Comprehension Is a Double-Edged Sword: Over-Interpreting Unspecified Information in Intelligible Machine Learning Explanations

YUEQING XUAN and EDWARD SMALL, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Australia

KACPER SOKOL, Department of Computer Science, ETH Zurich, Switzerland

DANULA HETTIACHCHI and MARK SANDERSON, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Australia

Automated decision-making systems are becoming increasingly ubiquitous, which creates an immediate need for their explainability. However, it remains unclear whether users know what insights an explanation offers and, more importantly, what information it lacks. To answer this question we conducted an online study with 200 participants, which allowed us to assess explainees' ability to realise *explicated information* – i.e., factual insights communicated by an explanation – and *unspecified information* – i.e, insights that are not presented – in four representative explanation types: model architecture, decision surface visualisation, counterfactual explainability and feature importance. Our findings uncover that highly comprehensible explanations, e.g., feature importance and decision surface visualisation, are highly susceptible to misinterpretation since users tend to infer spurious information that is outside of the scope of the explanation. Additionally, while the users gauge their confidence accurately with respect to the information explicated by these explanations, they tend to be overconfident when misinterpreting the explanations. Our work demonstrates that human comprehension can be a double-edged sword since highly accessible explanations may convince users of their truthfulness while possibly leading to various misinterpretations at the same time. Machine learning explanations should therefore carefully navigate the complex relation between their full scope and limitations to maximise understanding and curb misinterpretation.

Additional Key Words and Phrases: Human-centred computing, Comprehension, Evaluation, Explainability, Machine learning, Artificial intelligence.

Highlights

- Users are ignorant of explanations' limitations and over-generalise factual insights.
- Highly comprehensible explanations are more susceptible to user misinterpretation.
- Users are overconfident in misinterpretation of information missing from explanations.
- Easy-to-understand explanations are misleading despite subjective comprehensibility.

Source Code https://github.com/xuanxuanxuan-git/hcxai
DOI https://doi.org/10.1016/j.ijhcs.2024.103376
Cite us Let me figure it out.... Dont forget to add "Appendix".

1 INTRODUCTION

Artificial Intelligence (AI), and Machine Learning (ML) in particular, can deliver diverse benefits across the economy and society. These technologies support diagnosis and early detection of dangerous health conditions in hospitals [18, 33], improve at-home healthcare services [70], and support personalised learning and teaching [5, 10]. However, with

Authors' addresses: Yueqing Xuan, yueqing.xuan@student.rmit.edu.au; Edward Small, edward.small@student.rmit.edu.au, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Australia; Kacper Sokol, kacper.sokol@inf.ethz.ch, Department of Computer Science, ETH Zurich, Switzerland; Danula Hettiachchi, danula.hettiachchi@rmit.edu.au; Mark Sanderson, mark.sanderson@rmit.edu.au, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Australia.

automated decision-making tools becoming increasingly sophisticated, and hence opaque, we risk creating and using systems that we do not fully understand or control. This situation not only has ethical implications [17], but also raises concerns with respect to accountability [39], safety [14] and liability [37].

Researchers and regulatory bodies have urged providers of AI tools to explain predictions and decisions output by these systems to the public, especially the affected individuals [25, 26, 84]. The European Union's General Data Protection Regulation (GDPR) introduced a right to explanation, which stipulates that all individuals should be able to obtain "meaningful explanations of the logic involved" in automated decision-making [84]. The Canadian Artificial Intelligence and Data Act (AIDA) requires high-impact AI systems to be transparent by providing information that is sufficient to allow the public to understand their capabilities, limitations and potential impact [25]. However, it remains unclear what constitutes a "meaningful" explanation [24]. To complicate matters further, explanations of AI systems can be deceptive, e.g., they can be manipulated to intentionally mislead the public [15, 21, 46]. While previous work has evaluated whether different machine learning explanations are understandable to diverse stakeholders [8, 11, 36, 45], research on whether users are aware of the limitations of the information explicated by ML explanations is still largely missing.

The availability of AI explainability tools has soared in recent years. Popular explanatory mechanisms include deploying inherently transparent models such as tree-based predictors [68] or generating post-hoc explanations of black-box models [26, 38, 85]. While being essential for high stakes domains [68], transparent models may not necessarily engender understanding, especially in lay audiences [1, 55, 79], thus appropriate explanatory mechanisms may need to be deployed to facilitate user comprehension. Genuine understanding of an explanation requires the users not only to internalise the information that it offers but also to appreciate the information that is unspecified. The latter aspect concerns over-generalisation of explanations beyond their scope, which is of particular importance in view of the illusion of explanatory depth, i.e., people's limited knowledge and their misleading intuitive epistemology make them feel they understand a topic with far greater depth than they really do [67].

In this paper, we study user comprehension of *information explicated* and *unspecified* by machine learning explanations. The former is understood as factual insights that are provided in an explanation; the latter encompasses information that the explanation is not designed to communicate but may be misconstrued by the users. In practice, information about ML models is often selectively presented in explanations to maximise comprehensibility; however, the line separating explicated from unspecified information is often blurry, which may unintentionally contribute to users over-generalising their interpretation. Consequently, users can internalise invalid explanatory insights even when neither deception nor malice were intended. We refer to such explanations – for which unspecified information can be easily misconstrued and over-interpreted – as *misleading*.

As an example, consider counterfactual explanations, which inform users about the smallest possible change to a feature vector that results in a desirable outcome [85]. While they report a subset of features whose values need to be changed, they do not communicate the (local or global) importance of these features for the model's decision. Nonetheless, such an explanation can be misleading to some users if they misunderstand the explanation and infer insights about feature importance, which information is not communicated, i.e., unspecified. In such a case, the explanation could potentially be harmful because it is intrinsically faithful and trustworthy, which is likely to convince a user of its truthfulness and utility [30], while at the same time possibly leading to various misinterpretations.

Current research has covered different types of *misleading* explanations. Lakkaraju and Bastani [46] manipulated the explanation generation process and created explanations that misrepresent black-box models to disguise their lack of fairness. The authors evaluated whether these unfaithful explanations can deceive users into believing that such black-box models are, in fact, fair. Our paper differs from their work as we generate genuine and faithful explanations without any intention to deceive the users. We also focus on misinterpretation of unspecified information, which is a different facet of explanations being misleading. Eiband et al. [21] explored the impact of *placebic* explanations – which are constructed artificially and provide no useful information – on user trust. Our work differs from their approach as we use meaningful explanations that are popular in the literature. Our work is complementary to the aforementioned papers since users can have high trust in unfaithful explanations regardless of their level of comprehension, and highly comprehensible explanations do not necessarily guarantee user trust.

In this work, we designed and executed an online user study with 200 participants to systematically investigate four representative explainability approaches: model architecture, decision surface visualisation, counterfactual explainability and feature importance. Each explanation is generated for two inherently transparent predictive models: logistic regression and decision tree (we justify our explanation choices in Section 3). Specifically, to assess user comprehension of both explicated and unspecified information, we designed two types of comprehension statements for each explanation: a *statement of explicated information*, which assesses whether users can understand the insights communicated by an explanation; and a *statement of unspecified information*, which gauges whether users know that certain information is unspecified by an explanation, thus can appreciate its limitations. In our study, the participants were asked to judge these two statements for each explanation of one of the ML models. We also asked the participants to report the confidence of their judgements as well as the perception of whether an explanation was easy to understand and sufficiently detailed. We further collected demographic information. With this set up (the details of which are described in Section 4), we seek to answer the following research questions:

RQ1 How do user comprehension of explicated information and, more importantly, misinterpretation of unspecified information differ across diverse machine learning explanations?

RQ2 How do user confidence in their (mis)interpretation, their perception of explanation difficulty and their appreciation of explanation information richness differ across explanations?

RQ3 How do individual characteristics – such as education level, algorithm literacy, technical background, graph literacy and experience with explainable AI user studies – affect user (mis)interpretation of explanations?

We find that it was easier for participants to identify explicated information when compared to unspecified information across all four explanation types. Additionally, when the explicated information offered by an explanation was highly comprehensible, users were easily misled as they tended to more readily misconstrue unspecified information, thus misinterpret the explanation. For example, the participants were significantly more likely to understand the information explicated by feature importance and decision surface visualisation in comparison to the other two explanation types, but these two explanation types were also more misleading. Counterfactual explainability and model architecture were less intelligible in terms of explicated information, but also less misleading. Furthermore, the participants who identified *explicated information* correctly reported higher confidence in their interpretation than their peers who did not; for *unspecified information*, the participants who understood the limitations of the explanations were less certain than their overconfident peers who misinterpreted the explanations. Full study results are presented in Section 5, followed by an in-depth discussion in Section 6 and a conclusion in Section 7.

In summary, our work demonstrates that comprehension of the information explicated by ML explanations can be a double-edged sword as it may prompt users to misconstrue explanatory insights and be overconfident in their incorrect beliefs. Specifically, the contributions of our work are three-fold.

(1) We introduce the novel concept of *unspecified explanatory information*, which is misconstrued by explainees based on a sound and faithful ML explanation. We then highlight the importance of assessing explicated and unspecified information in explainable AI – i.e., the scope of explanatory insights and the (implicit) limitations thereof – which can lead to their unintended misinterpretations. We demonstrate this phenomenon through a comprehensive user study run on a collection of representative machine learning explanations that are faithful to the underlying predictive model, thus ensuring the ecological validity of our study.

(2) We demonstrate that comprehension is a double-edged sword. An explanation can at the same time be highly comprehensible and susceptible to misinterpretation as users are less attuned to the limitations of "easy-to-understand" explanations.

(3) We design a highly flexible and reusable framework for evaluating whether an explanation is comprehensible in view of its explicated and unspecified information, thus the degree to which an explanation can mislead explainees. This contribution enables others to evaluate different types of explanations situated in distinct domains, allowing them to identify and curb any form of their misinterpretation, e.g., by explicitly indicating their limitations and introducing complementary explanations (to the same effect).

2 RELATED WORK

Evaluating the intelligibility of different post-hoc explainability and ante-hoc interpretability approaches across diverse stakeholders is a popular research topic. This paper builds upon prior work in eXplainable AI (XAI) and human comprehensibility of such techniques, a review of which follows.

2.1 Explainable AI

There has been increasing interest in safety-critical industries to leverage machine learning models for high stakes predictions. Many machine learning models are constructed as black boxes whose internals are either unknown to the observer or impossible to interpret by humans [68]. A range of XAI tools are commonly used to explain the logic of automated decision-making. For example, transparent models that are self-explanatory are a form of ante-hoc interpretability in which the model itself constitutes an explanation [79]. Post-hoc explanations – defined as an "interface" between humans and a predictor that is both comprehensible to humans and an accurate proxy of the predictor [26] – provide a different mechanism to inspect the behaviour of a model. For example, a representative post-hoc explainability tool – counterfactual explanations – familiarises humans with unknown model behaviour by simulating the hypothetical input circumstances under which the output changes; consequently, trust can be achieved through such counterfactual thinking [19]. Such insights, however, may not be reliable and truthful with respect to the underlying predictor even if they are of high fidelity [68, 77]. For example, Aïvodji et al. [3] and Lakkaraju and Bastani [46] manipulated explanations of black-box models to hide their unfair decision-making by approximating them with seemingly fair transparent models from which explanations were derived. In such cases, if users misunderstand a model after consuming an explanation, it is unclear whether this is caused by misinterpretation of a post-hoc explanation or its false representation of the underlying model [73]. In this work, we take a step back from sophisticated models and their post-hoc explainability, and instead look at inherently transparent models and their ante-hoc interpretability, which is guaranteed to be truthful and faithful.

Additionally, explanations differ in their scope – global or local – and the type of information they communicate [75]. Global explanations describe the overall behaviour of a machine learning model; local explanations pertain to individual predictions. An inherently interpretable model constitutes a global explanation since we are able to understand the logic of an entire model, nonetheless it also offers local insights when following its reasoning for individual instances [26]. Feature-based explanations, e.g., feature importance or influence, show how much a feature impacts a prediction (locally) or how important it is for the model as a whole (globally) or its individual prediction (locally). As a result, they differ in general strengths, weaknesses and applicability to real-world use cases [65].Given that multiple tools of varying scopes are developed to help users understand predictive models from different perspectives, it is crucial to investigate the effectiveness of these explanations. This does not only require the explainees to correctly identify the scope, but also to be aware of the limitations of an explanation. In this paper, we explore both of these perspectives with a strong focus on whether users can recognise what information an explanation lacks, i.e., its limitations, to avoid their over-generalisation.

2.2 User Comprehension

Explanations are intended to assist users in understanding the general functioning as well as details of an ML model from varying perspectives. Because of the breadth and scope of these objectives current literature lacks a consensus on what actually constitutes user comprehension [53, 78]. Cheng et al. [11] considered user understanding in terms of whether explainees can pick up the influence of features on a model's output and simulate predictions of a model given feature changes. Bove et al. [7] captured user understanding from the perspective of identifying feature influence on a single instance and the scope of an explanation. In another work, which explored the intelligibility of counterfactual explanations, Bove et al. [8] assessed whether users can identify that the provided information was a counterfactual.

In addition to post-hoc explainability, researchers have also explored the comprehensibility of inherently transparent models, i.e., ante-hoc interpretability [2]. Huysmans et al. [29] explored the comprehensibility of a number of such models by considering whether users can derive the model's output and identify feature influence. Similarly, Bell et al. [4] measured interpretability as users' ability to anticipate the output of an ML model or identify its most important feature. In another work, Lage et al. [45] evaluated how well users can simulate the predictions of different transparent models. The authors asked participants to derive and verify outputs as well as anticipate changes to outputs given alterations to the input values.

User-based evaluation of explainability and interpretability can be premised on objective or subjective assessment of explainees' understanding of predictive systems [80]. The former allows for a quantitative approach that employs a questionnaire consisting of a collection of curated questions about a predictive system that are aligned with a selected definition of comprehension [7, 8, 11]. Subjective understanding, on the other hand, is often based on the Explanation Satisfaction Scale [28], which asks users whether an explanation is understandable and fulfils their needs, among others [7, 8, 72]. Our user study builds upon both of these paradigms and extends them beyond evaluating the comprehension of information explicated by an explanation, also measuring whether users can recognise (often implicit) explanation limitations, thus avoid misconstruing unspecified information. In this paper, we employ a questionnaire to evaluate objective understanding and subjective perception of explanations accounting for both of the aforementioned perspectives.

A related line of research has studied *misleading* explanations, but the definition of this property is inconsistent. Eiband et al. [21] explored whether users would misplace trust in explanations that contain void information; their main focus was on meaningless and invalid explanations. Lakkaraju and Bastani [46] assumed misleading explanations to be those that deliberately misrepresent black-box models to disguise their unfairness; they created misleading explanations so that users cannot access the underlying issue even if they interpret those explanations correctly. In contrast, we experiment with well-established (ante-hoc) interpretability techniques that are designed to maximise intelligibility of (inherently transparent) predictive models, focusing on over-interpretation of *unspecified* information.

