

# From Interpretability to Clinically Relevant Linguistic Explanations: The Case of Spinal Surgery Decision-Support

Alexander Berman<sup>1</sup><sup>a</sup>, Eleni Gregoromichelaki<sup>1</sup><sup>b</sup> and Catharina Parai<sup>2,3</sup><sup>c</sup>

<sup>1</sup>Centre for Linguistic Theory and Studies in Probability, Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden

<sup>2</sup>Department of Orthopedics, Sahlgrenska University Hospital, Gothenburg, Sweden

<sup>3</sup>Sahlgrenska Academy, University of Gothenburg, Sweden

{alexander.berman, eleni.gregoromichelaki}@gu.se, catharina.parai@vregion.se

**Keywords:** Interpretable AI, Machine Learning, Explanations, Linguistics, Argumentation Theory, Decision-Support Systems, Spinal Surgery.

**Abstract:** Interpretable models are advantageous when compared to black-box models in the sense that their predictions can be explained in ways that are faithful to the actual reasoning steps performed by the model. However, interpretability does not automatically make AI systems aligned with how explanations are typically communicated in human language. This paper explores the relationship between interpretability and linguistic explanation needs of human users for a particular class of interpretable AI, namely generalized linear models (GLMs). First, a linguistic corpus study of patient-doctor dialogues is performed, resulting in insights that can inform the design of clinically relevant explanations of model predictions. A method for generating natural-language explanations for GLM predictions in the context of spinal surgery decision-support is then proposed, informed by the results of the corpus analysis. Findings from evaluating the proposed approach through a design workshop with orthopaedic surgeons are also presented.

## 1 INTRODUCTION


In research concerning how to explain outputs from AI systems, two main paradigms have evolved. Post-hoc explanations methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) give some insight into how deep neural networks and other black-box models make their inferences. In contrast, predictions from interpretable models (or so-called “glass-box AI”) operate according to reasoning steps that are, at least in principle, comprehensible for humans (Rudin, 2019; Rudin et al., 2022).


It is sometimes argued that interpretable models are superior to black-box models in the sense that predictions can be explained in ways that are inherently faithful to the actual reasoning steps executed by the model, making interpretable models more adequate in high-stakes applications (Rudin, 2019). Nevertheless, research concerning interpretable models largely focuses on efforts to develop well-performing mod-


els (see, e.g. (Rudin et al., 2022)), leaving the relationship between interpretability and users’ explanation needs in AI-assisted decision-making largely unexplored. This gap in previous research concerns both how model interpretability can be leveraged to obtain linguistic explanations that meet users’ needs, and how such fulfilment of such needs depends on formal properties of interpretable models (sparsity, monotonicity, etc.).

This paper takes a step towards bridging this gap. Specifically, the paper focuses on a particular class of interpretable AI, namely generalized linear models (GLMs), in the context of spinal surgery decision-support. As its main contribution, the paper proposes a method for generating concise and clinically relevant linguistic explanations for predictions from GLMs, informed by communicative strategies observed in doctor-patient conversations.

The proposed method is applied in the context of a web-based instrument used by spine clinics in Sweden. The purpose of the instrument is to assist doctors and patients during medical consultations where decisions concerning choice of treatment (usually surgical or non-surgical treatment) are made. Based

<sup>a</sup>  <https://orcid.org/0000-0003-0513-4107>

<sup>b</sup>  <https://orcid.org/0000-0002-6933-5314>

<sup>c</sup>  <https://orcid.org/0000-0002-8332-0426>

on GLMs, the tool predicts two patient-reported outcomes of hypothetical surgery, as well as length of in-hospital stay, for patients with degenerative spinal disorders. The present study explores how the currently deployed instrument, which does not offer patient-specific explanations, can be modified and extended to meet doctors' and patients' clinical needs related to explainability.

The rest of the paper is organized as follows. Section 2 situates the work in relation to previous approaches to generating linguistic explanations for interpretable models. Section 3 is devoted to a linguistic corpus study, where explanations for medical judgements are collected from existing corpora and analysed in terms of communicative explanatory strategies. Implications of the analysis for the design of clinically relevant linguistic explanations are also discussed. In Section 4, a method for generating linguistic explanations of predictions from GLMs is proposed, informed by the findings from the corpus study. The section presents technical details concerning the proposed method, as well as a preliminary evaluation of the proposed method through a design workshop with orthopaedic surgeons. Finally, Section 5 offers conclusions and discusses future work.

## 2 RELATED WORK

The perhaps earliest example of natural-language explanations in the context of interpretable AI is SHRDLU (Winograd, 1971), a system which can explain its rule-based reasoning. For example, when the user asks why the system picked up a certain object, it may respond: "To get rid of it"; when asked why it got rid of it, it may respond: "To clean off the red cube", etc. In a similar vein, the more recent system DAISY (Wahde and Virgolin, 2023) can (exhaustively) explain how its use of hand-crafted or interactively learned procedures yields specific results. For example, when explaining how it concluded that the largest city in France is Paris, it states: "I retrieved all items in the city category", "Then I found all items belonging to France", etc. Methods for producing enthymematic (logically incomplete) explanations for answers inferred on the basis of facts and rules are proposed by (Xydis et al., 2020; Breitholtz, 2020; Maraev et al., 2021). For example, if the user asks why the system described by (Maraev et al., 2021) recommends a particular route, it responds: "Because the route is the shortest", thereby stating a fact whose relevance with respect to the explanandum hinges on the implicit premise that short routes are better than long ones. In contrast to these approaches, this paper

targets explanations for predictions from statistical models. Various such approaches have been proposed for black-box models, based on post-hoc explanation techniques (see, e.g. (Forrest et al., 2018; Kaczmarek-Majer et al., 2022; Slack et al., 2023)). One popular explanation strategy is to rank the most important features. For example, the system presented by (Slack et al., 2023) generates explanations on the form "For [this prediction], the importance of the features have the following ranking, where 1 is the most important feature: 1: glucose, 2: bmi, 3: age ..." Presumably, the lack of easily identifiable warrants makes such explanations difficult to understand, or cause a false sense of understanding when explainees identify warrants that do not reflect the actual inner workings of the model (Berman, 2024b).

