Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning
Reasoning, Learning & Decision-Making Track

Advancing Interactive Explainable AI via Belief Change Theory

Antonio Rago[1] and **Maria Vanina Martinez**[2]
[1]Department of Computing, Imperial College London, UK
[2]Artificial Intelligence Research Institute (IIIA-CSIC), Spain
a.rago@imperial.ac.uk, vmartinez@iiia.csic.es

Abstract

As AI models become ever more complex and intertwined in humans' daily lives, greater levels of interactivity of explainable AI (XAI) methods are needed. In this paper, we propose the use of belief change theory as a formal foundation for operators that model the incorporation of new information, i.e. user feedback in interactive XAI, to logical representations of data-driven classifiers. We argue that this type of formalisation provides a framework and a methodology to develop interactive explanations in a principled manner, providing warranted behaviour and favouring transparency and accountability of such interactions. Concretely, we first define a novel, logic-based formalism to represent explanatory information shared between humans and machines. We then consider real world scenarios for interactive XAI, with different prioritisations of new and existing knowledge, where our formalism may be instantiated. Finally, we analyse a core set of belief change postulates, discussing their suitability for our real world settings and pointing to particular challenges that may require the relaxation or reinterpretation of some of the theoretical assumptions underlying existing operators.

1 Introduction

To achieve the safe, regulated and trustworthy deployment of AI while maximising its potential, a number of applications benefit from interactive explanations, where a human provides feedback to the AI model (see (Wu et al., 2022) for a recent overview). Interactivity has also been recognised as a core tenet of ensuring that AI is *contestable* (Hirsch et al., 2017; Lyons, Velloso, and Miller, 2021), as recommended by design principles such as those of the ACM[1] and enforced by legal regulations such as the GDPR[2]. Meanwhile, the field of explainable AI (XAI), with its overarching objective of fostering trust in AI models, predominantly focuses on static explanations which do not support such interactivity (see (Ali et al., 2023) for an overview). Some XAI methods provide interactivity via user feedback, e.g. in human-in-the-loop reinforcement learning (Retzlaff et al., 2024), recommender systems (Rago et al., 2021) and text classification (Arous et al., 2021), where explainability has been said to be beneficial, but this research area remains relatively unexplored. Further, formal frameworks for interactivity in XAI are lacking, despite their crucial role in trustworthiness (Marques-Silva and Ignatiev, 2022), giving scant prospect for regulations on interactive XAI to be defined and systematically enforced.

In this paper, we propose the use of belief change theory (Alchourrón, Gärdenfors, and Makinson, 1985) within the modelling of interactive explanations for data-driven classifiers. We assume as given a set of explanations about a classifier, in the form of rules, e.g. as in (Guidotti et al., 2018; Ribeiro, Singh, and Guestrin, 2018; Shih, Choi, and Darwiche, 2018; Grover et al., 2019; Ignatiev, Narodytska, and Marques-Silva, 2019), and we envisage the possibility of users providing feedback thereon, also in the form of rules. We see *revision operators* as being particularly well-suited to modelling the process of feedback incorporation, as evidenced in the related setting of editing multi-label classifiers (Coste-Marquis and Marquis, 2021).

We argue that such formalisations lay the groundwork for the design and development of interactive explanations that promote transparency, interpretability and accountability in human-machine interactions. As an example of the importance of this topic, the recently endorsed AI Act[3] regulatory framework for the European Union, guarantees the *right of consumers to launch complaints and receive meaningful explanations*. Such legal requirements make it evident that novel methodologies and tools are needed to provide formal guarantees about not only AI models' behaviour but also about all related human-machine interactions.

After covering the related literature (§2), we make the following contributions:

- We define a novel, logic-based formalism to represent how information is shared between humans and machines, specifically classification models, in XAI (§3).
- We consider a set of real world scenarios of interactive XAI where our formalism may be instantiated with different prioritisations of new and existing knowledge (§4).
- We instantiate a core set of belief revision postulates in our formalism, discussing their strengths and weaknesses (§5), before looking ahead to what is required for belief revision to make advancements in interactive XAI (§6).