Our idea about user misinterpreting explanations is also closely related to the folk concept of behaviour [52] and the illusion of explanatory depth [67]. Jacovi et al. [30] attributed the failure where user understanding differed from what the explanation attempted to communicate to folk concepts of behaviour. Furthermore, Chromik et al. [13] examined if users fell for an illusion of explanatory depth when interpreting additive local explanations. However, both pieces of work focused exclusively on explanations about local feature attribution, and observed that users inferred global information from local explanations. Our work deploys a comprehensive set of explanations including both local and global explanations, and examines the opposite tendency of over-interpretation, i.e., inferring local information from global explanations. We also find that users infer excessive amount of local information that is not intended to be communicated in a local explanation, which results are complementary to those reported by Chromik et al. [13].To the best of our knowledge, we present the first comprehensive user study that identifies what information is missing from representative explanations and evaluates users' comprehension of information that is unspecified by these explanations.

2.3 Explanation Effectiveness

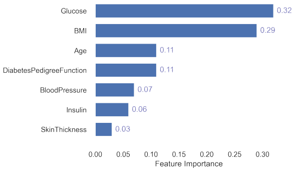
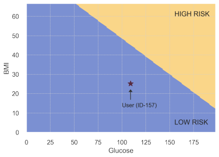
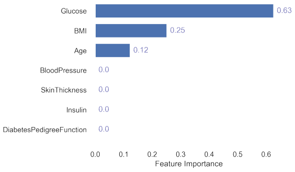
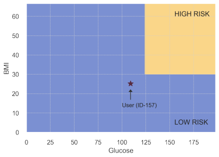
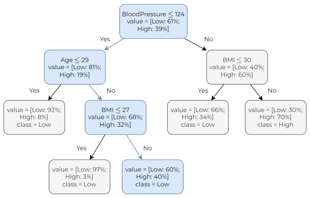
Extensive research has been done to evaluate the effectiveness of ML explanations in improving user understanding. Cheng et al. [11] studied the ability of different explanation interfaces to increase people's understanding of a data-driven university admissions system. They concluded that interactive and ante-hoc explanations increase participants' objective understanding of the algorithm, whereas their subjective understanding does not increase in the latter explanation prototype, possibly due to information overload. Bell et al. [4] compared two interpretable models – linear regression and decision tree – with a black-box model applied to public policy domains. Their result indicated that black-box models could be just as explainable as inherently interpretable models, with one possible reason being user confusion caused by the overwhelming amount of information given by interpretable models. Bove et al. [8] found that plural counterfactual examples increase objective understanding and satisfaction of explainees. In another work, Bove et al. [7] showed that augmenting local feature importance explanations with contextual information improves their subjective understanding.

Some early XAI work focused exclusively on interpretable models and explanations thereof. Huysmans et al. [29] empirically compared the comprehensibility of predictive models based on decision tables, trees and rules, and demonstrated that decision tables provide a significant advantage in terms of comprehensibility. Lim et al. [49] investigated which of “Why?”, “Why not?”, “How to?” and “What if?” explanations are more effective in helping users to understand a decision tree model. Their results showed that “Why?” explanations yield the best understanding. Despite a wealth of literature on the comprehensibility of different explanations, little effort has been made to compare local and global explanations, or explore whether users can identify their limitations. Our work fills this gap by focusing on local and global ante-hoc inherently interpretability of two transparent predictive models, thus guaranteeing the correctness of these insights.

3 STUDY DESIGN

Before reporting our results, we discuss the choice of machine learning models and their interpretability techniques as well as use case scenarios in our experiments.

Table 1. Overview of four explanations for two predictive models used in our study: logistic regression and decision tree. In the experiment, these explanations are paired with an informative description as outlined in B. The logistic regression model is displayed with specific weights and accompanied by a coefficient table to assist the users in interpreting the equation.

	Feature Importance	Decision Surface	Counterfactual	Model Architecture
Log. Reg.			*Had your Glucose been 150 and BMI been 30, you would have been predicted with low risk.*	$z(x) = \omega_0 + \omega_1 x_1 + \cdots + \omega_7 x_7$ $f(z) = \frac{1}{1+e^{-z}}$ $\text{Result} = \begin{cases} \text{high} & \text{if } f(z) \geq 0.5 \\ \text{low} & \text{otherwise} \end{cases}$
Decision Tree			*Had your Glucose been 150 and BMI been 29, you would have been predicted with low risk.*	 BloodPressure ≤ 124 value = [Low: 61%; High: 39%] Yes No Age ≤ 29 value = [Low: 81%; High: 19%] BMI ≤ 30 value = [Low: 40%; High: 60%] Yes No value = [Low: 92%; High: 8%] class = Low BMI ≤ 27 value = [Low: 68%; High: 32%] Yes No value = [Low: 66%; High: 34%] class = Low value = [Low: 30%; High: 70%] class = High Yes No value = [Low: 97%; High: 3%] class = Low value = [Low: 60%; High: 40%] class = Low

3.1 Predictive Models

In this study, we explore the intelligibility of explanations for *decision tree* and *logistic regression* models. We select these two ML models because they are inherently transparent and each of their parameters is directly interpretable, therefore we can generate a diverse range of faithful and reliable ante-hoc explanations in addition to using the model itself as an explanatory artefact [26, 31]. These two models are commonly used as explanatory artefact given their structural simplicity and because they provide the most comprehensive (full) information about the model's functioning. Additionally, these predictive models are commonly used in high stakes domains because of their transparency and interpretability [50, 68]. Both models are capable of predicting binary outcomes, they can be trained as to maintain low complexity, and they tend to achieve comparable predictive performance after appropriate hyper-parameter tuning, yet they operate in a distinct manner [26]. For our experiments, we use scikit-learn [61] to train the decision tree and logistic regression models, and perform hyper-parameter tuning to ensure their high predictive performance.

We choose transparent ML models to ensure perfect fidelity of their explanations, i.e., the explanations are guaranteed to accurately reflect the internals of a model. As such, we are able to explicitly verify the quality of the generated explanations, thus control for the confounding factor where a user misunderstands an explanation because it lacks fidelity rather than intelligibility. In contrast, post-hoc explanations of black-box models are generated by only approximating their behaviour, which can result in unfaithful explanations that fail to precisely capture the functioning of the explained model [20]. Given the opaqueness of black-box models, it is technically challenging to design their explanations with perfect fidelity; we are thus unable to control for the aforementioned confounding factor. In addition, explanations of black-box models can be easily manipulated due to the nature of their generation process – they only approximate the black box [46, 76]. Contrarily, ante-hoc interpretations of transparent models cannot be easily doctored, which ensures that user misunderstanding can be easily attributed to a concrete part of the explanatory pipeline.

3.2 Explainers

We choose four explanation types – *model architecture*, *decision surface visualisation*, *feature importance* and *counterfactual explainability* – that are commonly used to assist humans in understanding ML models. Each of these explanations can be appealing to a different stakeholder regardless of their level of expertise. Table 1 shows examples of the explanations used in our user study for the two ML models. These explanations differ in their scope, hence the amount of information they communicate. We use feature importance as a global explanation that conveys how important each feature is for the overall performance of an ML model. We employ decision surface visualisation to display local behaviour of an ML model, i.e., its change in prediction, when varying two features for a single data point while other features remain unchanged, which allows us to plot it in two dimensions. Given the difficulty of visualising a high-dimensional decision surface in an explanation (seven dimensions for our dataset), it is necessary to explicitly communicate that the remaining features are kept constant to enable the generation of this two-dimensional explanation. This assumption is explicitly provided in the description of the decision surface explainer to ensure that the explainee clearly understands what is being shown.Counterfactual explainability communicates local behaviour of a model for a given data point by informing the explainee about the smallest change to the feature vector that results in a desirable outcome [85]. The architecture of transparent models is often used as an explanation itself and allows the users to simulate its predictive behaviour (local) or grasp its overall mechanics (global) [4, 29, 49].

To ensure (ecological) validity of our findings, we use existing XAI tools to generate these explanations. Feature importance for logistic regression is based on the absolute value of the model's coefficients; for decision trees, it is based on the decrease in Gini impurity calculated for the training data [9]. To unify the representation of feature importance across these two models, we normalise the importance values such that they sum up to 1 and sort feature importance in descending order. To visualise the decision surface we randomly select two features with non-zero importance and display how altering their values, while keeping all the other features unchanged, affects the prediction of a selected data point. We generate realistic, feasible and actionable counterfactuals with a state-of-the-art explainer called FACE and display them as text [63]. As a result, all of our explanations are perfectly faithful with respect to the underlying ML models; because the models themselves are transparent, the explanation validity can be easily verified.

Given that the explanations differ in their scope and the type of information they contain, it is impractical to communicate them through a single modality while simultaneously ensuring that the information is presented clearly and concisely. Therefore, each explanation is represented with its most common modality as reported in existing literature. Feature importance is shown as a bar chart, as used by the popular SHAP [51] and LIME [66] explainers. We use a two-dimensional visualisation to show a decision surface, following the common practice [23, 29]. Counterfactual explanations are predominantly offered in the form of a textual statement [69, 85]. For the model structure of a decision tree, we use its canonical, hierarchical representation [4, 49]. To display the structure of a logistic regression model, we present both its mathematical formulation (equations) as well as a table with its coefficients to assist users who lack mathematics background, following similar practice reported in the XAI literature [4].

Since these explanations cannot be delivered in the same modality, we set the explanation technique as one of the independent variables in our study. We are not interested in finding the explanation modality that is the most understandable, but rather in investigating whether different types of explanation techniques, which convey different information, are susceptible to over-interpretation. If we fix the modality across explanation types, their quality becomes compromised and their intelligibility degraded, which incurs the risk of creating unintelligible explanations for the sake of controlling their modality. Additionally, using a modality of an explanation that differs from the one used in real life is likely to impair the generalisability and ecological validity of our study. We therefore present each explanation in its most common modality.

3.3 Mixed Factorial Design

Following the discussion in Sections 3.1 and 3.2, we use a mixed factorial design in our experiment, with two ML models – logistic regression or decision tree – as between-subject factors, and four explanation types – model architecture, decision surface visualisation, feature importance and counterfactual explainability – as within-subjects factors.

By varying the ML models, we aim to explore if the model type plays a role in the comprehensibility of their explanations. Among four explanations for the same ML model, we use the model architecture as the baseline to study the extent to which the model-agnostic explanations improve or deteriorate comprehensibility compared with presenting the raw model itself.

3.4 Use Case

We place our user study in the context of chronic disease diagnosis, where data-driven predictive models have high impact and are subject to regulation [25], thus requiring validation and certification [84]. Predictive models in clinical diagnostics is further subject to additional regulatory scrutiny such as the European In Vitro Diagnostic Medical Devices Regulation, which establishes the need for XAI tools in the medical domain [56].Similar framing has also been used in previous XAI research [6, 71]. Additionally, using inherently interpretable models is desirable in high stakes domains such as healthcare [22, 34, 68]. Evaluating the intelligibility of explanations in this context is therefore informative and ensures the ecological validity of our findings.

With the emergence of digital healthcare systems, it is important to make such systems informative and accessible to their lay users. Our use case adheres to a common scenario where an at-home healthcare system relies on easy-to-collect biomarkers to help a user keep track of their health status. Specifically, a data-driven model informs a person if they are at high risk of developing a chronic disease, which is intended to prompt them to seek professional medical advice. By explaining such algorithmic predictions, users do not need to understand the physiological processes that underlie a particular medical condition but rather know how to manage it. Therefore, data features should be familiar to non-experts and their number manageable. The target audience of our study is thus the lay population (not medical professionals) who do not necessarily have knowledge of medicine or AI.

To support the external validity and reproducibility of the study, we use the UCI diabetes dataset [74], which is a real-world dataset frequently used in machine learning research [12, 42]. The target variable in the original dataset is to predict whether a patient has diabetes based on eight diagnostic measurements such as number of pregnancies, BMI, insulin level and age. To generalise the context of our study, we remove the *number of pregnancies* feature – since it is a gender-specific variable – leaving us a total of seven features. This feature set size is compatible with the human cognitive capacity as according to Miller [54] humans can simultaneously hold about seven items in working memory. In our experiments, we also modify the target variable in the binary classification task from *having diabetes* (label 1) or not (label 0) to *having a high risk* (label 1) or low risk (label 0) of developing the disease as we believe the latter task is closer to a real-world use case of data-driven tools in healthcare.We further reuse the models trained on the diabetes dataset across a series of scenarios where we introduce them as being deployed for the diagnosis of different chronic diseases, and specifically, to predict whether a person has high or low risk of developing such a condition. This set-up allows us to generalise the scope of our user study and avoid the priming effect (see Section 4.5 for a detailed discussion). The logistic regression model uses all seven features and achieves 81% accuracy; the decision tree model achieves the highest accuracy of 77% for which it only requires three of the features. We generate each explanation for a data point sourced from the dataset that is classified correctly by the underlying model, which ensures the validity of the explanatory insights.

3.5 Pilot Study

We conducted two rounds of pilot studies with 12 researchers from our research centre to fine-tune the explanations and the questionnaire used to evaluate user comprehension with the help of the participants' feedback. We did this to validate the participants' understanding of the survey procedure and to refine the wording of our questions, which helped us to control the difficulty of each question and eliminate any confusion. We also assessed their perception of the survey workload to determine the appropriate number of questions to ensure that sufficient data can be collected without the risk of user fatigue.