Post-hoc explanation methods can also be used for interpretable models, by treating them as black boxes (see, e.g. (Ahmed et al., 2024)). As for approaches that instead leverage interpretability, (Baaj, 2022) shows how explanations can be generated for possibilistic and fuzzy rule-based systems. For example, a justification for the judgement that a patient's blood sugar level will not be low can be generated as: "This is mainly due to the fact that it is quite certain that the activity consists of drinking coffee, lunch or dinner and that the current blood sugar level is medium or high." A method for explaining predictions by decision trees and fuzzy rules is presented by (Alonso and Bugarín, 2019), enabling explanations such as "Beer is type Porter because its strength is standard and its color is brown".

The method presented in this paper builds on work by (Berman, 2024a), who proposes an interactive method for generating explanations for predictions by linear additive models, based on Toulmin's theory of argumentation (Toulmin, 2003). The method supports generation of both "data" (case-specific facts) and "warrants" (general statistical patterns). For example, if the model predicts that a person is introverted (based on her music preferences), the most important datum can be expressed as "The person likes high-energy music", with the corresponding warrant "Statistically, people that like high-energy music are more likely to be introverted." As detailed in Section 4, the present work extends this approach to handle a broader range of feature types (rather than only continuous features). Furthermore, the present paper shows how feature encoding can be jointly optimized for performance and linguistic intelligibility.

Almost none of the previous approaches have been empirically validated with end users. The only exception is (Slack et al., 2023) who let participants solve explainability-related tasks using two different

tools. However, the purpose of this validation was to compare the authors' conversational tool with a graphical interface; the extent to which the tasks or generated explanations were deemed clinically relevant by participants was not studied.

In contrast to previous approaches, the present work grounds the proposed natural-language generation method in a linguistic analysis of human explanatory interaction, and evaluates the method with end users in a clinically relevant scenario.

### 3 LINGUISTIC CORPUS STUDY

To inform the design of linguistic explanations, a qualitative linguistic corpus study is conducted by collecting and analysing examples of how doctors and patients explain (or support, or argue for/against) judgements (e.g. certain treatments) in clinical settings. Two empirical sources of clinical dialogues were chosen: the Norwegian Corpus of Doctor-Patient Consultations from Ahus (Gulbrandsen et al., 2013) (henceforth abbreviated Ahus), and a Swedish textbook in medicine focusing on the encounter between patient and doctor (Lindgren and Aspegren, 2004) (henceforth abbreviated L&A). The choice of empirical material is primarily motivated by the topics and types of situations that it encompasses. Furthermore, while one of the corpora (Ahus) is descriptive and contains transcripts of actual consultations, the other (L&A) is prescriptive and conveys communicative norms. Both types of linguistic data were deemed relevant for the purposes of the research.

The linguistic analysis builds on Toulmin's theory of argumentation (Toulmin, 2003). Toulmin identifies elements of arguments, including the *claim* (corresponding to the explanandum), *data* (specific facts that support the claim), and *warrants* (general norms or rules of thumb that justify *how* facts support claims). For example, the claim "you have a cold" can be supported (explained) by the datum "you have a runny nose", which in turn rests on warrants such as "runny nose is a symptom of a common cold".

Specifically, the corpus study aims to answer the following research questions:

1. To what extent do interlocutors (doctors and patients) *explicitly* convey argumentative elements (claims, data, and warrants)?
2. To what extent do interlocutors *implicitly* convey argumentative elements?
3. How do interlocutors linguistically indicate the relationship between argumentative elements?

The RQs were purposely formulated in relation to the intended downstream application of the results, i.e., design of linguistic explanations of predictive models. Specifically, claims are conceived to be potentially analogous with statistical predictions, data with feature values, and warrants with statistical patterns learned by predictive models. In other words, it is assumed in principle conceivable that doctors reason in ways that are analogous (to some extent) with how machine-learning models make predictions.

#### 3.1 Corpus Data

Occurrences of explanations (or related phenomena such as arguments or justifications) pertaining to medical judgements were identified using a search procedure. In the case of Ahus, which contains transcriptions of 220 consultations, this was done by searching for the word "why" ("hvorfors" in Norwegian); for L&A, the corpus was small enough to permit a manual search of the entire empirical material.

The topic of interest (medical judgements) primarily encompasses diagnosis (judging that a patient has a certain condition) and recommendations (judging that a particular action or intervention is adequate). The selection procedure resulted in two dialogues from each corpus, spanning a total of 88 utterances.

#### 3.2 Annotation

To address the RQs, the data was annotated with the following labels (hypothetical examples in parentheses):

- claim ("I think that *you have a cold*")
- datum ("since *you have a runny nose*")
- warrant ("since *runny nose is a symptom of a common cold*")

Note that arguments are not assumed *a priori* to be marked with particular syntactic constructions or particles such as "since", "because", or "therefore" (Sbisà, 1987).

A complete annotation was first done by one of the authors (with expertise in cognitive science, linguistics, and machine learning), and then reviewed by the two other authors (with expertise in linguistics and medicine respectively). During the reviews, annotations were open for collaborative amendments.

#### 3.3 Analysis

The analysis reveals that in the empirical material, claims are explicitly supported by up to three pieces

of data, whereas warrants are rarely communicated explicitly. For example, in one of the dialogues, the doctor expresses an intent to have the patient's lungs x-rayed (claim), which is justified with reference to the patient's low levels of oxygen in the blood (datum). There is no explicit mention of how a person's levels of oxygen explain the recommendation to perform lung x-ray (warrant).

In cases where warrants are made explicit, they are sometimes causal in nature. For example, in one dialogue, the patient expresses a wish to have her heart checked-up since she gets very dizzy and wonders if this is due to low blood pressure. In response, the doctor explains that the patient's dizziness can be caused by her diabetes. However, in many cases, arguments seem compatible with either causal or statistical warrants. For example, when a doctor judges that the patient has no respiratory illness partly on the basis that the "chest X-ray was completely normal"<sup>1</sup>, this is compatible with either a causal warrant (e.g. respiratory illnesses cause abnormalities that can be detected in a chest X-ray) or a statistical one (e.g. a normal chest X-ray correlates with absence of respiratory illness).