[1]https://www.acm.org/media-center/2022/october/tpc-statement-responsible-algorithmic-systems

[2]https://gdpr-text.com/read/article-22/

[3]https://artificialintelligenceact.eu/

2 Related Work

Within the area of belief revision the work of Falappa, Kern-Isberner, and Simari (2002) proposes a non-prioritised revision operator based on the use of explanations by deduction. The epistemic input is accompanied by an explanation supporting it and beliefs are dynamically qualified as defeasible or undefeasible and revised accordingly. Recently, Coste-Marquis and Marquis (2021) proposed a belief change operator, called a *rectification operator*, that aims to modify, according to some available background knowledge, a Boolean circuit that exhibits the same input-output behaviour as a multi-label classifier. The operation ensures that the rectified circuit complies with the background knowledge through different notions of compliance. Though this proposal also aims to model modifications to logical representations of classifiers through belief change operators, there exist significant differences. First, we assume partial and approximate knowledge of the classifier's behaviour and therefore a potentially incomplete and not coherent logical representation of it; this has a direct impact on the analysis of suitable postulates. Second, the classifier's representation and the feedback provided by users are specified by means of rules rather than propositional logical sentences. We believe this encoding provides greater interpretability from a user's point of view. Third, instead of prioritising the input or feedback, we study alternatives according to different scenarios of interactive explanations, allowing for the possibility for the logical representation to gradually differ from the original classifier specification as feedback is incorporated. Finally, the work from Schwind, Inoue, and Marquis (2023) proposes a series of operators that determine how a Boolean classifier should be edited whenever it does not label a data point in the correct way. The paper studies the incorporation of positive, negative and combined (positive and negative) instances. Besides only focusing on Boolean classifiers, the differences mentioned above for the multi-label approach also hold in this case.

3 Formalising Classifiers and Explanations

In this section, we formalise classifiers' outputs and explanations based on propositional logic and formal rules, extending the language from (Amgoud, 2023).[4]

We assume a single-label classification problem where $F = \{f_1, \ldots, f_m\}$ is the set of $m > 1$ features, where each $f_i \in F$ has a discrete domain $\mathcal{D}(f_i)$, and $C = \{c_1, \ldots, c_n\}$ is the set of $n > 1$ possible classes or classification labels. We let $V = \mathcal{D}(f_1) \times \ldots \times \mathcal{D}(f_m)$ be the (combinatorial) set of all possible data points, i.e. assignments of values to all features. Straightforwardly, we then let a dataset be a set of data points $D \subseteq V$. Then, a classifier $\mathcal{M} : D \to C$ is a total mapping[5] such that for any $\mathbf{x} \in D$, we say that $\mathcal{M}$ predicts class $c \in C$ iff $\mathcal{M}(\mathbf{x}) = c$. For a given $\mathbf{x}$, $\mathbf{x}^i$ is the value $v \in \mathcal{D}(f_i)$ assigned to feature f_i.

Syntax. To model a classifier, we assume a propositional language based on two finite alphabets $\mathcal{F} = f_1, \ldots, f_m$ and $\mathcal{C} = c_1, \ldots, c_n$, representing elements in F and C, resp. For each symbol f in $\mathcal{F}$, we assume a discrete set of constants $\mathcal{D}(f)$ corresponding to the domain ($\mathcal{D}(f)$) of feature $f \in F$.

A *feature atom* is of the form (f, v), where $f \in \mathcal{F}$ and $v \in \mathcal{D}(f)$; a feature literal is either a feature atom a or $\neg a$. On the other hand, a *classification atom* is of the form c, with $c \in \mathcal{C}$. Intuitively, a feature atom represents the fact that value v is assigned to feature f, while a classification atom represents a set of classes (in particular, an atom represents a singleton, as we will see later).

A feature (classification, resp.) formula is any logical formula built from feature (classification, resp.) literals using classical connectives $\neg, \wedge, \vee$. We use $\mathbb{F}$ ($\mathbb{C}$, resp.) to denote the set of all feature (classification, resp.) formulas.