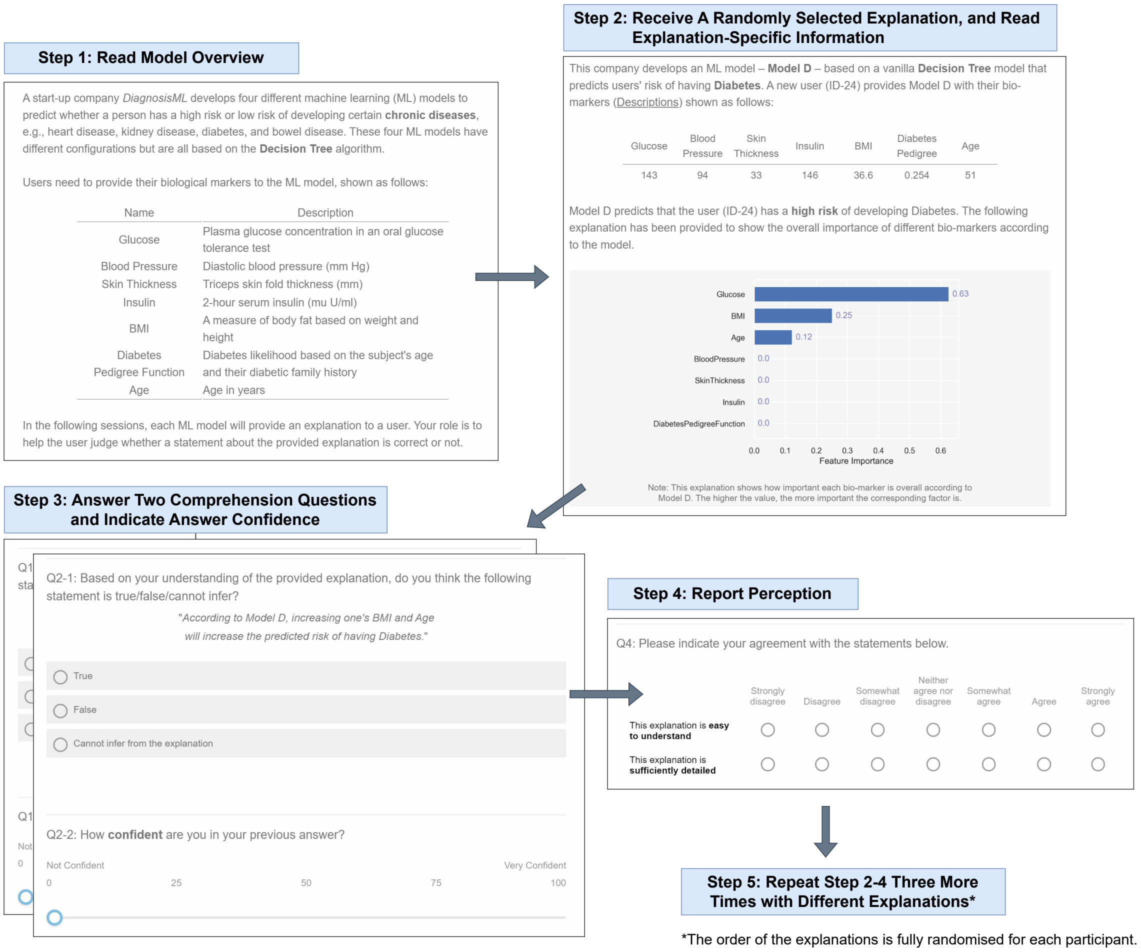
During the initial design phase of our experiment, we included two distinct scenarios, both in the healthcare domain but of different impact level: a high stakes scenario focused on chronic disease diagnosis and a lower stakes scenario concerned with a general health check. We attempted to evaluate whether user perception of explanations would diverge when the impact of the model differed while controlling for the confounding factor of the application domain. Most of the pilot study participants pointed out that they perceive both scenarios as high stakes despite the chronic disease diagnosis having bigger ramifications than advice to undergo a general health check. This feedback prompted us to drop the lower stakes use case in the final study. Finally, we ran two additional rounds of the pilot study on the crowd-sourcing platform Prolific with a total of 15 participants to validate the survey procedure and collect feedback from our target demographics.

4 USER STUDY

As discussed in Section 3.3, our user study has eight independent variables. Given that each variable requires a sample size $n = 20$ to achieve high statistical power [83], and to be conservative, we decided to recruit 200 participants. Before reporting our findings, we describe our participants, the study procedure and the design of our comprehension questions. The study has been approved by our institutional Human Ethics Advisory Committee.

4.1 Participants

Participants were recruited through the Prolific platform. To ensure sufficient response quality, we restricted participation to subjects with an acceptance rate of 95% or above, following the procedure outlined by van Berkel et al. [82]. We also limited the study to crowd workers based in the United Kingdom (UK). We used quotas for gender to ensure an even distribution of male and female participants. After accepting the task, the participants were directed to the Qualtrics platform, which hosted our experiment. Adhering to the UK minimum wage of £10.42 an hour at the time of our study and an expected completion time of 15 minutes (determined based on our pilot data), we compensated each participant with £2.75 (equivalent to £11 per hour) for completing our study. We included three attention-check questions in the experiment; the answers to these questions were presented on the same page so that no memory recall was required. An example of such a question is "Which information is provided by this user to the ML model?", with possible responses "(A) Insulin", "(B) Air Pollution", "(C) Smoking" and "(D) Vitamin Supplements". We did not use responses from participants who failed two or more – out of three – attention-check questions, but they were still compensated for the task.



Step 1: Read Model Overview

A start-up company *DiagnosisML* develops four different machine learning (ML) models to predict whether a person has a high risk or low risk of developing certain **chronic diseases**, e.g., heart disease, kidney disease, diabetes, and bowel disease. These four ML models have different configurations but are all based on the **Decision Tree** algorithm.

Users need to provide their biological markers to the ML model, shown as follows:

Name	Description
Glucose	Plasma glucose concentration in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-hour serum insulin (mu U/ml)
BMI	A measure of body fat based on weight and height
Diabetes Pedigree Function	Diabetes likelihood based on the subject's age and their diabetic family history
Age	Age in years

In the following sessions, each ML model will provide an explanation to a user. Your role is to help the user judge whether a statement about the provided explanation is correct or not.

Step 2: Receive A Randomly Selected Explanation, and Read Explanation-Specific Information

This company develops an ML model – **Model D** – based on a vanilla **Decision Tree** model that predicts users' risk of having **Diabetes**. A new user (ID-24) provides Model D with their bio-markers (Descriptions) shown as follows:

Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree	Age
143	94	33	146	36.6	0.254	51

Model D predicts that the user (ID-24) has a **high risk** of developing Diabetes. The following explanation has been provided to show the overall importance of different bio-markers according to the model.

Note: This explanation shows how important each bio-marker is overall according to Model D. The higher the value, the more important the corresponding factor is.

Step 3: Answer Two Comprehension Questions and Indicate Answer Confidence

Q2-1: Based on your understanding of the provided explanation, do you think the following statement is true/false/cannot infer?

"According to Model D, increasing one's BMI and Age will increase the predicted risk of having Diabetes."

○ True

○ False

○ Cannot infer from the explanation

Q2-2: How **confident** are you in your previous answer?

Not Confident — Very Confident

0 25 50 75 100

Step 4: Report Perception

Q4: Please indicate your agreement with the statements below.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
This explanation is **easy to understand**	○	○	○	○	○	○	○
This explanation is **sufficiently detailed**	○	○	○	○	○	○	○

Step 5: Repeat Step 2-4 Three More Times with Different Explanations*

*The order of the explanations is fully randomised for each participant.

Fig. 1. Overview of the user study workflow employed to assess user comprehension and their perception. In the actual survey, Step 1 is shown in a separate screen; once a participant clicks Next, Steps 2–4 are shown in a new screen simultaneously. Clicking Next again leads to an updated screen showing Steps 2–4 for a new explanation. For each explanation, the participant can hover their mouse over "Descriptions" (in Step 2), which will trigger a drop-down page showing information from Step 1. Therefore, the participant can always revisit model information in case they forget the details.

4.2 Procedure

Participants first saw the consent form that described the study and the tasks involved. After consenting, participants were assigned to one of two conditions at random: logistic regression or decision tree. We restricted participants from taking part in our experiment more than once, yielding 101 participants in the logistic regression and 99 in the decision tree condition. A sample questionnaire received by our participants is shown in Figure 1.[1] The high-level survey flow can be found in A, with the main steps outlined below.

[1] The survey data and code used for our analysis are available at https://github.com/xuanxuanxuan-git/hcxai.

(1) **Model overview.** Participants first read an introduction and a high-level summary of the predictive task – chronic disease diagnosis – along with information about four different configurations of an ML model – either logistic regression or decision tree – that were developed; each configuration corresponds to a different disease diagnosis to mitigate the priming effect (see Section 4.5 for a detailed discussion). We also provided a description of the data features (see Step 1 in Figure 1). Participants were then informed about the tasks they needed to perform.

(2) **Evaluation questionnaires.** We evaluate each explanation – specifically, its intelligibility – from two perspectives: users' objective understanding and subjective assessment, which is popular in the XAI literature [7, 8, 11].

- **Read explanation-specific information.** Participants were presented with one of the four ML explanation types selected at random – model architecture, decision surface visualisation, feature importance or counterfactual explainability – and a short description of the purpose of this explanation (Step 2 in Figure 1). Each explanation was generated to explain a model output corresponding to a random data profile, i.e., a data point, such that each explanation delivered different information. More details about how each explanation is introduced are provided in Section 4.3.Participants were then asked to complete the following questionnaire.
- **Comprehension questions.** We employed comprehension questions (described below) to assess participants' objective comprehension and misinterpretation of ML explanations. Each comprehension question consisted of an assertion that the participants were asked to judge based on their understanding of an explanation (Step 3 in Figure 1). Specifically, we asked: "Based on your understanding of the provided explanation, do you think the following statement is true/false/cannot infer?", with the possible answers being "true", "false" and "cannot infer from the explanation". As discussed in Section 3.2, the information explicated by an explanation was determined by its scope and type, therefore each explanation required bespoke comprehension questions. We designed two types of questions – comprehension of *explicated* and *unspecified* information – to capture two aspects of user understanding – explanation comprehension and over-interpretation.
  - **Comprehension of *explicated* information.** We measured whether participants can understand the insights communicated by an explanation through assertions about *explicated* information. The factual correctness of these assertions can be determined solely based on the information provided by the corresponding explanations. In other words, the answers to assertions about *explicated* information can only be "true" or "false". Examples of such assertions are:
    - ∗ "model D uses all 7 biomarkers while making predictions, and Glucose is the most important factor for model D" for *feature importance*; and
    - ∗ "assuming that all other biomarker values remain the same (including BMI), increasing this person's Glucose to 135 will change their prediction from low risk to high risk" for *decision surface visualisation.*

    For these examples, the former statement was "true" while the latter was "false"; participants were assigned to either a "true" or a "false" statement at random for this type of question.
  - **Comprehension of *unspecified* information.** We also measured whether participants know what information an explanation does not communicate to assess whether they can recognise the limitations of explanatory insights. To this end, we designed assertions that, while relevant to the

explanations, could not be answered based on their content since information needed to judge these assertions was unspecified by the explanations. Examples of such assertions are:

- ∗ "BMI and Glucose are the MOST influential factors (among all 7 factors) in determining this person's result" for *decision surface visualisation*; and
- ∗ "according to Model D, increasing one's BMI and Insulin would increase the predicted risk of having Diabetes" for *feature importance*.

The correct answer to all questions in this category is "cannot infer from the explanation".

In summary, for each explanation, participants were shown one question about explicated information (to which the correct answer was either "true" or "false") and one question about unspecified information (to which the correct answer was "cannot infer"). The order in which the two questions were displayed was randomised. The complete list of comprehension questions for each explanation is provided in B. The design principles of our comprehension statements are discussed in Section 4.4.

- **Answer confidence.** Every quantitative (objective) comprehension question was accompanied by a request to report the participant's confidence in their answer on a range of 0 to 100 (Step 3 in Figure 1), following a similar practice in other XAI user studies [29].
- **Explanation perception.** We asked participants to report their agreement with "this explanation is *sufficiently detailed*" on a 7-point Likert scale, ranging from "strongly disagree (1)" to "strongly agree (7)", which question was adapted from a similar user study [28]. Participants were also asked to report their perceived difficulty of comprehending the explanation by rating their agreement with "the explanation is *easy to understand*" using the same scale, which question was adapted from a similar user study [45]. See Step 4 of Figure 1 for reference.
- **Repeat.** Participants repeated all the actions listed under the *evaluation questionnaires* item with a different and randomly selected explanation for a total of four times per experiment. Each explanation was introduced in a distinct medical context – diagnosis of a different disease – for a different individual – previously unseen data instance – to prevent participants from accumulating information about the underlying predictive model after interacting with previous explanations.

(3) **Graph literacy.** We assessed the graph literacy of the participants using the Short Graph Literacy (SGL) scale [59], following the set-up reported in prior XAI research [82]. SGL is a validated questionnaire consisting of four questions, each based on a visual graph.

(4) **Demographic information.** We asked participants to report their age, gender and educational attainment. In addition to the basic demographic information, we also asked participants to report their literacy in machine learning algorithms, English proficiency level, technical background and their experience with similar XAI user studies. To evaluate *algorithm literacy*, we asked participants to report their knowledge of ML algorithms on a 5-point Likert scale; we adapted this question from a prior study [86]. To measure users' *technical background*, we asked participants to report their participation in the field of Science, Technology, Engineering and Mathematics (STEM); we adapted this question from a prior study as well [64]. We also asked users to report their *English proficiency level*, which was categorised into six proficiency levels according to the Common European Framework of Reference for Languages (CEFR) [62]. Lastly, we asked participants to report their previous participation in similar XAI user studies, ranging from "None (0)" to "A lot (probably more than 15)".

Table 2. Overview of explicated and unspecified information for each explanation assessed by our comprehension assertions.

	Feature Importance	Decision Surface	Counterfactual	Model Architecture
Scope	global	local	local	global
Explicated Information	global feature importance	change in a prediction based on values of *two* features	effect of a feature vector change on the prediction	model structure (simulatability)
Unspecified Information	local feature influence	(local) importance of the two displayed features	(local) importance of the two listed features	(combined) impact of features

4.3 Explanation Display

Given that each explanation is generated for a specific ML predictor and an individual prediction result, we always provided the context and usage of an explanation before introducing it to the participants. For each explanation, we first described the predictor and its usage (e.g., predicting the risk of having diabetes). Then, we provided the data profile of an anonymous individual and the prediction result they received from the ML model (e.g., high risk or low risk of having diabetes). Next, we provided an explanation along with a short textual description of the explanation in a grey box for improved visualisation; for example, feature importance was described as: "This explanation shows how important each biomarker is overall according to Model D. The higher the value, the more important the corresponding factor is." Within the same web page, the participants were asked to complete the questions for objective and subjective assessment. Clicking the *Next* button directed the participants to the next explanation, which dealt with a new ML predictor and data profile. The participants could not go back to see previous explanations as these are independent of each other and the information they provide should be analysed separately. All the explanations used in our study and their description are shown in B. An example of a complete questionnaire distributed to a participant is shown in Figure 10 displayed in A.

4.4 Comprehension Questions

The statements used in the questionnaire to evaluate objective comprehension were distinct for each explanation type. These questions quizzed participants about their core understanding of both the information explicated and unspecified by an explanation. Specifically, statements about *unspecified information* asked participants about details that were relevant to an explanation but cannot be uniquely determined. For example, our decision surface visualisation communicates how changes in the values of two features influence the prediction of the underlying model for a particular user, but such an explanation lacks information about the importance of these two features. For this explanation, we employed a question to assess whether participants can predict the new output of the model given different values of these two features as *explicated information*; and another question about whether they incorrectly inferred *unspecified information* – feature importance – from this specific explanation. The rationale behind the latter question is that unspecified information can be easily misconstrued in this case. Our design – informed by an XAI expert and improved iteratively during the pilot studies – ensures that the questions resemble each other and are comparably difficult to judge across explanation types.

Given that local explanations – such as decision surface visualisation – provide insights that are limited in scope, a wide range of information remains (implicitly) unspecified by this explanation type. Since all four explanation types used in our study entailed insights about features, we designed comprehension questions purely around this explanatory aspect. Specifically, the statements either quizzed the participants about the output of the model given a particular change to the feature vector or about feature importance and influence. An overview of explicated and unspecified information relevant to each explanation type is displayed in Table 2. While focusing predominantly on feature-based explanatory information implicitly limits the scope of our results, our study offers a principled evaluation of explicated and unspecified explanatory information that is otherwise overlooked by the current literature.