Although warrants are rarely verbalized, data are very frequently conveyed in ways that indicate what *kind* of warrant the speaker might have in mind. In one excerpt, the doctor says that the patient's oxygen levels are "a bit lower than one would expect". Although no warrant is explicitly conveyed, the words "lower" and "expect" both allude at a warrant such as "unexpectedly low levels of oxygen in the blood can indicate lung disease". In another example, the doctor explains a recommended change of medication as follows: "since you have had [the medication for] over two months and have increased the dosage and not had any effect"; here, the lexical choices "over", "increased", and "not ... any" trigger a warrant such as "having used a medication for a long time without any effects motivates trying another medication".

Generally, two types of warrant triggers can be observed in the data: scalar/gradable and norm/expectation related. Examples of scalar triggers include *lower* (levels of oxygen in the blood than one would expect), *over* (two months of medication use), *increased* (dosage), and *no* (effect of medication). Examples of norm-related warrant triggers include *expected* (levels of oxygen in the blood), *normal* (lung x-ray), *abnormal* (nothing abnormal in patient's lungs), *should* (nothing observed that shouldn't be there), and *good* (cholesterol levels).

Although linguistic triggers help explainees to

<sup>1</sup>Cited excerpts from the empirical material have been translated to English by the authors.

identify potentially relevant warrants, a certain amount of argumentative underspecification (ambiguity) can be observed. For example, conveying oxygen level in the blood as lower than expected is compatible with a warrant that posits a *monotonically decreasing* relation between oxygen level in the blood and the risk of lung disease. However, it is also compatible with a *non-monotonic* relationship, i.e. that too *high* oxygen levels in the blood also indicate a higher risk of disease. Similarly, when multiple pieces of data are presented in support of a claim (such as having used a medication for a long time with a high dosage), potential interactions between data remain unstated. This kind of underspecification can potentially be understood as serving the purposes of relevance and brevity, i.e. only presenting information that is deemed relevant for the patient, and not providing more information than needed in the context (cf. Grice's maxims of relation/relevance and quantity (Grice, 1975)).

As an additional finding, we observe that interlocutors sometimes discuss mutually opposing claims. For example, in one dialogue, the patient expresses a wish to have her heart checked-up, while the doctor argues that the patient's dizziness can be caused by her diabetes and that her blood pressure is fine, thereby constructing a counter-claim that a heart check-up is not needed.

### 3.4 Implications for Linguistic Design of Model Explanations

When using the results of the corpus study to inform the design of clinically relevant explanations for model predictions, several guiding principles can be distilled. First, given the limited amount of conveyed data per claim in human-human dialogues, it is advisable to focus specific (local) explanations only on those features that are most important to the predicted outcome.

Second, given the consistent use of warrant triggers in the presentation of data, and the explanatory function that these triggers can be assumed to carry, it is advisable to formulate data in ways that allude to the corresponding statistical patterns learned by the model. For scalar triggers, this can be done by choosing a suitable modifier. For example, if the model has learned that older age is associated with a lower probability of being satisfied with surgery, and the patient's age is statistically low, the patient's age can be presented as "relatively young". Norm/expectation triggers can also be conceived, e.g. "*unexpectedly* high pain levels in the arm". However, in the present clinical context, this was not deemed applicable.

Third, given that interlocutors sometimes exchange claims and counter-claims, and give reasons both for and against judgements, it is advisable to convey circumstances that both support and contradict a certain prediction (cf. (Miller, 2023)).

Furthermore, to provide more detailed information about the statistical patterns learned by the model and thereby help resolve potential ambiguities, explicit warrants could be offered. However, given that, in the studied corpora, warrants were primarily conveyed implicitly, warrants should only be presented on demand.

Finally, it is worth noting that the observed usage of causal warrants in human-human dialogues may indicate that statistical explanations may not always be satisfactory or sufficient for users (as previously argued by e.g. (Miller, 2019)). For example, a correlation between low disability and high probability of successful surgery might be difficult to comprehend without resorting to causal reasoning of some kind (such as disability causing depression or other psychological states that in turn influence pain perception). In principle, generated explanations for model predictions could include such causal links, potentially collected from domain experts and built into the tool. However, such explanations could invite false inferences concerning how the model actually reasons. For this reason, we argue that generated warrants should only convey actual statistical patterns learned by the model, leaving causal matters open for interpretation.

## 4 GENERATING CLINICALLY RELEVANT EXPLANATIONS

Swedish spine clinics have access to an AI-based “Dialogue Support” tool whose purpose is to assist patients and doctors in their decision-making concerning treatment options for four different types of degenerative spinal disorders: disc herniation and spinal stenosis in the lumbar spine respectively, or chronic low back pain, as well as cervical radiculopathy (Fritzell et al., 2022). The tool is used by the doctor and patient together during brief (approx. 20 minutes) medical consultations. Based on sociodemographic information (age, gender, etc.) and other answers provided by the patient in a questionnaire, the tool presents predictions for three types of outcomes of a hypothetical surgery:

- **Satisfaction:** the probability of responding to the question “What is your attitude towards the result of the surgery?” with one of the first two options

in the Likert scale *satisfied, hesitant, dissatisfied* one year after surgery.

- **Global Assessment (GA):** the probability of responding to the question “How is your pain today as compared to before the surgery?” with one of the first two options in the Likert scale *completely pain-free, much better, somewhat better, unchanged, worse* one year after surgery.
- **Length of stay:** the number of days of hospitalization in connection with the surgery.

The predictions are based on three different types of GLMs trained on patient data from the national quality registry Swespine. Currently, the Dialogue Support tool explains its predictions “globally”, with information about features, sample size, etc. No case-specific (“local”) explanations are presented.