We distinguish the following set of feature formulas, intuitively to link each of them to a specific data point in V.

Definition 1. *A* (data) instance x *is conjunction of feature atoms such that each feature $f \in \mathcal{F}$ appears exactly once. We will call $\mathcal{V}$ the set of all possible data instances.*

Intuitively, feature formulas represent sets of data points in V, while a data instance represents a specific data point in V. On the other hand, classification formulas represent sets of classification labels. The concept of a rule, defined below, allows us to map feature formulas into classification formulas, which ultimately seek to represent a mapping between data points and a set of potential classification labels.

Definition 2. *A* rule r *is of the form $\phi \Rightarrow \psi$, where $\phi \in \mathbb{F}$ and $\psi \in \mathbb{C}$. We call ϕ the body of r, denoted $body(r)$, and ψ the head of r, denoted $head(r)$. If ϕ is a data instance (i.e. $\phi \in \mathcal{V}$) and c is a positive literal, we call r an* instance rule.

Intuitively, we use a rule r to establish the set of classes, defined by the classification formula in the head of the rule, that is assigned to the set of data instances characterised by the feature formula $body(r)$. Note that when a rule establishes that a certain set of data points are assigned to a non-singleton set of classes, we interpret that any of those classes in the set could be assigned, but only one of them.

Semantics. Function $\mathcal{I}_f : \mathbb{F} \to 2^V$ maps feature formulas to sets of data points in V. Formally, for feature formulas ϕ, ψ:

- if $\phi = (f_i, v)$, then $\mathcal{I}_f((f_i, v)) = \{\mathbf{x} \in V \mid \mathbf{x}^i = v\}$
- $\mathcal{I}_f(\phi \wedge \psi) = \mathcal{I}_f(\phi) \cap \mathcal{I}_f(\psi)$
- $\mathcal{I}_f(\phi \vee \psi) = \mathcal{I}_f(\phi) \cup \mathcal{I}_f(\psi)$
- $\mathcal{I}_f(\neg\phi) = \{\mathbf{x} \in V \mid \mathbf{x} \notin \mathcal{I}_f(\phi)\}$
- $\mathcal{I}_f(\bot) = \varnothing, \mathcal{I}_f(\top) = V$

The semantics for classification formulas is defined with $\mathcal{I}_c : \mathbb{C} \to 2^C$, which maps classification formulas to sets of classes in C. Formally, for classification formulas ϕ, ψ:

- if $\phi = c$, where c is a classification atom, then $\mathcal{I}_c(c) = \{c\}$
- $\mathcal{I}_c(\phi \wedge \psi) = \mathcal{I}_c(\phi) \cap \mathcal{I}_c(\psi)$
- $\mathcal{I}_c(\phi \vee \psi) = \mathcal{I}_c(\phi) \cup \mathcal{I}_c(\psi)$
- $\mathcal{I}_c(\neg\phi) = \{c \in C \mid c \notin \mathcal{I}_c(\phi)\}$
- $\mathcal{I}_c(\bot) = \varnothing, \mathcal{I}_c(\top) = C$

To interpret rules in this setting, we define function $\mathcal{I}$ such that, given a rule r, $\mathcal{I}(r) = (\mathcal{I}_f(body(r)), \mathcal{I}_c(head(r)))$.

[4] In the supplementary material (available in the extended version of the paper at https://arxiv.org/abs/2408.06875) we provide example illustrations of our approach, as well as a proof of Theorem 1.

[5] Note that $\mathcal{M}$ is a total mapping wrt D, i.e. the data points for which the classes predicted by the classifier are known. Here, D may represent any dataset, e.g. that used for training.

Here, $\mathcal{I}$ maps the set of data points represented by the formula $body(r)$ into a set of classes determined by $head(r)$.

Using the language defined above, we can logically model classifiers, and explanations therefor (defined later), by means of rules since they express mappings of sets of data points into sets of classification labels. In addition to extending the representation language from (Amgoud, 2023), note that the spirit of the aforementioned paper is different to ours. In that work, the authors formally define functions that generate different types of explanations and study their properties in relation to existence and correctness. In this work, we assume explanations from a classifier have already been provided in the form of rules and we model the interactions with the model's users through operations that could update such rules as a result of the user's feedback.