4.5 Priming Effect

Given that the explanation type is a within-subject factor, it is crucial to address the priming effect in our study design. To this end, we fully randomised the order of four explanation types presented to each participant. To remove the risk of participants accumulating information across tasks, we framed each explanation in the context of a different disease diagnosis. Specifically, for each participant, the ML model was presented in four different configurations, each under a different pseudo-name such as “Model B” and tasked with diagnosing one of the four diseases: diabetes, heart disease, kidney disease and bowel disease. We further introduced a new user profile with its unique biomarkers for which an explanation was generated for each explanation type. Each such profile was assigned with a different user ID, e.g., ID-24. The use case description also emphasised that each explanation was provided in a different context. By doing so, participants had limited chances of learning between explanations. After collecting the survey responses, we further checked the priming effect and verified that participants’ performance on the comprehension questions did not increase as they progressed through the survey (see D for more details).

5 RESULTS

201 participants completed the study from which 200 responses were included in the final analysis because one participant failed the attention check. The demographic information of our participants is summarised in Table 3. The average completion time of our study was 13.35 minutes, with a minimum completion time of 4.72 minutes. Participants’ average graph literacy score – assessed through the Short Graph Literacy scale – was 2.46 (*SD*=1.06, ranging from 0 to 4), which is close to the average score of 2.2 observed by the original study done on the US population [59]. 79% of the participants indicated that they have “no knowledge” or “negligible knowledge” of machine learning algorithms; 69.5% of the participants indicated that they do not have STEM background. The population of our respondents is aligned with the envisaged target audience, namely stakeholders who lack technical expertise.

5.1 User Comprehension

In order to identify the effect of *ML model type* and *explanation type* on the participants’ comprehension of explicated and unspecified information, we constructed statistical regression models to analyse the correlations. In each statistical model, the ML model and explanation type are the independent variables, and the dependent variable is whether a participant’s answer to the comprehension question about explicated or unspecified information is correct or not. The modelling outcomes are summarised in Table 4. Models 1 and 3 examine the main effects of ML model and explanation type on user comprehension of explicated and unspecified information respectively. Models 2 and 4 include the interaction effects between ML model and explanation type. To ensure the validity of statistical modelling, we checked for the existence of multicollinearity among model parameters. We found that the variance inflation factors of the parameters are in the 1–1.5 range, which is below the threshold of 5 or 10 indicative of multicollinearity [27].

Model 1 in Table 4 shows that feature importance and decision surface visualisation have significant main effects on the user comprehension of *explicated information*, which is visualised in Figure 2a. Additional McNemar’s tests

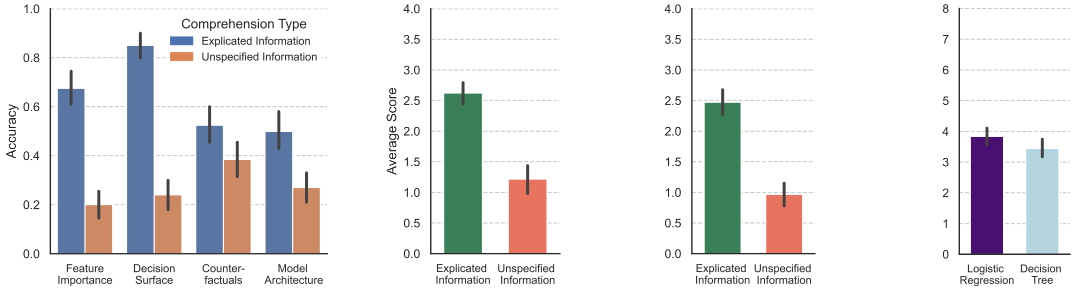
Table 3. Overview of major personal characteristics for our 200 participants.

Characteristic	Values	Frequency	Percentage
Gender	Female	99	49.5%
	Male	96	48.0%
	Non-binary	5	2.5%
Age	18–24 years old	37	18.5%
	25–34 years old	49	24.5%
	35–44 years old	44	22.0%
	45–54 years old	31	15.5%
	55–64 years old	22	11.0%
	65+ years old	17	8.5%
Educational attainment	Less than high school degree	4	2.0%
	High school graduate	38	19.0%
	College degree	30	15.0%
	Bachelor's degree	86	43.0%
	Graduate or professional degree	42	21.0%
ML algorithm literacy	No knowledge	95	47.5%
	Negligible knowledge	63	31.5%
	Some knowledge	35	17.5%
	Moderate knowledge	7	3.5%
Technical background	Yes (STEM-related education/employment)	61	30.5%
	No	139	69.5%
Experience with XAI user study	None (0)	125	62.5%
	A few (roughly 1–5)	69	34.5%
	A fair amount (around 6–15)	4	2.0%
	A lot (probably more than 15)	2	1.0%

Table 4. Coefficients, standard errors (in brackets) and significance indicators (* for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$) of statistical regression models used to assess the effect of ML model type and explainability approach on user comprehension. Models 1 and 3 show main effects on user comprehension of explicated and unspecified information respectively; Models 2 and 4 include interaction terms. Note that these statistical regression models are used to explore the correlation between variables (i.e., model type and explainability approach) and user comprehension, not for predicting user comprehension, therefore we focus more on their coefficients and p-values rather than the goodness of their fit (Pseudo-R^2).

	Explicated Information		Unspecified Information	
	Model 1	Model 2	Model 3	Model 4
Model Type: Logistic Regression	0.18 (0.15)	0.04 (0.28)	0.32 (0.16)*	0.59 (0.33)
Explanation Type: Feature Importance	0.74 (0.21)***	0.50 (0.29)	-0.39 (0.24)	-0.34 (0.37)
Explanation Type: Decision Surface	1.75 (0.24)***	1.75 (0.35)***	-0.16 (0.23)	0.00 (0.35)
Explanation Type: Counterfactual	0.10 (0.20)	0.08 (0.29)	0.53 (0.22)*	0.84 (0.32)**
Logistic Regression + Feature Importance	—	0.50 (0.42)	—	-0.11 (0.49)
Logistic Regression + Decision Surface	—	-0.02 (0.49)	—	-0.29 (0.47)
Logistic Regression + Counterfactual	—	0.04 (0.40)	—	-0.58 (0.44)
Pseudo-R^2	0.07	0.08	0.03	0.03

between pairs of explanations confirm that participants were significantly more likely to have accurate comprehension of explicated information communicated by feature importance (μ=0.68, *SD*=0.47) and decision surface visualisation (μ=0.85, *SD*=0.36); this is in contrast to counterfactual explainability (μ=0.53, *SD*=0.50) and model architecture (μ=0.50,



(a) Explanation type. (b) Logistic regression. (c) Decision tree. (d) Overall score.

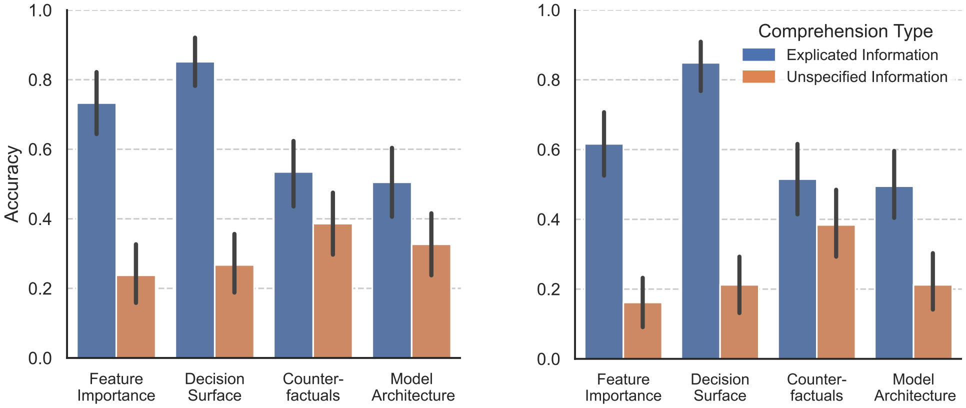
Fig. 2. Accuracy of the participants' answers to the question about explicated and unspecified information stratified by explanation type is shown in Panel (a). Participants were significantly more likely to understand explicated information shown in feature importance and decision surface visualisation but less attuned to their unspecified information, in contrast to counterfactuals and model architecture. The average user comprehension score for the four questions about explicated information and the four questions about unspecified information grouped by explanation type is shown in Panel (b) for logistic regression and Panel (c) for decision tree. Participants were significantly more likely to have correct comprehension of the information unspecified by the explanations of logistic regression compared to decision tree. Average score for all comprehension questions (including all eight questions about explicated and unspecified information) stratified by the type of ML model is shown in Panel (d). All error bars indicate 95% confidence interval.

Table 5. $\mathcal{X}^2$ and significance indicators (* for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$) of McNemar's test used to assess whether user comprehension of explicated and unspecified information is identical across explanation types.

	Explanation Type			
	Feature Importance	Decision Surface	Counterfactual	Model Architecture
Logistic Regression	39.06***	49.01***	3.95 *	4.50*
Decision Tree	35.53***	57.52***	3.07	17.82***
All	74.59***	106.31***	7.0**	18.24***

SD=0.50). Statistics of our pairwise McNemar's tests can be found in Table 9 included in C. Model 2 shows no significant interaction effects between explanation type and ML model on comprehension of explicated information.

Model 3 shows that the ML model type has a significant main effect on user comprehension of *unspecified information*, which is visualised in Figures 2b & 2c. Our results indicate that participants were significantly more likely to recognise the limitations of logistic regression explanations, i.e., their unspecified information. This means that when participants were working with ML explanations of logistic regression, they were significantly more likely to identify the constraints of these explanations (but less likely to do so for decision tree). We do not find a significant difference in comprehension of explicated information between the two ML models. The overall comprehension of explanations between these ML models was not significant either as shown in Figure 2d. Model 3 also indicates that counterfactual explainability has a significant effect on comprehension of unspecified information – see Figure 2a for a visual representation. Additional McNemar's tests confirm a significantly higher comprehension level for counterfactuals (μ=0.39, *SD*=0.49) compared to model architecture ($\mathcal{X}^2$=5.82, p=0.02), decision surface visualisation ($\mathcal{X}^2$=16.49, p <0.001) and feature importance ($\mathcal{X}^2$=18.75, p <0.001). No significant interaction effects are observed for comprehension of unspecified information according to Model 4.



(a) Logistic regression.

(b) Decision tree.

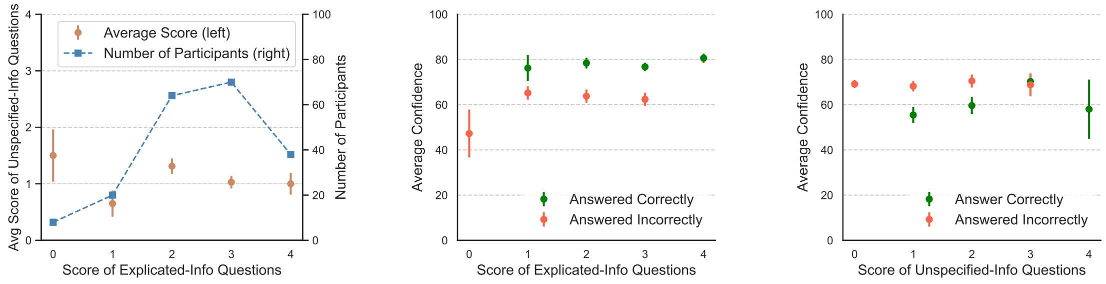
Fig. 3. Accuracy of participants' answers to comprehension questions about explicated and unspecified information grouped by explanation type, separately for our two ML models. Error bars indicate 95% confidence interval. Participants' comprehension of explicated information is substantially more accurate than their comprehension of unspecified information for every type of explanation of logistic regression and three explanations of decision tree (other than counterfactual explainability).

In addition to statistical regression modelling, we also performed statistical tests to measure whether user comprehension of explicated and unspecified information is identical for each explanation separately. The statistics of McNemar's tests with continuity correction are summarised in Table 5. Test results show that for every explanation of logistic regression the proportion of participants who correctly identified explicated information was significantly different from the proportion of participants who successfully recognised unspecified information. This significant difference is visualised in Figure 3a and indicates that participants' comprehension of explicated information is substantially more accurate than their comprehension of unspecified information for every explanation type of logistic regression. A similar phenomenon can be observed for explanations of decision tree other than counterfactual explainability as demonstrated in Figure 3b. After aggregating the participants' responses for the two ML models, McNemar's tests confirm that the proportion of participants who have developed correct comprehension of explicated information was statistically different from the proportion who accurately identified unspecified information for every explanation type.

5.2 Comprehension and Confidence

We further compare the participants' performance on the comprehension questions about explicated and unspecified information and their confidence level. As shown in Figure 4a, regardless of how well the participants performed in the questions about explicated information, they had consistently worse performance in identifying unspecified information. One-way ANOVA test confirms that the difference in average performance on the questions about unspecified information was insignificant across participants with different comprehension levels of explicated information ($F_{(200)}$=1.98, p=0.10).

The participants' confidence in their answers to the comprehension questions is visualised in Figure 5a. A paired Student's t-test (t=6.90, p <0.001) shows that participants were significantly more confident in their answers to the comprehension questions about explicated information (μ=72.36, SD=17.24) in comparison to the questions pertaining to unspecified information (μ=66.67, SD=17.04) regardless of the correctness of their comprehension. Additional paired t-tests confirm that the aforementioned observation is significant for decision surface visualisation (t=6.95, p <0.001),

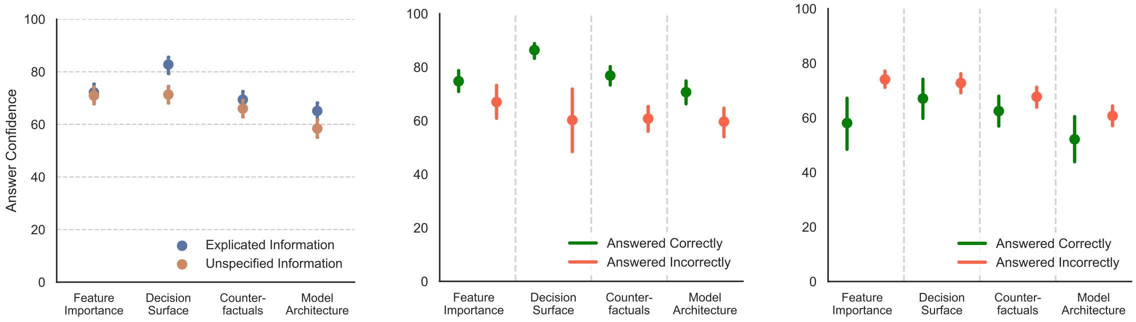


(a) Performance comparison.

(b) Explicated information.

(c) Unspecified information.

Fig. 4. Overview of participants' performance for questions about explicated and unspecified information and their confidence level. Panel (a) suggests that the participants with different levels of performance on the questions about explicated information attained comparable average score for the questions about unspecified information. Panel (b) shows participants with different levels of performance for questions about *explicated information* and the confidence in their answers. Panel (c) shows participants with different levels of performance for questions about *unspecified information* and the confidence in their answers.



(a) Aggregated.

(b) Explicated information.

(c) Unspecified information.