In the subsequent sections, a method for generating linguistic explanations for predictions from GLMs will be proposed, with the purpose of enabling an alternative Dialogue Support tool based on the same patient data and types of models. The proposed method builds on previous work by (Berman, 2024a), which is extended to support needs and desiderata identified in the corpus study, as well as feedback collected in a design workshop with orthopaedic surgeons (see Section 4.7).<sup>2</sup>

### 4.1 Model Specification

GLMs estimate an outcome on the basis of a linear combination of predictors (independent variables) and a link function that transforms the linear combination to an outcome:

$$\mathbb{E}[Y | X] = g^{-1} \left( \beta_0 + \sum_i X_i \beta_i \right)$$

where  $\mathbb{E}[Y | X]$  is the expected value of the outcome  $Y$  given the intercept (bias) term  $\beta_0$ , the predictors  $X_i$ , the coefficients  $\beta_i$ , and the link function  $g$ . For the purposes at hand,  $g$  is assumed to be monotonic, which is typically the case. Specifically, in the study at hand,  $g$  is either

- *logit* (i.e. logistic regression), for estimating satisfaction,
- *threshold function for the cumulative distribution function* (i.e. ordered probit), for estimating GA, or
- *rounding to non-negative integer* (linear (Ridge) regression adapted to counts), for predicting length of stay.

<sup>2</sup>While the prototype currently supports Swedish, this paper uses English translations.

One model is used for each combination of diagnosis and task (type of outcome). Since there are 4 diagnoses and 3 tasks, a total of 12 models is used.

## 4.2 Datasets

Historical patient data was obtained from the Swe-spine registry in the form of one dataset per diagnosis. The 4 datasets together encompass 37 features. Table 1 presents the feature types, with examples of features.

For the purposes of the study, features were selected to jointly optimize for performance and sparsity. This was done using backward elimination, with area under the ROC curve (AUC) as performance metric for satisfaction and GA, and mean absolute error (MAE) for length of stay. Among the best-performing feature sets (i.e. performance not worse than the best performance minus a tolerance threshold), the feature set with the smallest number of features was selected.

## 4.3 Feature Encoding

Different feature types/encodings afford distinct linguistic expressions of data and warrants. Numeric features enable the use of scalar triggers such as “*young* age” or “*no* pain in the back”. For example, if the patient’s young age is presented as a positive factor, this invites the inference that the model generally associates a lower age with a more favourable outcome.

Binary features enable self-evident warrant-triggering contrasts. For example, if the patient being female is presented as a positive factor, this implies that female patients are generally estimated to have a more favourable outcome than male patients.

In the case of multinomial categorical features, warrants are not as straightforwardly triggered. For example, if the fact that the patient is “treated at a university clinic” (rather than a public or private clinic) is presented as indicative of a negative outcome, a corresponding warrant cannot easily be identified. Specifically, the formulation does not convey whether the prediction would be more positive if treated in a public or private clinic. Since not many multinomial categoricals were among the most predictive features in the studied case, this problem was not addressed.

As for ordinals, the picture is more nuanced. At least two approaches can be conceived: to treat them either as numeric or as multinomial categorical features. In line with the reasoning above, numeric encoding is more favourable from the perspective of linguistic intelligibility. However, a categorical en-

coding may yield better performance. To this end, the choice of encoding for ordinals was made in a flexible way. Specifically, high-cardinality ordinals (> 5 values) were encoded numerically, since it was deemed highly unintuitive to treat each value on a 10-item pain scale as its own category. As for low-cardinality ordinals, a data-driven approach was employed to jointly optimize for linguistic intelligibility and performance. Among the best-performing encoding candidates (i.e. performance not worse than the best performance minus a tolerance threshold), the feature set with the largest amount of numeric encodings was selected.<sup>3</sup>

The proposed strategy for choosing feature encodings is summarized in Table 1.

## 4.4 Interface Design

Local explanations are presented in a waterfall chart, where the estimated outcome is visualized in terms of the outcome for an average patient, and the cumulative effect of data (see Figure 1), grouped into positive and negative, and ordered by decreasing importance. For example, a moderate probability of a successful outcome can be explained by the fact that the patient has no other illnesses (positive factor) and relatively severe back pain (negative factor). A maximum of three positive and negative factors respectively is shown by default; additional data can be revealed by clicking “Show more”.

Outcomes are visualized along a probability scale (for satisfaction and GA) or integer scale (for length of stay), while data (feature contributions) are visualized without any explicit scale. In other words, mathematically, the waterfall chart informally conveys:<sup>4</sup>

$$\mathbb{E}[Y | \bar{X}] + \sum \beta_i (X_i - \bar{X}_i) \approx \mathbb{E}[Y | X]$$

Warrants conveying information about the statistical patterns learned by the model, e.g. that lower disability is associated with a higher estimated probability of a successful outcome, are presented inside a widget titled “More information”. The widget also contains information about the sample size and training data, which argumentatively can be said to back the warrants. The content of the widget is collapsed by default, but can be expanded as needed.

<sup>3</sup>Predictive performance after interpretability optimizations was AUC 0.65-0.69 for satisfaction, AUC 0.62-0.69 for GA, and mean absolute error 0.27-1.06 for length of stay.

<sup>4</sup>A completely faithful visualization would need to account for non-linearity of the link function. This degree of faithfulness was not deemed motivated for the purposes at hand.

Table 1: Features types among the datasets used to train the predictive models. The reasoning behind the choice of encodings is described in Section 4.3.