Based on the semantics, we can now define notions that help us establish relationships among rules. The first is *enforcement*: intuitively, a set of rules enforces another set of rules whenever every possible assignment of data points to a class that the enforced set of rules represents is also an assignment that is represented by the enforcing set.

Definition 3. *Given sets of rules $\mathcal{R}_i$ and $\mathcal{R}_j$, $\mathcal{R}_i$* enforces *$\mathcal{R}_j$, denoted $\mathcal{R}_i \sqsupseteq \mathcal{R}_j$, iff $\forall r_j \in \mathcal{R}_j$, $\forall \mathbf{x} \in \mathcal{I}_f(body(r_j))$, $\exists r_i \in \mathcal{R}_i$, such that $\mathbf{x} \in \mathcal{I}_f(body(r_i))$ and $\mathcal{I}_c(head(r_i)) \subseteq \mathcal{I}_c(head(r_j))$.*

The second notion we define is *consistency*, requiring that sets of rules do not assign incompatible labels to data points.

Definition 4. *Given a set of rules $\mathcal{R}$, $\mathcal{R}$ is* consistent *iff $\forall \mathbf{x} \in \bigcup_{r \in \mathcal{R}} \mathcal{I}_f(body(r))$, $\forall r_i, r_j \in \mathcal{R}$ such that $\mathbf{x} \in \mathcal{I}_f(body(r_i))$ and $\mathbf{x} \in \mathcal{I}_f(body(r_j))$ then $\mathcal{I}_c(head(r_i)) \cap \mathcal{I}_c(head(r_j)) \neq \varnothing$. Otherwise, $\mathcal{R}$ is* inconsistent.

The notion of *coherence* defined below aims to capture the relationship between rules and models. Intuitively, given a threshold $\tau \in [0, 1]$, a set of rules is τ-coherent with a model iff the proportion of instances captured by the body of every rule, such that the model's classification of the instance is included in the head of the rule, is at least τ. This generalises the notion of *compatibility* in (Amgoud, 2023) allowing a percentage of the classifications described by the set of rules to differ from the classifications provided by the classifier.

Definition 5. *Given a classifier $\mathcal{M}$ and a* threshold *$\tau \in [0, 1]$, we say that rule r is τ-*coherent *with $\mathcal{M}$ iff $\mathcal{I}_f(body(r)) \cap D = \varnothing$ or:*

$$\frac{|\{\mathbf{x} \in \mathcal{I}_f(body(r)) \cap D | \mathcal{M}(\mathbf{x}) \in \mathcal{I}_c(head(r))\}|}{|\mathcal{I}_f(body(r)) \cap D|} \geq \tau$$

*We say that a set of rules $\mathcal{R}$ is τ-*coherent *with $\mathcal{M}$ iff $\forall r \in \mathcal{R}$, r is τ-*coherent*. Whenever $\tau = 1$, we drop the τ prefix and say that a (set of) rule(s) is* coherent *with $\mathcal{M}$.*

Lemma 1. *Given a classifier $\mathcal{M}$, a rule r is coherent with $\mathcal{M}$ iff $\forall \mathbf{x} \in \mathcal{I}_f(body(r)) \cap D$, $\mathcal{M}(\mathbf{x}) \in \mathcal{I}_c(head(r))$.*

Next, we formalise whether a set of rules completely (and exclusively) represents the set of known data points D.

Definition 6. *Given a classifier $\mathcal{M}$, a set of rules $\mathcal{R}$ is* complete *for $\mathcal{M}$ iff $\mathcal{R}$ is a set of instance rules such that $|\mathcal{R}| = |D|$ and $\forall \mathbf{x} \in D$, $\exists r \in \mathcal{R}$ where $\mathcal{I}(r) = (\{\mathbf{x}\}, \{\mathcal{M}(\mathbf{x})\})$.*

We now represent the knowledge we have about a classifier by means of rules as follows.