Fig. 5. Overview of participants' confidence in their answers to comprehension questions about explicated and unspecified information is shown in Panel (a). Answer confidence of participants who identified *explicated information* correctly (displayed in green) and incorrectly (displayed in red) is shown in Panel (b). Answer confidence of participants who identified *unspecified information* correctly (displayed in green) and incorrectly (displayed in red) is shown in Panel (c). Participants who answered the questions about *explicated information* correctly reported significantly higher confidence in their answers than their peers who did not. On the other hand, participants who answered the questions about unspecified information correctly were less confident than those who answered them incorrectly. Error bars indicate 95% confidence interval.

counterfactual explainability (t=2.07, p=0.04) and model architecture (t=4.41, p <0.001); however, we do not observe a significant difference in answer confidence for feature importance.

We further compare answer confidence between participants who exhibited correct comprehension and those who did not, which difference is visualised in Figures 5b & 5c. We find that participants who answered the questions about *explicated information* correctly reported significantly *higher confidence* in their answers than their peers who did not. This is confirmed with additional independent t-tests for each explanation type: feature importance (t=2.2, p=0.03), decision surface visualisation (t=6.12, p <0.001), counterfactual explainability (t=5.27, p <0.001) and model architecture (t=3.13, p=0.002). On the other hand, participants who answered the questions about *unspecified information* correctly were *less confident* than those who answered them incorrectly. Independent t-tests show that this difference in

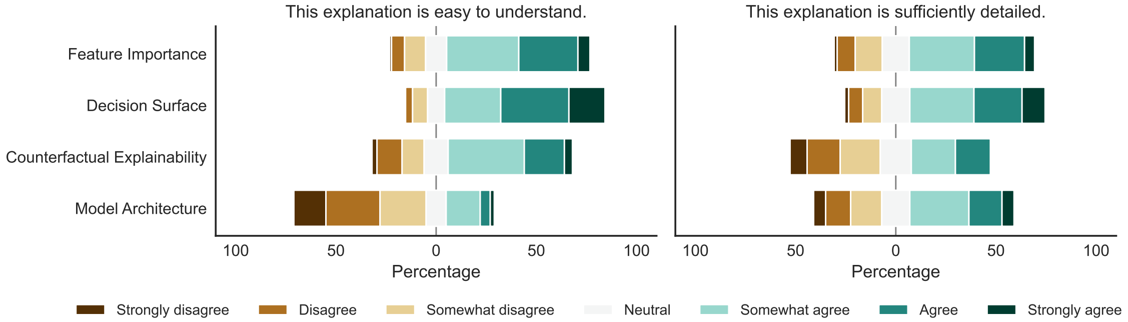


Fig. 6. Response overview for the Likert scale questions stratified by explanation type. Across both questions, the perception was significantly more negative towards model architecture and counterfactual explainability than for feature importance and decision surface visualisation.

confidence of unspecified information was significant for feature importance (t=-4.02, p <0.001) and model architecture (t=-2.20, p=0.03), but insignificant for decision surface visualisation and counterfactual explainability.

We also explore if participants with varying objective comprehension performance have different confidence levels in their answers to the questions about *explicated* and *unspecified information*. As shown in Figures 4b & 4c, participants with different levels of comprehension had consistent levels of confidence across their answers to both question types. This indicates that a higher level of comprehension does not guarantee better confidence calibration.

5.3 Subjective Assessment

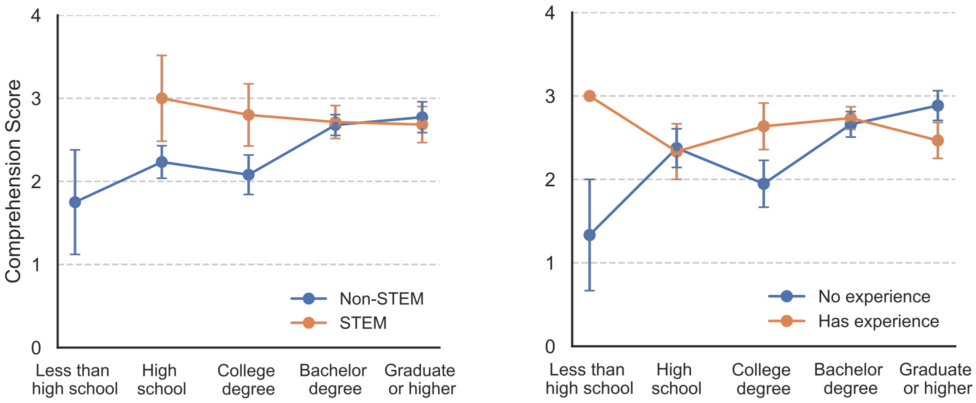
Next, we analyse the participants' Likert-scale responses to "this explanation is easy to understand" and "this explanation is sufficiently detailed" stratified by explanation type, which are shown in Figure 6. A repeated measures ANOVA test shows that the user perception of comprehension difficulty was significantly diverged across explanation types ($F_{(3,591)}$=121.61, p <0.001).[2] Specifically, perception of model architecture and counterfactual explainability was significantly more negative than of feature importance and decision surface visualisation (see Tukey's post-hoc paired test results reported in Table 10 provided in C). Additional repeated measures ANOVA tests also show that responses regarding the difficulty of understanding were significantly different across explanations if we evaluate them separately for logistic regression ($F_{(3,300)}$=101.0, p <0.001) and decision tree ($F_{(3,294)}$=35.79, p <0.001) – see Figure 11 in C for more details.

In terms of the participants' appreciation of the explanations' richness of information, repeated measures ANOVA tests indicate that the responses were significantly diverged across explanation types, both at an aggregate level ($F_{(3,591)}$=26.90, p <0.001) as well as separately for logistic regression ($F_{(3,300)}$=13.54, p <0.001) and decision tree ($F_{(3,294)}$=15.10, p <0.001). Specifically, the participants were significantly more discontent with counterfactual explainability compared to other explanation types – see Table 11 in C for more details.

[2] While parametric tests such as ANOVA are in principle only applicable to continuous data – and Likert scale offers discrete, ordinal data – Norman [58] showed that parametric tests offer meaningful results that are generally more robust than non-parametric tests in such situations, which practice has been adopted by other XAI studies [6]. Sullivan and Artino Jr [81] also noted that ordinal variables with five or more categories can often be used as continuous without any negative impact on the analysis.

Table 6. Coefficients, standard errors (in brackets) and significance indicators (* for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$) of the generalised linear models used to analyse the effect of individual characteristics on user comprehension scores. Models 5 and 7 show the main effect; Model 6 includes interaction terms.

	Model 5 (Aggregated)	Model 6 (Explicated)	Model 7 (Unspecified)
Age	0.00 (0.07)	0.25 (0.18)	0.01 (0.05)
Gender: Male	0.12 (0.21)	-0.22 (0.53)	0.01 (0.16)
STEM: Yes	0.39 (0.24)	2.31 (0.77)**	0.26 (0.18)
Education	0.33 (0.10)**	0.50 (0.21)*	0.17 (0.07)*
Machine Learning Literacy	0.14 (0.14)	-0.62 (0.40)	0.10 (0.11)
Experience with XAI User Studies: Yes	-0.10 (0.21)	1.35 (0.59)*	-0.15 (0.16)
Education + Age	—	-0.07 (0.05)	—
Education + Male	—	0.08 (0.14)	—
Education + STEM	—	-0.56 (0.19)**	—
Education + Machine Learning Literacy	—	0.17 (0.10)	—
Education + Experience with XAI User Studies	—	-0.35 (0.16)*	—
R^2	0.117	0.111	0.070
Adjusted R^2	0.089	0.058	0.040



(a) Education vs. STEM background.

(b) Education vs. user study experience.

Fig. 7. Average score for the four questions about *explicated information* stratified by education level and (a) STEM background or (b) XAI user study experience. On average, participants who had a STEM background or experience with XAI user studies achieved *lower* comprehension scores as their education level increased. On the contrary, participants without STEM background or XAI user study experience attained *higher* scores as their education level increased. Error bars indicate standard error.

5.4 Individual Difference

We further explore the effect of individual characteristics on explanation comprehension. The major personal characteristics we consider in our user study include age, gender, STEM background, education level, machine learning literacy and experience with XAI user studies. We treat age, ML literacy and education as ordinal data; gender as categorical data; and the remaining variables as binary data. Our sample consists of five non-binary participants, who we exclude from our analysis in this section due to their limited representation in our sample. Table 6 describes the generalised linear models that we use to analyse the effect of personal characteristics on the participants' scores for the comprehension questions about explicated and unspecified information as well as scores for all the questions. (Recall that each participant answered four questions about explicated information and four questions about unspecified information during the experiment, thus the score for each set of questions is in the 0–4 range and the total score is in the 0–8 range.)

Models 5 and 7 – detailed in Table 6 – describe main effects of individual characteristics on *overall comprehension* and *comprehension of unspecified information* respectively. Additional models were constructed to explore interaction effects in this setting, but since no significant interaction effects were observed we do not include these results. Model 6 explores main effects of individual characteristics on *comprehension of explicated information*; it identified significant main effect of education level on this type of comprehension. We further explore interaction effects between education level and the remaining five personal characteristics on comprehension of explicated information; the results are displayed in Table 6 as Model 6.

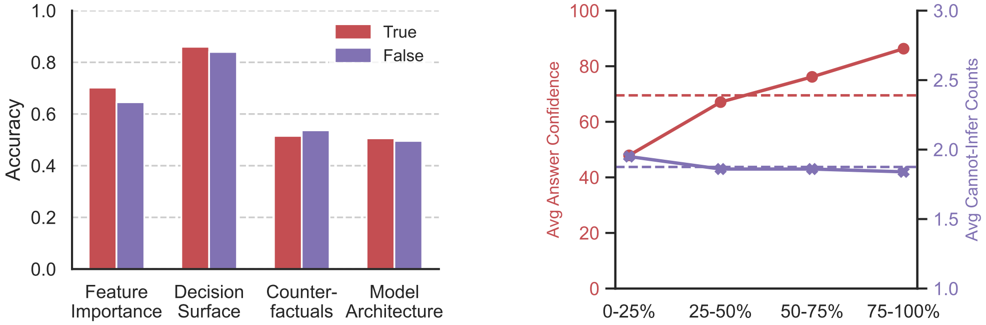
Model 5 shows that the participants' education level was positively correlated with overall comprehension, with the average comprehension score increasing as the attained education level becomes higher (*coef*=0.33, p <0.01). Models 6 and 7 confirm the significantly positive correlation between education level and comprehension for explicated (*coef*=0.50, p <0.05) and unspecified (*coef*=0.17, p <0.05) information. Model 6 shows that technical background had a significantly positive effect on the comprehension of *explicated information* (coef=2.31, p <0.01), with participants who had STEM background achieving higher score (μ=2.74, *SD*=1.00) than those without it (μ=2.46, *SD*=1.04). Experience with XAI user studies was also positively correlated with comprehension of *explicated information* (*coef*=1.35, p <0.05), with veteran participants achieving higher scores (μ=2.50, *SD*=0.88) than their inexperienced peers (μ=2.51, *SD*=1.12).

Model 6 also highlights two other significant interaction effects on user comprehension of *explicated information*: between education and STEM background (*coef*=-0.56, p <0.01), and between education and XAI user study experience (*coef*=-0.35, p <0.05). We visualise these two interaction effects in Figure 7. On average, participants with STEM backgrounds achieved lower comprehension scores as their education level increased. On the contrary, participants without STEM backgrounds attained higher scores as their education level increased. Similarly, participants who have never before participated in XAI user studies scored higher when their education level was higher; this trend is reversed – worse performance was associated with higher education levels – for participants who interacted with XAI user studies before. We discuss the implications of this finding in Section 6.4.

Finally, we analyse the relationship between participants' graph literacy and their XAI comprehension. Pearson correlation tests show that the correlation between the graph literacy score and correct answers to the questions about *explicated information* was significant, with the average comprehension score decreasing as the participants' graph literacy increased (*coef*=-0.20, p <0.01). We do not observe any significant relationship between graph literacy and overall comprehension score or comprehension score for unspecified information alone.

5.5 Participant Engagement Analysis

To verify the validity and robustness of our results, we conduct additional data analysis to study the participants' engagement with the user study. Given that the "true" assertions regarding counterfactual explanations are negatively framed, which may lead to the concern that they confuse participants, it is important to ensure a comparable accuracy between the "true" and "false" assertions across the explanations. We confirm this in Figure 8a, which shows that the participants who received the "true" assertion questions exhibited similar performance to those who received the "false" assertions. This suggests that the negatively framed assertions had negligible impact on user comprehension.



(a) Accuracy of the "true" and "false" explicated-information questions.

(b) Frequency of selecting "cannot infer" as the answer among different participant groups.

Fig. 8. Overview of the participants' engagement analysis. Panel (a) shows the accuracy of the answers to the "true" or "false" assertions grouped by explanation type. Both the "true" and "false" assertions are gauging user comprehension of explicated information. For each explanation, the difference in answer accuracy to the "true" and "false" questions is insignificant. Panel (b) shows the frequency of selecting the "cannot infer" option as the answer to the comprehension questions. We group participants into quartiles based on their average answer confidence, with each group having an equal number of participants. The first quartile (0–25%) consists of the participants with the lowest confidence and the fourth quartile with the highest confidence. The red dotted line indicates the average answer confidence over all of the participants; the purple dotted line indicates the average number of times a participant chooses "cannot infer" as the answer out of the eight comprehension questions. The figure shows that the participants whose confidence was low opted for the "cannot infer" option more frequently than average, however the difference is not significant.

We further explore if the participants actively engaged in more demanding tasks (e.g., detecting unspecified information) or whether instead they frequently opted for the "cannot infer" option as their default (passive) answer. Figure 8b shows that participants chose "cannot infer" option 1.875 times on average out of the eight questions they received. For participants with low confidence in their answers, they tended to choose "cannot infer" more frequently than average (1.95 times). However, the difference in frequency is small among the participants with different confidence levels. This suggests that participants engaged actively in detecting unspecified information instead of defaulting to the "cannot infer" option.

Lastly, we analyse the time that the participants spent on each explanation. We find that the participants spent more time on attempting to understanding global explanations, which communicate more information and involve more features: 3.16 minutes on model architecture and 3.11 minutes on feature importance. On the other hand, the participants spent less time on local explanations, which only present information about two features; specifically, 3.08 minutes on decision surface visualisation and 2.94 minutes on counterfactual explanation. This suggests that the participants invested more efforts in understanding more complex explanations as expected.

6 DISCUSSION

In this paper, we set out to understand whether users can correctly interpret factual insights conveyed by machine learning explanations and, at the same time, be aware of their inherent limitations. Current literature emphasises that bespoke explanatory mechanisms are needed to facilitate understanding of data-driven predictive systems as transparency by itself cannot guarantee intelligibility [29, 32, 45, 79]. Our findings complement this perspective by demonstrating that the **comprehension of explanations is a double-edged sword as highly intelligible**

explanations are likely to be misinterpreted despite their inherent truthfulness. In this section, we discuss our results with a particular focus on user comprehension and misinterpretation of explanations. Then, we suggest better explanation modelling strategies to reduce the prevalence of such undesired effects.