Feature type	Encoding	Example(s)
numeric	standardization (continuous value)	age, BMI
binary	binary	gender, employment status
multinomial categorical	one-hot (dummy variables)	clinic type (public, private, or university hospital)
high-cardinality ordinal	standardization (continuous value)	pain levels (10-item scale from 0 (no pain) to 10 (worst conceivable pain))
low-cardinality ordinal	one-hot (dummy variables) or standardization (continuous value)	walking distance (5-item scale from 0-100 meters to more than 2 years)

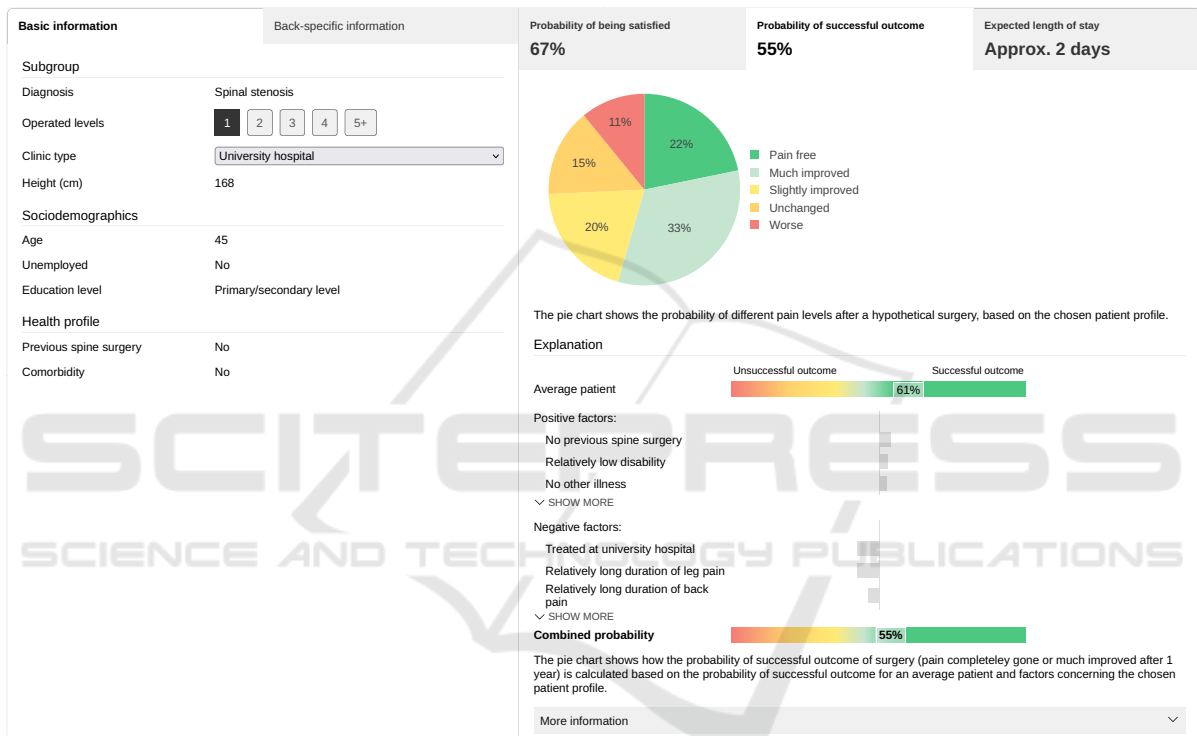


Figure 1: Screenshot of the prototype, with a hypothetical patient profile.

#### 4.5 Choice of Reference

As elaborated above, data are visually and linguistically presented in relation to a reference. In the proposed prototype, the reference is chosen as the mean historical patient with the same diagnosis ( $R = \bar{X}$ ), although such a notion may be perceived as abstract. The main reason for not settling with the intercept/bias as reference is that this would introduce an undesired bias for binary features; e.g. the gender encoded as 0 (in this case being male) would never be highlighted as a factor.

The frequently observed comparison with expectation/norm in the corpus (see Section 3.3) may suggest using “healthy patient” as a reference in medical contexts where it can be clearly established *a priori*

what constitutes a “healthy” feature value. Allowing reference to be chosen interactively may also be an option.

#### 4.6 Generation of Data and Warrants

The proposed method for generating linguistic explanations for GLMs consists of general (domain- and language-independent) functions for generating data and warrants, which depend on a domain- and language-specific grammar containing functions that produce linguistic surface realizations. The overall architecture of the proposed method is illustrated in Figure 2.

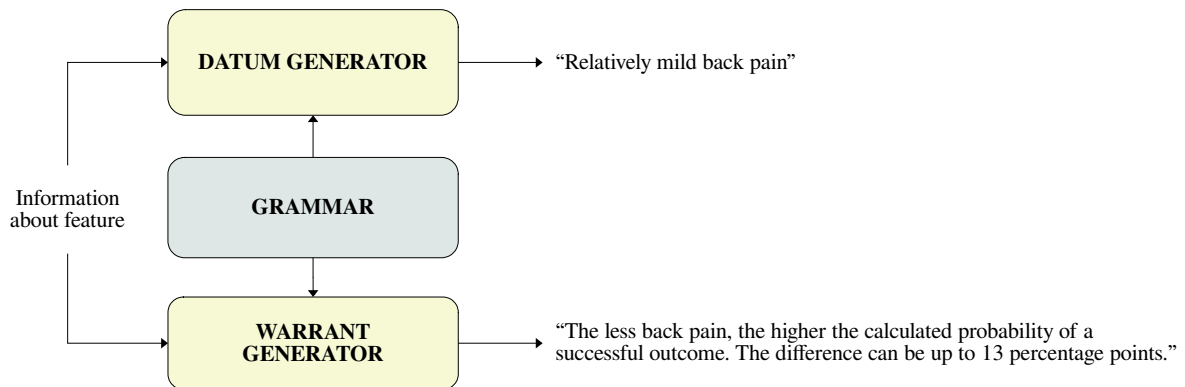


Figure 2: Overall architecture of the proposed method for generating linguistic explanations for GLMs, with output examples for the feature BackPain. Input to the datum and warrant generators consists of information such as feature type, feature value, mean feature value, and coefficient.