Definition 7. *Given a classifier $\mathcal{M}$, an* explanation knowledge base *for $\mathcal{M}$ is a set $\mathcal{K}_\mathcal{M} = \mathcal{K}_d \cup \mathcal{K}_e$, where $\mathcal{K}_d$ is a set of instance rules, called the* data*, and $\mathcal{K}_e$ is a set of general rules, called the* explanations*.*

Our intention is for $\mathcal{K}_d$ to represent data points for which the classification is known, these may come either from training or evaluation phases or from previous use of the classifier. $\mathcal{K}_d$ logically represents the classifier, such that they encode exactly the same classifications. In addition to this, there exist different methods in the literature to elicit behavioural patterns from classifiers, often expressed as rules functioning as explanations, we use $\mathcal{K}_e$ to represent that kind of knowledge. Although consistency is generally expected, a priori we impose no restrictions of coherence of $\mathcal{K}_e$ with $\mathcal{M}$, as they represent tentative knowledge obtained from potentially imprecise methods. These explanation rules may have been extracted by existing formal methods for explaining (discrete) classifiers from the literature, such as (Guidotti et al., 2018; Ribeiro, Singh, and Guestrin, 2018; Shih, Choi, and Darwiche, 2018; Grover et al., 2019; Ignatiev, Narodytska, and Marques-Silva, 2019). Moreover, τ-coherence corresponds to the notion of precision in (Ribeiro, Singh, and Guestrin, 2018), and could be used to allow for tolerance in the correctness of explanations. However, for this paper we will assume that $\tau = 1$.

Theorem 1. *Given a classifier $\mathcal{M}$ and an explanation knowledge base $\mathcal{K}_\mathcal{M} = \mathcal{K}_d \cup \mathcal{K}_e$, where $\mathcal{K}_d$ is complete for and coherent with $\mathcal{M}$, a set of rules $\mathcal{R}$ is coherent with $\mathcal{M}$ iff $\mathcal{R} \cup \mathcal{K}_d$ is consistent.*

This result shows that preserving consistency with $\mathcal{K}_d$ preserves coherence with $\mathcal{M}$ (see the consistency postulate in §5) whenever $\mathcal{K}_d$ is complete and coherent with $\mathcal{M}$. Note, however, that we do not make this assumption in general, since, as discussed later, we aim for a framework that is tolerant to inconsistency and in which the logical representation of $\mathcal{M}$ may differ from $\mathcal{M}$ due to feedback incorporation.

4 Interactive Explanations

We now demonstrate how interactive explanations may be modelled with our formalism, and consider how such explanations may be deployed in real world settings. We consider interactive explanations which give users the ability to provide feedback to classifiers in a number of ways, in the form of rules (Definition 2), which we call here *feedback*. When a rule is provided as feedback (we limit to single rules), the goal is to analyse how/if this knowledge can be incorporated, possibly modifying both the explanation knowledge base and the feedback itself. This type of feedback mirrors rule-based explanations from XAI (as we discuss in §3) that intuitively represent knowledge in any domain and easily translate to and from natural language.

We define the following basic desiderata for this process:

- *Constrained Inconsistency*: specific scenarios may require some tolerance to inconsistency, e.g. requiring only $\mathcal{K}_d$ to be kept consistent after an interaction.

- *Bounded model incoherence*: while we expect $\mathcal{K}_d$ to be coherent with $\mathcal{M}$, the weaker notion of τ-coherence of the explanations $\mathcal{K}_e$ with $\mathcal{M}$ could be accepted, for a given τ.
- *Minimal information loss*: the information contained in $\mathcal{K}_\mathcal{M} \cup \{r\}$ should be modified or removed minimally, and only when it jeopardises the above desiderata.

Belief revision incorporates new information following two main principles: consistency (preservation) and minimal change. Our desiderata for interactive explanations coincide with these aims in minimising the amount of information loss. However, we relax the notion of consistency, and allow the classifier and its logical representation to drift in a restricted manner through the notion of τ-coherence.

We now give a (non-exhaustive) set of real world application settings where interactive explanations may be deployed. We base our settings on those proposed in (Retzlaff et al., 2024) for human-in-the-loop systems.