6.1 Explanation Misinterpretation

Our results demonstrate that **participants were good at correctly interpreting the information explicated by explanations, but struggled to identify the unspecified information**. For example, feature importance informs explainees of how much a model relies on each feature. Figure 2a shows that participants achieved a high comprehension level for the information explicated by feature importance compared to other explanation types. Nonetheless, participants also (incorrectly) inferred local feature influence – i.e., whether a feature contributes positively or negatively to a particular prediction – from such explanations despite this information being unavailable. We also find that users tend to identify the features used in the explanations based on decision surface visualisation and counterfactual explainability as the most important, even though such explanations do not (explicitly) account for this information. We hypothesise that participants may exhibit confirmation bias – a well-known cognitive error – which in this case led the explainees to assume that the feature affecting a particular prediction, as shown by the explanation, must be the most important one [22, 41]. Our results also indicate that users tend to identify the feature used by the root node of a decision tree as the most important one, which is not necessarily correct given that this property is determined by an impurity criterion rather than the tree structure. This finding aligns with the phenomenon observed by Bell et al. [4].

In terms of specific explainability approaches, our results show that participants were significantly more likely to understand explicated information when it is communicated by feature importance and decision surface visualisation rather than counterfactual explainability and model architecture. A similar finding was reported by Cheng et al. [11], who showed that white-box explanations (similar to our feature importance) increase explainees' objective understanding.[3] We additionally find that the former two explanation types were also more likely to be misinterpreted, i.e., the participants were prone to deduce unspecified information from them, evidenced by a low level of comprehension of unspecified information shown in Figure 2a. A similar finding was observed by Chromik et al. [13] who indicated that users overestimate the understanding they gain from local feature explanations because of the illusion of explanatory depth. Counterfactual explainability and model architecture were less intelligible to participants in terms of explicated information, but the participants were also less likely to be misled by these explanations. To summarise, our results show that, among the four explanation types, **when the information explicated by an explanation was more comprehensible, unspecified information was also more likely to be misconstrued by the explainees**.

We also find that participants did not achieve a high level of comprehension of explicated information for model architecture when compared to the remaining three explanation types – as shown in Figure 2a – even though this approach discloses the complete model structure, thus allows its in-vivo simulation. This aligns with the argument put forth by Sokol and Vogt [79] who noted that transparency is a prerequisite for understanding, but this property alone is insufficient for comprehensibility. Similarly, Bell et al. [4] argued that disclosing more information does not necessarily yield more comprehension as the effectiveness of information processing is also crucial. Furthermore, participants were less likely to correctly recognise the unspecified information in explanations produced for decision trees and more acute to it for logistic regression models. Given that the explicated information of the explanations generated for both models was equally comprehensible – refer to Figures 2b & 2c – this finding suggests that the explainees were more likely to

[3] The definition of objective understanding used by Cheng et al. [11] corresponds to our notion of comprehension of explicated information.

misinterpret explanations for predictors that were seemingly easy to understand. This aligns with the argument by Bell et al. [4] that a less complex predictor could be more misleading than a complex one. In summary, validating user comprehension of explicated information when evaluating the intelligibility of user-centred explanations is *insufficient*, it is equally critical to assess the degree of confusion and misinterpretation that XAI may lead to.

6.2 Over-generalisation and Over-confidence

Through the objective assessment of comprehension, we find that participants did not know how much they knew; in other words, they cannot correctly assess their state of knowledge. As shown in Figure 4a, having a high level of comprehension of explicated information did not stop users from over-generalising the explanations. We also incorporated subjective assessment results, and paired participants' subjective answer confidence with objective answer accuracy to evaluate whether the self-assessed confidence in their judgement exceeded the correctness of their comprehension. It is highly desirable for participants to correctly understand the factual insights conveyed by an explanation while also having high confidence in their comprehension. The opposite, however, should be avoided: participants should not overlook the limitations of an explanation and be confident in the misconstrued information.

In this work, we find that **participants who interpreted explicated information correctly also possessed higher confidence in their correct interpretation than others who incorrectly interpreted explicated information** across all explanation types. This observation suggests that the explanations led to high confidence that is well calibrated for explicated information. In contrast, **participants became overconfident in their incorrect interpretation of the information unspecified by explanations**, especially so for feature importance and model architecture. This indicates that not only the limitations of these explanations were difficult to identify, but also that the information unspecified by these two explanation types appeared convincing enough for the users to develop high confidence in their misinterpretation.

Prior work on overconfidence noted that humans' confidence exhibits low levels of calibration when individuals face difficult tasks [40, 48]. In this work, we observe a similar phenomenon where participants had a high degree of overconfidence when working on a more difficult task, namely answering questions about unspecified information, for which the correct response rate was significantly lower as demonstrated by Figure 2. This suggests that identifying the limitations of explanations is a more challenging task than just attempting to understand explicated information, in which case user confidence is more likely to be poorly calibrated. As Darwin [16] noted decades ago, "ignorance more frequently begets confidence than does knowledge". Our findings confirm that people who are ignorant of the limitations of an explanation are more overconfident than their peers who are more competent.

6.3 User Perception

In general, the **participants agreed that feature importance and decision surface visualisation were easier to understand and sufficiently detailed as compared to the other two explanation types**. The explainees believed that counterfactual explainability was harder to understand and not sufficiently detailed, despite this XAI approach being praised in the literature for its appeal, simplicity and brevity [35, 79, 85]. This suggests that simple explanations do not necessarily provide the information expected or needed by explainees, and the right balance between information load and simplicity should be considered on case-by-case basis. Model architecture was perceived as hard to understand and insufficiently detailed. This suggests that providing all information about an ML model does not necessarily guarantee that users find an explanation to be sufficiently detailed; one possible reason is the details being unintelligible due to the explanation complexity, which can be seen in Figure 2a as low comprehension of explicated information.

The participants' perception of whether an explanation is easy to understand corresponded to their comprehension level of explicated information. As shown in Figure 6, the participants' subjective agreement with easiness of understanding was high for decision surface visualisation but low for model architecture. By additionally considering Figure 2a, we find that the explanations for which people achieved high comprehension of explicated information but low comprehension of unspecified information were found to be universally easier to understand. In summary, **whenever an explanation was believed to be "easy to understand", it was subjectively more intelligible but at the same time also more misleading**.

6.4 Individual Characteristics

Correct and truthful understanding of an explanation relies on human reasoning [78, 79], but reasoning capacity varies between individuals, which in turn affects the degree of understanding. Human factors are therefore of paramount importance when studying user comprehension. Our results indicate that the **participants with higher educational attainment developed a more accurate understanding of both the information that an explanation explicates and the information that is unspecified**. Those with lower education levels were more prone to develop various misinterpretations. This finding agrees with the well-known Dunning–Kruger effect, which states that those with limited knowledge in a domain not only reach incorrect conclusions but their incompetence prevents them from realising their mistakes [40]. Similarly, Cheng et al. [11] found that explainees' education level has a main effect on their understanding of predictive algorithms.

We also observe that **users with technical background developed a more accurate interpretation of explicated information**, which confirms findings in prior work [43]. Furthermore, **participants who have taken part in XAI user studies before performed better in grasping the information explicated by explanations**. This suggests that ML explanations could be an effective educational tool to increase user knowledge of AI-related concepts [57]. On the other hand, Human–Computer Interaction researchers studying explainability should be cautious of crowd workers with prior XAI user study experience, and ideally include their historical participation in such experiments as a confounding factor. Since our user study strictly followed the standard participant recruitment protocol popular in current XAI literature, our results offer a warning that this factor may influence the participants' performance. While the individuals involved in user studies might gain XAI knowledge from their exposure, this nuanced factor cannot be easily captured through demographic questionnaires employed in current study designs. Future work could also explore the nature of this type of experience.

6.5 Explanation Complementarity

Providing information that is incomplete, partial or oversimplified may engender a false sense of knowledge and lead to overconfidence [44]. Our work uncovered that explanations for which the users struggled to identify their limitations and unspecified information were likely to be misinterpreted (with poor confidence calibration). Such misleading explanations could also possibly be misuse [60] and result in unwarranted trust [30]. Furthermore, according to Leichtmann et al. [47], a high level of trust in ML models should not be a goal in itself as excessive trust may lead to judgement errors and other unintended consequences due to unjustified overreliance. Therefore, when users' only interaction with predictive models is through explanations, it is crucial to help them develop an appropriate, well-calibrated level of confidence and trust in their comprehension. To this end, we posit that practitioners should use explanations – or combinations thereof that are complementary – with well-understood limitations that are either known to the explainees, easy to identify or explicitly communicated.

One solution, suggested by van Berkel et al. [82], is for researchers to explicitly indicate the information that is unspecified by (explanatory) artefacts shown to humans. However, we argue that simply disclosing the limitations of an explanation may not suffice to engender understanding and well-calibrated trust. This is because numerous explanations are available, each with a different scope – global, local or cohort – and information content – attribute importance, feature influence, counterfactual explainability, model architecture, and the like. It is thus impractical to identify and communicate all the pieces of information unspecified by a single explanation or their collection. Additionally, it remains to be seen whether explainees can understand and correctly apply such limitations and how to best communicate them.

As one possible approach to address this challenge, we suggest to **consider the alignment of different explanations, thus identify the ones that provide complementary information, covering the shortcomings and limitations of each other**. Complementary explanations promise to offer non-overlapping information without overwhelming users. As an example, our participants were likely to infer feature importance from decision surface visualisation and counterfactual explainability; by accompanying such explanations with feature importance, we can prevent incorrect insights from developing. In this context, we can consider feature importance to be complementary to decision surface visualisation and counterfactual explainability since it minimises the chances of explainees inferring insights that remain unspecified. Future work could investigate if complementary explanations can mitigate this type of misinterpretations as compared to standalone explanations.

6.6 Limitations and Future Work

We recognise several limitations of our work that should be considered when interpreting our results. (1) We only investigated four common explainability approaches, looking at the degree to which participants can identify explanatory information and its limitations. Future work could evaluate other XAI techniques and identify explanation types that lead to a high level of comprehension for both explicated and unspecified information. (2) As discussed in Section 3.1, we only explored inherently interpretable and transparent models, setting aside complex AI models for which we could not ensure perfect explanation fidelity. Future work could extend our study to black-box models, accounting for the imperfect fidelity of their explanations. (3) The scenario presented to our participants – chronic disease diagnosis – covered only the healthcare domain. Future work could explore other scenarios and application domains that span diverse levels of risk associated with data-driven decision-making. (4) We limited the design of our comprehension statements to quiz-style questions pertaining exclusively to information about data features. However, an explanation could provide multi-faceted information depending on its type and scope, hence users may deduce different kinds of unspecified information. Future work could thus expand the set of comprehension questions to cover more types of explicated and unspecified information. This approach would help to develop a comprehensive view of the information captured by each explanation and the limitations thereof. Such a framework could be used to better align explanations to provide complementary and reliable information. (5) The assertions vary across explanations as they are designed to capture the specific information communicated by different explanations. Thus one may find that the structure and difficulty of the assertions differ. As we followed the framework in Section 6.5 and analysed the impact of the assertion design in Section 5.5, we posit that the formulation of the assertions has a negligible impact on the study results. Future work could design a skeleton of assertion statements so that one can easily plug in different pieces of information to consistently formulate the assertion questions.(6) Finally, we note that our user study was limited to UK crowd workers and different demographics may produce different results.

7 CONCLUSION

Providing explanations that are intelligible to the general public, such that the recipients do not misconstrue the information that they carry, is of paramount importance. In this paper, we investigated whether common ML explanations provide intelligible insights and to what extent they are likely to be misinterpreted (when their limitations are overlooked). Specifically, we examined user comprehension of explicated and unspecified information for four representative explainability approaches applied to two popular types of predictive models: logistic regression and decision tree; this setup allowed us to assess user perception of explanation informativeness and difficulty. We found that comprehension can be a double-edged sword as highly intelligible explanations are often misinterpreted resulting in invalid beliefs. In particular, we showed that feature importance and decision surface visualisation offer highly comprehensible information but the explainees are ignorant of their limitations, thus misinterpret the insights that they provide. Further, our findings indicate that explanations perceived by the users as "easy to understand" are in fact objectively more comprehensible, but also more misleading. Accounting for the explainees' perception of the explicated and unspecified information can therefore inspire better design of the explanation content and presentation format, such that it provides complementary insights that address each other's limitations or address such shortcomings otherwise.

ACKNOWLEDGEMENTS

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (project number CE200100005), and funded by the Australian Government through the Australian Research Council. Additional funding was provided by the Hasler Foundation (grant number 23082).

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–18.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: The risk of rationalization. In *International Conference on Machine Learning*. PMLR, 161–170.

[4] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 248–266.

[5] Aditi Bhutoria. 2022. Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence* 3 (2022), 100068.

[6] Reuben Binns, Max van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 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*. 1–14.

[7] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *27th International Conference on Intelligent User Interfaces*. 807–819.

[8] Clara Bove, Marie-Jeanne Lesot, Charles Albert Tijus, and Marcin Detyniecki. 2023. Investigating the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: An Explanation User Interface Proposition and User Study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 188–203.

[9] Leo Breiman. 2017. *Classification and regression trees*. Routledge.

[10] Maud Chassignol, Aleksandr Khoroshavin, Alexandra Klimova, and Anna Bilyatdinova. 2018. Artificial Intelligence trends in education: A narrative overview. *Procedia Computer Science* 136 (2018), 16–24.

[11] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[12] Dilip Kumar Choubey, Sanchita Paul, Santosh Kumar, and Shankar Kumar. 2017. Classification of Pima Indian diabetes dataset using naïve Bayes with genetic algorithm as an attribute selection. In *Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016)*. 451–455.

[13] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think I get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[14] David Danks and Alex John London. 2017. Regulating autonomous systems: Beyond standards. *IEEE Intelligent Systems* 32, 1 (2017), 88–91.

[15] Valdemar Danry, Pat Pataranutaporn, Ziv Epstein, Matthew Groh, and Pattie Maes. 2022. Deceptive AI Systems That Give Explanations Are Just as Convincing as Honest AI Systems in Human-Machine Decision Making. *arXiv preprint arXiv:2210.08960* (2022).

[16] Charles Darwin. 1888. *The descent of man, and selection in relation to sex*. John Murray, Albemarle Street.

[17] Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 296–299.

[18] Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future Healthcare Journal* 6, 2 (2019), 94.

[19] Javier Del Ser, Alejandro Barredo-Arrieta, Natalia Díaz-Rodríguez, Francisco Herrera, Anna Saranti, and Andreas Holzinger. 2024. On generating trustworthy counterfactual explanations. *Information Sciences* 655 (2024), 119898.

[20] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[21] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[22] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.

[23] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* 24, 1 (2015), 44–65.

[24] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57.

[25] Government of Canada. 2023. Artificial Intelligence and Data Act. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act

[26] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–42.

[27] Joseph F Hair. 2009. Multivariate data analysis. (2009).

[28] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[29] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (2011), 141–154.

[30] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.

[31] Aishwarya Jakka and J Vakula Rani. 2023. An Explainable AI Approach for Diabetes Prediction. In *Innovations in Computer Science and Engineering: Proceedings of the Tenth ICICSE, 2022*. Springer, 15–25.

[32] Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 805–815.

[33] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. 2017. Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology* 2, 4 (2017).

[34] Jinglu Jiang, Surinder Kahai, and Ming Yang. 2022. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies* 165 (2022), 102839.