#### 4.6.1 Generation of Data

In the visualization of local explanations, data are expressed either with scalar triggers (e.g. “Relatively young/old age”) or a bare linguistic label (e.g. “Female”), depending on the type of feature. Formally, we define  $\text{DatumPhrase}(t, f, x, \bar{x})$  as a general function that, given information about a feature, returns a datum phrase. Specifically,

$$\text{DatumPhrase}(t, f, x, \bar{x}) = \begin{cases} \text{FeatureLevelPhrase}(f, \text{Level}(x, \bar{x})) & \text{if } t = \text{numeric,} \\ \text{Label}(f, x) & \text{if } t = \text{categor.} \end{cases}$$

where  $f$  is the feature symbol,  $x = X_i$  is the feature value, and  $\bar{x}$  is the mean value of the feature for the entire sample.  $\text{FeatureLevelPhrase}(f, l)$  is a grammar function that generates a feature-level phrase for feature  $f$  and level  $l$  (e.g. “Relatively young age” for  $f = \text{Age}$  and  $l = \text{low}$ ), while the grammar function  $\text{Label}(f, x)$  returns a bare linguistic label (e.g. “Female” for  $f = \text{Gender}$  and  $x = 1$ ). The general function  $\text{Level}(x, r)$  returns the level of the feature value  $x$  in relation to a reference value  $r = R_i$ . Specifically,

$$L(x, r) = \begin{cases} \text{zero} & \text{if } x = 0, \\ \text{low} & \text{if } 0 < x < r, \\ \text{high} & x \geq r \end{cases}$$

For the grammar functions  $\text{FeatureLevelPhrase}$  and  $\text{Label}$ , the present work uses simple mappings (see Figure 3).

Results of applying the proposed method on historical patients (i.e. instances in the datasets) are shown in Table 2 and Table 3, presenting the 10 most frequently occurring phrases for positive and negative factors respectively.

Table 2: The 10 most frequently occurring phrases describing positive factors for historical patients, as generated by the proposed method and aggregated across diagnoses and outcomes. “Occurring” here means that the phrase is included among the top three positive factors shown in the interface. “Positive” refers to higher estimated probability of satisfaction or successful outcome, or shorter duration of stay. Frequencies are relative to the total number of patients.

Data presented as positive	Freq.
Treated in private clinic	0.46
No other illnesses	0.35
No previous spine surgery	0.32
Relatively low disability	0.18
Relatively few operated levels	0.17
Relatively short duration of back pain	0.15
Has university education	0.13
Relatively mild back pain	0.13
Can walk relatively far in normal pace	0.12
No university education	0.12

Inferences that are invited by presenting data in the manner described above are mathematically guaranteed to correctly reflect the actual reasoning steps performed by the model. For example, if “relatively mild back pain” (compared to an average patient) is presented as a positive factor, this invites the inferences that the patient has some amount of back pain, but lower than an average patient, and that mild back pain is associated with a more favourable predicted outcome than severe back pain. Mathematically, this corresponds to conveying  $X_i < \bar{X}_i$  and that  $\beta_i(X_i - \bar{X}_i) > 0$  contributes positively to the outcome (assuming a monotonically increasing inverse link function), and linguistically inviting the factual inference  $0 < X_i < \bar{X}_i$  and the warrant inference  $\beta_i < 0$ . The correctness of the factual inference follows from the condition for selecting the word “mild” ( $L(X_i, \bar{X}_i) = \text{low}$  iff  $0 < X_i < \bar{X}_i$ ). As for the warrant inference,



$$\text{FeatureLevelPhrase}(f, l) = \begin{cases} \text{if } f = \text{BackPain} : & \begin{cases} \text{“No back pain”} & \text{if } l = \text{zero,} \\ \text{“Relatively mild back pain”} & \text{if } l = \text{low,} \\ \text{“Relatively severe back pain”} & \text{if } l = \text{high} \end{cases} \\ \text{if } f = \dots & \end{cases}$$

$$\text{Label}(f, x) = \begin{cases} \text{if } f = \text{Gender} : & \begin{cases} \text{“Male”} & \text{if } x = 0, \\ \text{“Female”} & \text{if } x = 1 \end{cases} \\ \text{if } l = \dots & \end{cases}$$

Figure 3: Examples of domain- and language-specific linguistic mappings belonging to an English grammar in the domain of spinal disorders.

Table 3: The 10 most frequently occurring phrases describing negative factors for historical patients, as generated by the proposed method and aggregated across diagnoses and outcomes. “Occurring” here means that the phrase is included among the top three negative factors shown in the interface. “Negative” refers to lower estimated probability of satisfaction or successful outcome, or longer duration of stay. Frequencies are relative to the total number of patients.

Data presented as negative	Freq.
Relatively high disability	0.19
Treated in public clinic	0.18
Relatively severe back pain	0.18
Has other illnesses	0.16
No university education	0.16
Relatively long duration of back pain	0.15
Relatively short height	0.13
Relatively many operated levels	0.13
Relatively old age	0.13
Relatively long duration of leg pain	0.10

its correctness can be verified as follows. Since  $\beta_i(X_i - \bar{X}_i) > 0$  and  $X_i - \bar{X}_i < 0$  (because  $X_i < \bar{X}_i$ ), the product  $\beta_i(X_i - \bar{X}_i)$  can only be positive if  $\beta_i < 0$ . In other words, the mathematical guarantee hinges on the monotonicity of the link function and on the linear additive treatment of features.

#### 4.6.2 Generation of Warrants

Warrants conveying correlations learned by the models are expressed in ways that communicate both effect size and, when relevant, polarity. For example, the warrant for a numeric feature can be formulated as: “The less back pain, the higher the calculated probability of a successful outcome. The difference can be up to 24 percentage points.” Effect sizes are calculated with respect to the magnitude of the coefficient and the absolute maximum slope of the inverse link function.<sup>5</sup>

Table 4 shows generated warrants for a particular

<sup>5</sup> $|\max((g^{-1})'(x))|$  is  $\frac{1}{4}$  for logistic regression,  $\frac{1}{\sqrt{2\pi}}$  for ordered probit, and 1 for linear regression.

diagnosis and outcome.