In the first setting, named **S1**, we envisage a classification model which is in **development**, e.g. being debugged by a developer as in (Lertvittayakumjorn, Specia, and Toni, 2020). Here, the user provides feedback to update, and correct, the model. In this case, the model's trust in the feedback can be regarded as **credulous**, since the model should be updated to align with the feedback, i.e. any feedback r takes priority over the existing knowledge (informally represented with $r \succ \mathcal{K}_\mathcal{M}$). For example, if a user provides r which contradicts an instance rule representing an existing data point in $\mathcal{K}_d$, e.g. due to the default settings of the model or changing preferences of the user, we would like to incorporate r and update the conflicting instance rule to align with the new conditions specified by the user.

In the second setting, **S2**, we introduce a model which is being refined in an **evaluation** stage by group of users, e.g. as in domain expert information fusion (Holzinger et al., 2021), before the model is deployed at scale. In this case, a single model is being updated by feedback from multiple users, and so the model's trust in the feedback must be **balanced** with that in the existing knowledge. Here, a single user's feedback should not necessarily take precedence over existing knowledge (informally, $r \simeq \mathcal{K}_\mathcal{M}$), and so both the new and the existing knowledge may be modified in order for consistency and coherence with the model to be maintained with minimal information loss. For example, if a user provides r which contradicts $\mathcal{K}_e$, it may be desirable to preserve $\mathcal{K}_d$ but modify $\mathcal{K}_e$ or r by weakening or rejection to incorporate as much of the new knowledge as possible.

In the final setting, **S3**, we consider a model which has already undergone commercial **deployment** at scale, but allows for feedback from the sizeable group of users for completing gaps in the knowledge, e.g. as in autonomous vehicles (Wu et al., 2023). Here, the model will be updated by users' feedback, but due to the size of this group and the fact that the model has already been deployed commercially, e.g. potentially raising legal issues, the trust in the feedback is **sceptical**, and it thus prioritises existing over new knowledge (informally, $r \prec \mathcal{K}_\mathcal{M}$). The new knowledge can thus be modified in order to ensure its consistency with the existing knowledge. For example, if the user provides some r which does not violate the consistency of $\mathcal{K}_e$ or the coherence of $\mathcal{K}_d$, then it may be incorporated to $\mathcal{K}_\mathcal{M}$ as is to minimise information loss. Meanwhile, if it contradicts $\mathcal{K}_\mathcal{M}$, then we may preserve $\mathcal{K}_\mathcal{M}$ while only part of r may be incorporated.

It is important to note that these modifications are not performed over the model itself but its logical representation, creating in each interaction a new knowledge base that may differ substantially from the original knowledge base (and the model). A distance between different versions of the knowledge base could be measured through τ-coherence or more conventional distance measures, and be used as a way of checking the effect of feedback, e.g. as an indicator for when the retraining of the model is required.

Having presented out motivational scenarios, in the next section we analyse the suitability of belief revision operators to model interactions with explanation knowledge bases.

5 Revision of Explanation KBs

One of the main contributions of the foundational models of belief change is the development of a style of research and development methodology based on providing axiomatic characterisations of the operators' behaviour in terms of postulates. The postulates focus on conditioning and constraining the inputs and the results of the operators, rather than providing insights into how the results are achieved. Representation theorems are used both to provide semantic characterisations for belief change operators, as well as linking these characterisations to computational implementations, providing provable guarantees on the behaviour of such algorithms. In the following we analyse a core set of postulates for belief base revision (Hansson, 1993), translate them in our logical setting and discuss their suitability with respect to the different scenarios of interactive explanations.

In this work, we adopt the approach to belief revision known as *base revision*, where existing knowledge is represented as a finite set of formulas (Hansson, 1993), which we call an explanation knowledge base $\mathcal{K}_\mathcal{M}$, as described in §3. The new information consists of a single rule r that is obtained from the interaction with the user(s) of the model. In the following analysis we use $\mathcal{K}_\mathcal{M} * r$ to describe the application (and the results) of operator $*$ over the existing knowledge base $\mathcal{K}_\mathcal{M}$ and input (feedback) r.