[35] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *Comput. Surveys* 55, 5 (2022), 1–29.

[36] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[37] John KC Kingston. 2016. Artificial intelligence and legal liability. In *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV 33*. Springer, 269–279.

[38] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*. PMLR, 1885–1894.

[39] Joshua Alexander Kroll. 2015. *Accountable algorithms*. Ph. D. Dissertation. Princeton University.

[40] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77, 6 (1999), 1121.

[41] Arie W Kruglanski and Icek Ajzen. 1983. Bias and error in human judgment. *European Journal of Social Psychology* 13, 1 (1983), 1–44.

[42] P Suresh Kumar and S Pranavi. 2017. Performance analysis of machine learning algorithms on diabetes dataset using big data analytics. In *2017 International Conference on Infocom Technologies and Unmanned Systems (ICTUS): Trends and Future Directions*. IEEE, 508–513.

[43] Samuli Laato, Miika Tiainen, AKM Najmul Islam, and Matti Mäntymäki. 2022. How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research* 32, 7 (2022), 1–31.

[44] Simone Lackner, Frederico Francisco, Cristina Mendonça, André Mata, and Joana Gonçalves-Sá. 2023. Intermediate levels of scientific knowledge are associated with overconfidence and negative attitudes towards science. *Nature Human Behaviour* 7, 9 (2023), 1490–1501.

[45] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 59–67.

[46] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

[47] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2023), 107539.

[48] Sarah Lichtenstein and Baruch Fischhoff. 1977. Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance* 20, 2 (1977), 159–183.

[49] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*. 2119–2128.

[50] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 56–67.

[51] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30 (2017).

[52] Bertram F Malle and Joshua Knobe. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology* 33, 2 (1997), 101–121.

[53] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management* 39, 1 (2022), 53–63.

[54] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 2 (1956), 81.

[55] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[56] Heimo Müller, Andreas Holzinger, Markus Plass, Luka Brcic, Cornelia Stumptner, and Kurt Zatloukal. 2022. Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. *New Biotechnology* 70 (2022), 67–72.

[57] Davy Tsz Kit Ng, Jac Ka Lok Leung, Kai Wah Samuel Chu, and Maggie Shen Qiao. 2021. AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology* 58, 1 (2021), 504–509.

[58] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education* 15 (2010), 625–632.

[59] Yasmina Okan, Eva Janssen, Mirta Galesic, and Erika A Waters. 2019. Using the short graph literacy scale to predict precursors of health behavior change. *Medical Decision Making* 39, 3 (2019), 183–195.

[60] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2 (1997), 230–253.

[61] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12 (2011), 2825–2830.

[62] Domain Locations Institutions Persons. 2001. Common European framework of reference for languages: Learning, teaching, assessment.

[63] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 344–350.

[64] Samuel Reeder, Joshua Jensen, and Robert Ball. 2023. Evaluating Explainable AI (XAI) in Terms of User Gender and Educational Background. In *International Conference on Human-Computer Interaction*. Springer, 286–304.

[65] Carl O Retzlaff, Alessa Angerschmid, Anna Saranti, David Schneeberger, Richard Roettger, Heimo Mueller, and Andreas Holzinger. 2024. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research* 86 (2024), 101243.

[66] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[67] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26, 5 (2002), 521–562.

[68] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[69] Chris Russell. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 20–28.

[70] Scale AI. 2022. Demand Forecasting and Supply Matching to Optimize Continuity of Care. www.scaleai.ca/funded-projects/demand-forecasting-and-supply-matching-to-optimize-continuity-of-care

[71] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel Van Den Bosch. 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies* 154 (2021), 102684.

[72] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.

[73] Edward Small, Yueqing Xuan, Danula Hettiachchi, and Kacper Sokol. 2023. Helpful, Misleading or Confusing: How Humans Perceive Fundamental Building Blocks of Artificial Intelligence Explanations. In *ACM CHI 2023 Workshop on Human-Centered Explainable AI (HCXAI)*.

[74] Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 261.

[75] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 56–67.

[76] Kacper Sokol and Peter Flach. 2020. Interpretable representations in explainable AI: From theory to practice. *arXiv preprint arXiv:2008.07007* (2020).

[77] Kacper Sokol and Peter Flach. 2020. LIMEtree: Consistent and faithful surrogate explanations of multiple classes. *arXiv preprint arXiv:2005.01427* (2020).

[78] Kacper Sokol and Peter Flach. 2021. Explainability is in the mind of the beholder: Establishing the foundations of explainable artificial intelligence. *arXiv preprint arXiv:2112.14466* (2021).

[79] Kacper Sokol and Julia E Vogt. 2023. (Un)reasonable Allure of Ante-hoc Interpretability for High-stakes Domains: Transparency Is Necessary but Insufficient for Explainability. *Workshop on Interpretable ML in Healthcare at ICML* (2023).

[80] Kacper Sokol and Julia E. Vogt. 2024. What Does Evaluation of Explainable Artificial Intelligence Actually Tell Us? A Case for Compositional and Contextual Validation of XAI Building Blocks. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*.

[81] Gail M Sullivan and Anthony R Artino Jr. 2013. Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education* 5, 4 (2013), 541–542.

[82] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[83] Tjeerd van der Ploeg, Peter C Austin, and Ewout W Steyerberg. 2014. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* 14, 1 (2014), 1–13.

[84] Paul Voigt and Axel von dem Bussche. 2017. The EU general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.

[85] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[86] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

A SURVEY FLOW

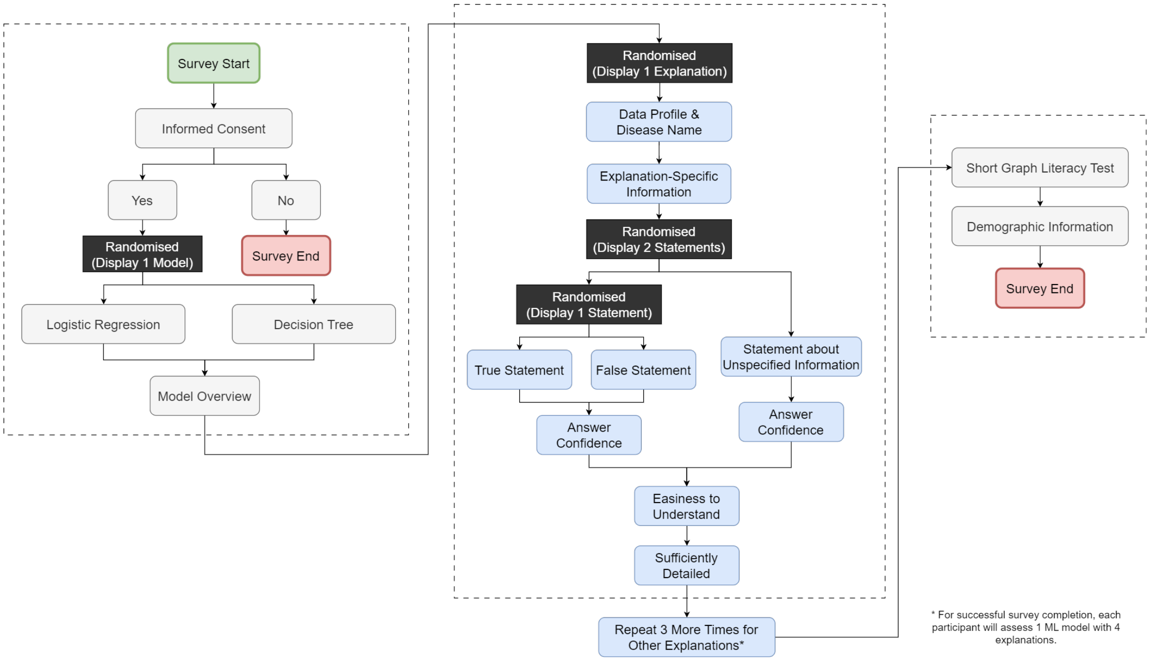
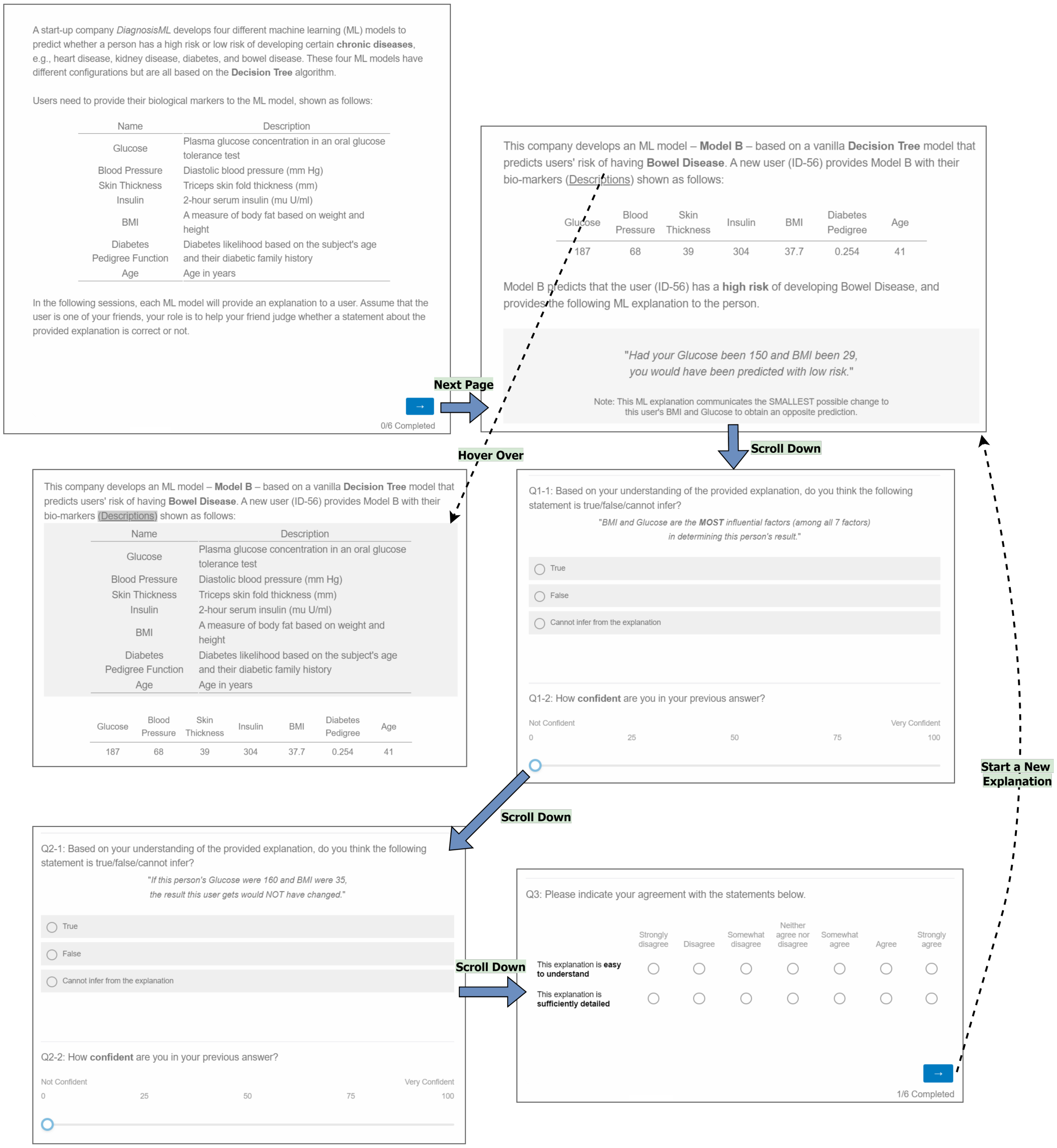

Survey Start
Informed Consent
Yes
No
Randomised (Display 1 Model)
Survey End
Logistic Regression
Decision Tree
Model Overview
Randomised (Display 1 Explanation)
Data Profile & Disease Name
Explanation-Specific Information
Randomised (Display 2 Statements)
Randomised (Display 1 Statement)
True Statement
False Statement
Statement about Unspecified Information
Answer Confidence
Answer Confidence
Easiness to Understand
Sufficiently Detailed
Repeat 3 More Times for Other Explanations*
Short Graph Literacy Test
Demographic Information
Survey End
* For successful survey completion, each participant will assess 1 ML model with 4 explanations.

Fig. 9. Diagram illustrating the flow of our survey.

B EXPLANATIONS AND COMPREHENSION STATEMENTS



A start-up company *DiagnosisML* develops four different machine learning (ML) models to predict whether a person has a high risk or low risk of developing certain **chronic diseases**, e.g., heart disease, kidney disease, diabetes, and bowel disease. These four ML models have different configurations but are all based on the **Decision Tree** algorithm.

Users need to provide their biological markers to the ML model, shown as follows:

Name	Description
Glucose	Plasma glucose concentration in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-hour serum insulin (mu U/ml)
BMI	A measure of body fat based on weight and height
Diabetes Pedigree Function	Diabetes likelihood based on the subject's age and their diabetic family history
Age	Age in years

In the following sessions, each ML model will provide an explanation to a user. Assume that the user is one of your friends, your role is to help your friend judge whether a statement about the provided explanation is correct or not.

→

0/6 Completed

This company develops an ML model – **Model B** – based on a vanilla **Decision Tree** model that predicts users' risk of having **Bowel Disease**. A new user (ID-56) provides Model B with their bio-markers (Descriptions) shown as follows:

Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree	Age
187	68	39	304	37.7	0.254	41

Model B predicts that the user (ID-56) has a **high risk** of developing Bowel Disease, and provides the following ML explanation to the person.

"Had your Glucose been 150 and BMI been 29, you would have been predicted with low risk."

Note: This ML explanation communicates the SMALLEST possible change to this user's BMI and Glucose to obtain an opposite prediction.

This company develops an ML model – **Model B** – based on a vanilla **Decision Tree** model that predicts users' risk of having **Bowel Disease**. A new user (ID-56) provides Model B with their bio-markers (Descriptions) shown as follows:

Name	Description
Glucose	Plasma glucose concentration in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-hour serum insulin (mu U/ml)
BMI	A measure of body fat based on weight and height
Diabetes Pedigree Function	Diabetes likelihood based on the subject's age and their diabetic family history
Age	Age in years

Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree	Age
187	68	39	304	37.7	0.254	41

Q1-1: Based on your understanding of the provided explanation, do you think the following statement is true/false/cannot infer?

*"BMI and Glucose are the **MOST** influential factors (among all 7 factors) in determining this person's result."*

○ True

○ False

○ Cannot infer from the explanation

Q1-2: How **confident** are you in your previous answer?

Not Confident — Very Confident

0 25 50 75 100

○

Q2-1: Based on your understanding of the provided explanation, do you think the following statement is true/false/cannot infer?

"If this person's Glucose were 160 and BMI were 35, the result this user gets would NOT have changed."

○ True

○ False

○ Cannot infer from the explanation

Q2-2: How **confident** are you in your previous answer?