## 4.7 Evaluation

An early prototype of the proposed solution was evaluated through a design workshop with orthopaedic surgeons. In the invitation, potential participants were informed that they would test a new alternative interface to the Dialogue Support tool, which they all had experience of using. 3 out of 5 invited respondents participated in the workshop on site, while one tested the prototype individually and then gave written feedback via email.<sup>6</sup>

In the first part of the workshop, participants individually accessed the prototype, where a randomized patient profile was shown.<sup>7</sup> They were then asked to imagine having a dialogue with a patient who wants to know why the computer estimates an X% probability of successful outcome, and to answer the patient’s question very briefly. Participants noted their answers and were then asked to voluntarily read them aloud.

In the second part, participants were asked: “Is there anything in the explanations that can be improved?”<sup>8</sup> The discussion was moderated by the organizer of the workshop (one of the authors).

The version of the prototype tested by participants only supported one diagnosis (disc herniation) and two outcomes (satisfaction and GA). Instructions were conveyed to participants verbally and via a beamer presentation. Feedback was recorded in notes and later organized according to themes.

<sup>6</sup>One of the authors participated in the workshop in the role of orthopaedic surgeon. This author had not been involved in the development of the prototype or the organization of the workshop.

<sup>7</sup>The randomization was done individually for each participant. Feature values were uniformly sampled from predefined ranges.

<sup>8</sup>Participants were also asked if there is anything else with the alternative interface that can be improved. Results from this part of the workshop are not reported here.

Table 4: Generated warrants conveying correlations for spinal stenosis and pain assessment (GA). Warrants are shown to users when they click “More information” in the interface.

<b>Disability</b>	The lower the disability, the higher the calculated probability of a successful outcome. The difference can be up to <b>33</b> percentage points.
<b>Leg pain duration</b>	The shorter the leg pain duration, the higher the calculated probability of a successful outcome. The difference can be up to <b>19</b> percentage points.
<b>Previous spine surgery</b>	Patients who have not undergone previous spine surgery are calculated to have a higher probability of a successful outcome. The difference can be up to <b>14</b> percentage points.
<b>Back pain</b>	The less back pain, the higher the calculated probability of a successful outcome. The difference can be up to <b>13</b> percentage points.
<b>Back pain duration</b>	The shorter the back pain duration, the higher the calculated probability of a successful outcome. The difference can be up to <b>11</b> percentage points.
<b>Comorbidity</b>	Patients with no other illnesses are calculated to have a higher probability of a successful outcome. The difference can be up to <b>8</b> percentage points.
<b>Unemployment</b>	Patients with employment are calculated to have a higher probability of a successful outcome. The difference can be up to <b>8</b> percentage points.
<b>Type of clinic</b>	Patients treated at private clinics are calculated to have the highest probability of a successful outcome. Patients treated at university hospitals are calculated to have the lowest probability of a successful outcome. The difference can be up to <b>7</b> percentage points.

#### 4.7.1 Results

In the first part, one of the participants gave the following answer: *“There is a 76% probability of being satisfied with surgery which is 10% worse than an average patient who is operated for disc herniation. The reason that it looks this way is that you have had back pain for a longer time and furthermore xxx diseases. Furthermore, your age gives you a statistically somewhat lower probability of a successful result.”* (our translation)

As for the second part, one participant commented that the textual explanations for the model and the specific outcomes were good and informative. Three suggestions related to global explanations were raised. First, it was recommended to use positive instead of negative wording; e.g. substituting expressions like *“The older the age, the lower the...”* with *“The younger the age, the higher the...”*. Second, the wording *“is estimated”* was advised to be replaced with alternatives such as *“is calculated,”* *“results have previously shown,”* or *“based on previous patients’ results of surgery”*. Third, it was suggested that features should be sorted by descending effect size. All these suggestions were subsequently accommodated.

A question was raised as to whether the correlation between age and satisfaction is in reality *“a curve rather than linear”*. The question highlights a discrepancy between the participant’s domain knowledge and monotonicity assumptions built into the model. The fact that this discrepancy surfaced can be seen as a positive finding, in the sense that the tool’s explanation enabled the user to form a correct mental model of how the AI reasons (and to contrast this model with

his/her own reasoning).

Two participants wondered if it would be adequate to tell the patient that the model makes its prediction based on data concerning other patients with similar characteristics. In Toulmin’s framework, this relates to *backing*, i.e. arguing for the general acceptability of a warrant. To enable more accurate backing, information was added in the tool to clarify that predictions are made on the basis of the entire available sample of patients with the diagnosis at hand.

One participant observed that some factors were missing from the global explanations; this was later attributed to a bug, which has since been resolved.

As an indirect finding, it was observed that no questions or critical comments were raised regarding the phrasing of data. Furthermore, when participants exemplified how they would respond to the patient’s request for an explanation, the participants seemed to have interpreted the data as intended. This can be taken as an indication that the generation of datum phrases is serving its intended purpose.

In summary, the design workshop resulted in generally positive findings regarding the proposed linguistic explanations, as well as suggested improvements that have been accommodated into the prototype.

## 5 CONCLUSIONS AND FUTURE WORK

Informed by a linguistic and argumentative analysis of doctor-patient dialogues, this paper has proposed

a method for generating linguistic explanations for predictions from GLMs in the context of a statistical instrument aiming to assist treatment decisions concerning degenerative spinal disorders. Unlike previous approaches to generating natural-language explanations for predictions from statistical models, the method proposed here is grounded in empirical findings concerning analogous human communicative strategies. Specifically, the proposed method linguistically formulates salient case-specific facts in ways that concisely indicate the underlying statistical patterns used by the model, without overloading the user with an exhaustive account of the model's entire reasoning logic. This way, the approach balances purposeful brevity with the possibility to get more detailed information that reduces potential ambiguity regarding the model's reasoning.

The simplicity of the concrete linguistic explanations obtained with the proposed method ("relatively young age", etc.) reflects a desirable alignment with how humans typically formulate explanations. Notably, this simplicity is not obtained at the cost of reduced faithfulness, unlike with popular post-hoc explanation methods for black-box models (such as LIME and SHAP).

