F. (2009). Does projection into use improve trust and exploration? an example with a cruise control system. *Safety Science, Volume 47, Issue 9*.
- Deck, L., Schomäcker, A., Speith, T., Schöffner, J., Kästner, L., and Kühl, N. (2024). Mapping the potential of explainable ai for fairness along the ai lifecycle.
- Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*.
- Gedikli, F., Jannach, D., and Ge, M. (2014). How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382.
- GeeksforGeeks (2024). Difference between statistical model and machine learning.
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2019). Metrics for explainable ai: Challenges and prospects.
- Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154.
- Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*.
- IBM (2017). What is demand planning?
- Kulesza, T., Stumpf, S., Burnett, M., Wong, W.-K., Riche, Y., Moore, T., Oberst, I., Shinsel, A., and McIntosh, K. (2010). Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 41–48.
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31(4):611–627.
- Liao, Q. V., Gruen, D., and Miller, S. (2020). Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15.
- Liao, Q. V. and Varshney, K. R. (2022). Human-centered explainable ai (xai): From algorithms to user experiences.
- Lim, B. Y., Dey, A. K., and Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- Lopes, P., Silva, E., Braga, C., Oliveira, T., and Rosado, L. (2022). Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19).



- Lount, A. B. M. and Lauzon, C. (2012). Are explanations always important? a study of deployed, low-cost intelligent interactive systems. *International Conference on Intelligent User Interfaces, Proceedings Iui*.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.
- Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113:103655.
- Mohseni, S., Zarei, N., and Ragan, E. D. (2020). A multidisciplinary survey and framework for design and evaluation of explainable ai systems.
- Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable machine learning—a brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 417–431. Springer.
- Moore, J. D. and Swartout, W. R. (1988). Explanation in expert systems: A survey. *University of Southern California*.
- Naqvi, M. R., Elmhadi, L., Sarkar, A., Archimede, B., and Karray, M. H. (2024). Survey on ontology-based explainable AI in manufacturing. *Journal of Intelligent Manufacturing*, 35(8):3605–3627.
- Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.
- Nourani, M., Kabir, S., Mohseni, S., and Ragan, E. D. (2019). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. *Proceedings of the Aaai Conference on Human Computation and Crowdsourcing*, 7:97–105.
- Pawlicka, A., Pawlicki, M., Kozik, R., Kurek, W., and Choraś, M. (2023). How explainable is explainability? towards better metrics for explainable ai. In *The International Research & Innovation Forum*, pages 685–695. Springer.
- Preece, A. (2018). Asking ‘why’ in ai: Explainability of intelligent systems – perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*.
- Retzlaff, C. O., Angers Schmid, A., Saranti, A., Schneeberger, D., Röttger, R., Müller, H., and Holzinger, A. (2024). Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research*, 86:101243.
- Seong, Y. and Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*.
- Singh, M. T. R. S. and Guestrin, C. (2016). ”why should i trust you?” explaining the predictions of any classifier.

*Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*.

- Suresh, H., Gomez, S. R., Nam, K. K., and Satyanarayan, A. (2021). Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.
- Van Lent, M., Fisher, W., and Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pages 900–907.
- Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*.
- Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *electronics*.

## APPENDIX

### 4.1 SME Interview

The full interviews can be made available upon request, if needed.

### 4.2 User Story Mapping