Not Confident — Very Confident

0 25 50 75 100

○

Q3: Please indicate your agreement with the statements below.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
This explanation is **easy to understand**	○	○	○	○	○	○	○
This explanation is **sufficiently detailed**	○	○	○	○	○	○	○

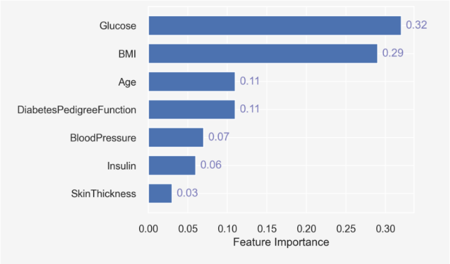
→

1/6 Completed

Fig. 10. Presentation of the task description and explanations in the survey. For brevity, we use only one explanation as an example.

Table 7. Explanations and comprehension assertions for the *logistic regression* model. The explanations are, from top to bottom, feature importance, decision surface visualisation, counterfactual explainability, and model architecture.

Explanation | Comprehension Assertion

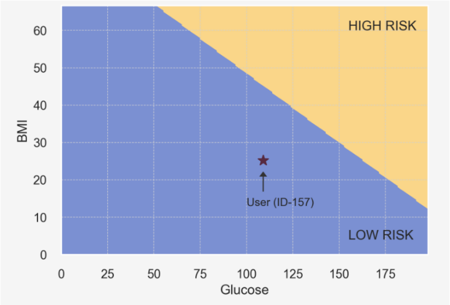


Note: This explanation shows how important each bio-marker is overall according to Model D. The higher the value, the more important the corresponding factor is.

True Model D uses all 7 biomarkers while making predictions, and Glucose is the most important factor for Model D.

False For Model D, Age is more important than DiabetesPedegreeFunction when predicting users' risk of diabetes.

Cannot Tell According to Model D, increasing one's BMI and Insulin would increase the predicted risk of having Diabetes.

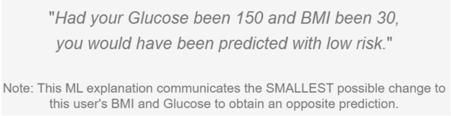


Note: This ML explanation shows the prediction that the user (ID-157) gets, and how it varies in relation to the two selected factors (while keeping other factor values unchanged).

True Assuming that all other biomarker values remain the same, increasing this person's Glucose to 150 and BMI to 40 will change their prediction from low risk to high risk.

False Assuming that all other biomarker values remain the same (including BMI), increasing this person's Glucose to 135 will change their prediction from low risk to high risk.

Cannot Tell BMI and Glucose are the MOST influential factors (among all 7 factors) in determining this person's result.



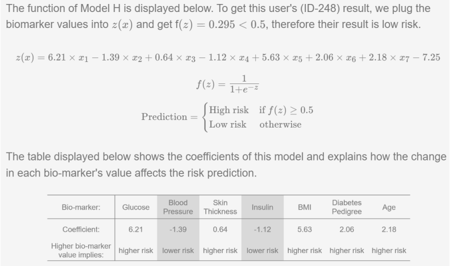
"Had your Glucose been 150 and BMI been 30, you would have been predicted with low risk."

Note: This ML explanation communicates the SMALLEST possible change to this user's BMI and Glucose to obtain an opposite prediction.

True If this person's Glucose were 160 and BMI were 35, the result this user gets would NOT have changed.

False If this person's BMI were 30 while all the other factors remained unchanged (including Glucose), the result this user gets would have changed to low risk.

Cannot Tell BMI and Glucose are the MOST influential factors (among all 7 factors) in determining this person's result.



The function of Model H is displayed below. To get this user's (ID-248) result, we plug the biomarker values into $z(x)$ and get f$(z) = 0.295 < 0.5$, therefore their result is low risk.

$$z(x) = 6.21 \times x_1 - 1.39 \times x_2 + 0.64 \times x_3 - 1.12 \times x_4 + 5.63 \times x_5 + 2.06 \times x_6 + 2.18 \times x_7 - 7.25$$

$$f(z) = \frac{1}{1+e^{-z}}$$

$$\text{Prediction} = \begin{cases} \text{High risk} & \text{if } f(z) \geq 0.5 \\ \text{Low risk} & \text{otherwise} \end{cases}$$

The table displayed below shows the coefficients of this model and explains how the change in each bio-marker's value affects the risk prediction.

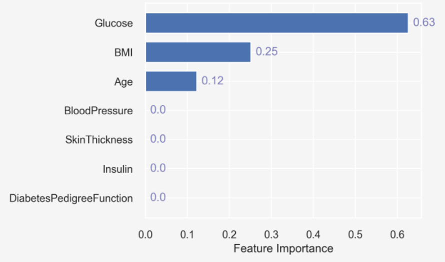
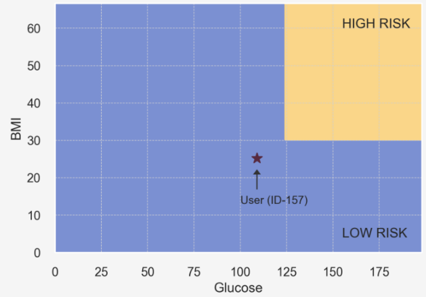
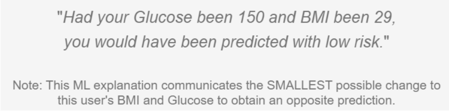
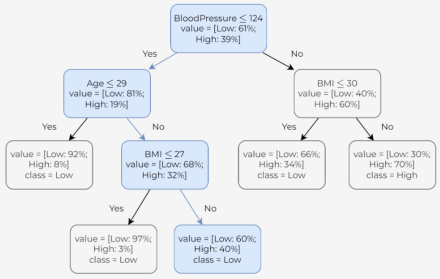
Bio-marker:	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree	Age
Coefficient:	6.21	-1.39	0.64	-1.12	5.63	2.06	2.18
Higher bio-marker value implies:	higher risk	lower risk	higher risk	lower risk	higher risk	higher risk	higher risk

True According to Model H, decreasing a person's Glucose and BMI will decrease their predicted risk.

False According to Model H, decreasing a person's Glucose and BMI will increase their predicted risk.

Cannot Tell According to Model H, increasing a person's Blood Pressure and BMI will decrease their predicted risk.

Table 8. Explanations and comprehension assertions for the *decision tree* model. The explanations are, from top to bottom, feature importance, decision surface visualisation, counterfactual explainability, and model architecture.

Explanation	Comprehension Assertion
 Note: This explanation shows how important each bio-marker is overall according to Model D. The higher the value, the more important the corresponding factor is.	**True** Model D only uses 3 features to predict users' risks of diabetes, and Insulin level does not affect the model's predictions. **False** The level of Insulin influences Model D's prediction for this user (ID-24) and all other users. **Cannot Tell** According to Model D, increasing one's BMI and Age will increase the predicted risk of having Diabetes.
 Note: This ML explanation shows the prediction that the user (ID-157) gets, and how it varies in relation to the two selected factors (while keeping other factor values unchanged).	**True** Assuming that all other biomarker values remain the same, increasing this person's Glucose to 135 and BMI to 36 will change their predicted result from low risk to high risk. **False** Assuming that all other biomarker values remain the same (including BMI), increasing this person's Glucose to 135 will change their result from low risk to high risk. **Cannot Tell** BMI and Glucose are the MOST influential factors (among all 7 factors) in determining this person's result.
 *"Had your Glucose been 150 and BMI been 29, you would have been predicted with low risk."* Note: This ML explanation communicates the SMALLEST possible change to this user's BMI and Glucose to obtain an opposite prediction.	**True** If this person's Glucose were 160 and BMI were 35, the result this user gets would NOT have changed. **False** If this person's BMI were 29 while all the other factors remained unchanged (including Glucose), the result this user gets would have changed to low risk. **Cannot Tell** BMI and Glucose are the MOST influential factors (among all 7 factors) in determining this person's result.
 BloodPressure ≤ 124 value = [Low: 61%; High: 39%] Yes No Age ≤ 29 value = [Low: 81%; High: 19%] BMI ≤ 30 value = [Low: 40%; High: 60%] Yes No Yes No value = [Low: 92%; High: 8%] class = Low BMI ≤ 27 value = [Low: 68%; High: 32%] value = [Low: 66%; High: 34%] class = Low value = [Low: 30%; High: 70%] class = High Yes No value = [Low: 97%; High: 3%] class = Low value = [Low: 60%; High: 40%] class = Low	**True** Assuming that all other biomarker values remain the same, increasing this person's Blood Pressure to 130 will change their result from low risk to high risk. **False** Assuming that all other biomarker values remain the same, increasing this person's BMI to 38 will change their result from low risk to high risk. **Cannot Tell** Blood Pressure has the most impact (i.e., is the most important) on the prediction this user (ID-248) gets.

C OUTCOMES OF ADDITIONAL STATISTICAL TESTS

Table 9. $\mathcal{X}^2$ and significance indicators (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$) of McNemar's test. The results show that feature importance and decision surface visualisation are more likely to lead to accurate comprehension of *explicated information* than counterfactual explainability and model architecture.

	Feature Importance	Decision Surface	Counterfactual	Model Architecture
Feature Importance	—	$18.28^{\star\star\star}$	$9.18^{\star\star\star}$	$14.41^{\star\star\star}$
Decision Surface	—	—	$44.47^{\star\star\star}$	$52.13^{\star\star\star}$

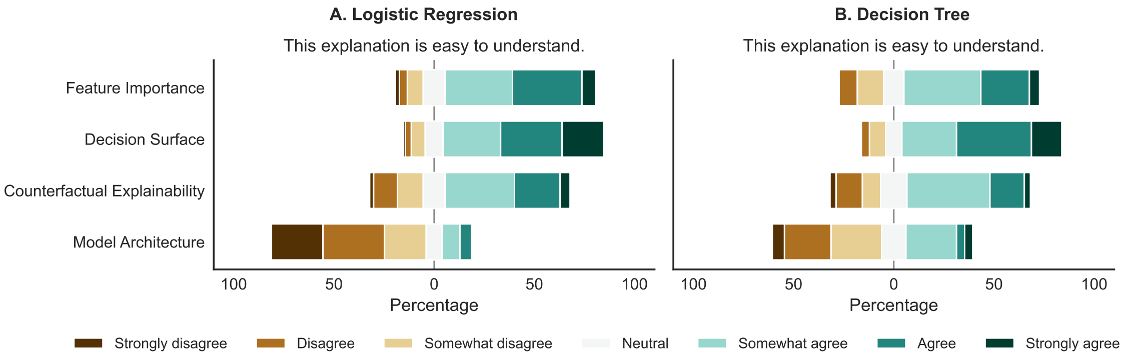


Fig. 11. Overview of Likert scale responses to "this explanation is easy to understand" stratified by explanation type for logistic regression in Panel (a) and decision tree in Panel (b).

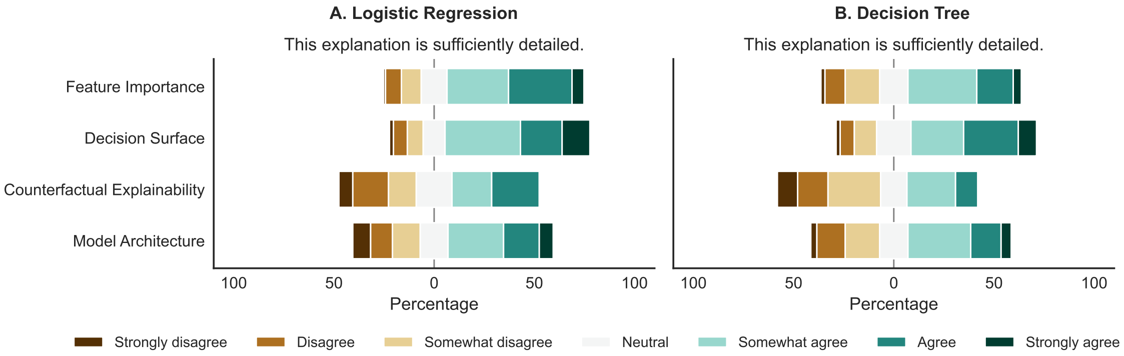


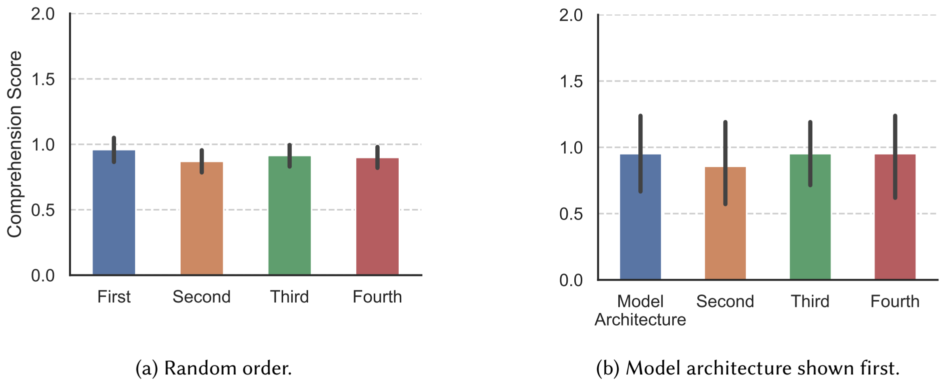
Fig. 12. Overview of Likert scale responses to "this explanation is sufficiently detailed" stratified by explanation type for logistic regression in Panel (a) and decision tree in Panel (b).

Table 10. Difference between the mean and significance indicator (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$) of Tukey's post-hoc paired test (the results are diagonally symmetrical). The outcomes show comparative differences between participants' responses to "this explanation is easy to understand" for different explanation types.

	Feature Importance	Decision Surface	Counterfactual	Model Architecture
Feature Importance	—	-0.475**	0.400*	1.785***
Decision Surface	—	—	0.875***	2.260***
Counterfactual	—	—	—	1.385***

Table 11. Difference between the mean and significance indicator (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$) of Tukey's post-hoc paired test (the results are diagonally symmetrical). The outcomes show comparative differences between participants' responses to "this explanation is sufficiently detailed" for different explanation types.

	Feature Importance	Decision Surface	Counterfactual	Model Architecture
Feature Importance	—	-0.235	0.830***	0.395*
Decision Surface	—	—	1.065***	0.630***
Counterfactual	—	—	—	-0.435*



(a) Random order.

(b) Model architecture shown first.

Fig. 13. Average number of questions answered correctly per explanation. We fully randomise the order of explanations in each survey to remove the impact of the priming effect. Panel (a) shows that the average performance for each explanation is comparable regardless of the order in which explanations were provided. Panel (b) demonstrates that for those users who were shown *model architecture* as their first explanation their performance on subsequent tasks did not improve consistently.

D ADDITIONAL ANALYSIS OF PRIMING EFFECT

As discussed in Section 4.2, we designed our user study in a way to eliminate the priming effect. In this appendix, we provide further evidence that the ordering effect did not play a role in the participants' performance as the result of our design choice. The performance results shown in Figure 13 suggest that participants did not gain an additional advantage when they were exposed to the *model architecture* explanation at the beginning, which could have been the case given that it is a global explanation that contains complete information about the model. More precisely, their performance in subsequent tasks remained stable (Figure 13b) and followed a similar pattern to the one found when the order of explanations was fully randomised (Figure 13a).