While the interface design and linguistic surface realization presented in this paper are adapted to a specific medical use case, the method as such is simple in nature and straightforwardly generalizes to other clinical use cases with similar needs, and potentially also to other domains. Despite recent developments in deep neural networks and generative AI, GLMs are still commonly used in high-stakes domains such as healthcare, in part due to their interpretability (Pantanowitz et al., 2024). From this perspective, one of the main implications of the presented work is to demonstrate how interpretability, conceived as pertaining to certain abstract or formal properties, can be leveraged in practice to obtain downstream value in the form of linguistic explanations that are aligned with how humans typically explain decisions and judgements. This can be potentially valuable not only for AI-based decision support, but also in situations where linguistically intelligible explanations for AI-based decisions are required for ethical or legal reasons (see, e.g. (Berman, 2024b)). On a theoretical level, the work also shows how practical applications of interpretable AI depend on specific formal properties of interpretable models – in this case monotonicity and linear additivity.

The proposed method for generating linguistic explanations has been tested and refined through a design workshop with orthopaedic surgeons. Overall, the findings from the workshop yielded generally pos-

itive feedback. Nevertheless, the extent to which the explanations meet needs of doctors and patients in real-world clinical settings has not been studied. In a planned next step, the approach will therefore be tested in a clinical study, where interactions between patient, doctor, and predictive tool will be recorded and analysed. Doctors' and patients' experiences of using the tool will also be investigated through interviews and questionnaires.

In future work, it might also be useful to conduct a more fine-grained argumentation analysis with all Toulmin's (Toulmin, 2003) argumentative elements (e.g. *backing* and *qualifier*). Moreover, in the corpus analysis it was observed that claims and their accompanying data/warrants might be complex in that they are structured with one argument element embedded in another, having a coordinative arrangement or other potential arrangements (see, e.g. (Eemeren et al., 2021)). More elaborate annotation will be required to bring out these forms of structuring and determine their potential relevance in relation to model interpretability and AI explanations.

## 6 ETHICS STATEMENT

Approval from the Swedish Ethical Review Authority was obtained (case number 2024-00839-01) for handling (de-identified) patient data from Swespine's registry. The Ahus data has previously been collected with informed consent and stored in anonymized form in accordance with approval from the Regional Ethics Committee for Medical Research (Gulbrandsen et al., 2013); since the data is anonymized, no ethical approval in Sweden was needed for using the data for the present study.

## ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for useful comments and suggestions. This work was supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

## REFERENCES

- Ahmed, S., Shamim Kaiser, M., Hossain, M. S., and Andersson, K. (2024). A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. *IEEE Access*, pages 1–1.

- Alonso, J. M. and Bugarín, A. (2019). ExpliClas: Automatic Generation of Explanations in Natural Language for Weka Classifiers. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE.
- Baaj, I. (2022). *Explainability of possibilistic and fuzzy rule-based systems*. Theses, Sorbonne Université.
- Berman, A. (2024a). Argumentative Dialogue As Basis For Human-AI Collaboration. In *Proceedings of HHA I 2024 Workshops*.
- Berman, A. (2024b). Too Far Away from the Job Market – Says Who? Linguistically Analyzing Rationales for AI-based Decisions Concerning Employment Support. *Weizenbaum Journal of the Digital Society*, 4(3).
- Breitholtz, E. (2020). *Enthymemes and Topoi in Dialogue: the use of common sense reasoning in conversation*. Brill.
- Eemeren, F. H., Garssen, B., and Labrie, N. (2021). *Argumentation between Doctors and Patients. Understanding clinical argumentative discourse*. John Benjamins Publishing Company.
- Forrest, J., Sripada, S., Pang, W., and Coghill, G. (2018). Towards making NLG a voice for interpretable machine learning. In *Proceedings of The 11th International Natural Language Generation Conference*. Association for Computational Linguistics (ACL).
- Fritzell, P., Mesterton, J., and Hagg, O. (2022). Prediction of outcome after spinal surgery—using The Dialogue Support based on the Swedish national quality register. *European Spine Journal*, pages 1–12.
- Grice, H. P. (1975). Logic and conversation. *Syntax and semantics*, 3:43–58.
- Gulbrandsen, P., Finset, A., and Jensen, B. (2013). Lege-pasient-korpus fra Ahus.
- Kaczmarek-Majer, K., Casalino, G., Castellano, G., Dominiak, M., Hryniewicz, O., Kamińska, O., Vessio, G., and Díaz-Rodríguez, N. (2022). Plenary: Explaining black-box models in natural language through fuzzy linguistic summaries. *Information Sciences*, 614:374–399.
- Lindgren, S. and Aspegren, K. (2004). *Kliniska färdigheter: informationsutbytet mellan patient och läkare*. Studentlitteratur AB.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- Maraev, V., Breitholtz, E., Howes, C., and Bernardy, J.-P. (2021). Why should I turn left? Towards active explainability for spoken dialogue systems. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 58–64.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.
- Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, page 333–342, New York, NY, USA. Association for Computing Machinery.
- Pantanowitz, L., Pearce, T., Abukhiran, I., Hanna, M., Wheeler, S., Soong, T. R., Tafti, A. P., Pantanowitz, J., Lu, M. Y., Mahmood, F., Gu, Q., and Rashidi, H. H. (2024). Nongenerative Artificial Intelligence in Medicine: Advancements and Applications in Supervised and Unsupervised Machine Learning. *Modern Pathology*, page 100680.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (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*, pages 1135–1144.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.
- Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16(none):1 – 85.
- Sbisà, M. (1987). Acts of explanation: A speech act analysis. *Argumentation: Perspectives and approaches*, pages 7–17.
- Slack, D., Krishna, S., Lakkaraju, H., and Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, 5(8):873–883.
- Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press.
- Wahde, M. and Virgolin, M. (2023). DAISY: An Implementation of Five Core Principles for Transparent and Accountable Conversational AI. *International Journal of Human-Computer Interaction*, 39(9):1856–1873.
- Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language*. PhD thesis, Massachusetts Institute of Technology.
- Xydis, A., Hampson, C., Modgil, S., and Black, E. (2020). Enthymemes in dialogues. In *Computational Models of Argument*, pages 395–402. IOS Press.