Success states that the epistemic input is always accepted, i.e. new knowledge is prioritised. This can be formalised in our framework by means of our notion of enforcement of the feedback rule, i.e. $\mathcal{K}_\mathcal{M} * r \supseteq \{r\}$. In setting S1, the success postulate can be used to enforce the feedback taking priority over the existing $\mathcal{K}_\mathcal{M}$ (in the presence of inconsistency). Prioritised revision operators are suitable for this setting, while this is not the case for (possibly S2 and) S3, where the existing knowledge should be prioritised. A first approach to define non-prioritised behaviour could be modelled by a simple *relative success* postulate (Fermé, Mikalef, and Taboada, 2003), which states that either the input is fully (explicitly) accepted or rejected, i.e. either $r \in \mathcal{K}_\mathcal{M} * r$ or $\mathcal{K}_\mathcal{M} * r = \mathcal{K}_\mathcal{M}$, resp. More fine-grained alternatives would allow for the specification of conditions under which the input could be fully or partially accepted. For instance, *weak success* (Resina et al., 2020) may state that if $\mathcal{K}_\mathcal{M} \cup \{r\}$ is

consistent then $\mathcal{K}_\mathcal{M} * r \sqsupseteq \{r\}$. Meanwhile, *proxy success* and *weak proxy success* (Resina et al., 2020) state that the revision should incorporate part of the input, e.g. to ensure all users' feedback plays a part in S2. Formally, proxy success could be defined requiring that $\exists r'$ such that $\{r\} \sqsupseteq \{r'\}$, $\mathcal{K}_\mathcal{M} * r \sqsupseteq \{r'\}$ and $\mathcal{K}_\mathcal{M} * r = \mathcal{K}_\mathcal{M} * r'$. In weak proxy success, r' is not conditioned by r: $\exists r'$ such that $\mathcal{K}_\mathcal{M} * r \sqsupseteq \{r'\}$ and $\mathcal{K}_\mathcal{M} * r = \mathcal{K}_\mathcal{M} * r'$. These weakened postulates seem appropriate for S2 and S3, where gaps in $\mathcal{K}_\mathcal{M}$ could be filled more often with these weaker constraints, but less so for S1, where success may be preferred given the trust in the user here. However, any version of success that allows for the incorporation of only part of a rule could induce bias in the dataset. A potentially problematic example could be when only a stricter version of a feedback rule, covering only a subset of a feature (e.g. an ethnic minority in a population), rather than its entirety, is incorporated to $\mathcal{K}_\mathcal{M}$.

Inclusion states that the only addition to the existing knowledge can be the feedback itself, instantiated in our setting as $\mathcal{K}_\mathcal{M} * r \subseteq \mathcal{K}_\mathcal{M} \cup \{r\}$. This raises issues in our settings, since it may be desirable that $\mathcal{K}_d$, $\mathcal{K}_e$ or both are modified, for instance making rules more specific. In S1, it is desirable that we incorporate r as is, but we may wish for $\mathcal{K}_e$ to be adapted to this new information. Also, in S2 and S3, we may want to incorporate only part of r, since it may be unrealistic to incorporate r in its entirety given the higher priority of $\mathcal{K}_\mathcal{M}$. An alternative is *weak inclusion* (Resina et al., 2020), which states that if $r \in \mathcal{K}_\mathcal{M} * r$, then $\mathcal{K}_\mathcal{M} * r \subseteq \mathcal{K}_\mathcal{M} \cup \{r\}$. This relaxation alleviates the second aforementioned issue, and we would thus posit that this is desirable in S3, where existing and new information is restricted from modification, e.g. from a legal standpoint if users have already seen it. However, in S2 we would expect that $\mathcal{K}_e$ being adapted to r would be more suitable. We thus propose three alternate formulations of inclusion based on our notion of enforcement, prioritising the suitable data in each setting. For S1, we suggest that $\mathcal{K}_\mathcal{M} * r \subseteq A \cup \{r\}$, where $\mathcal{K}_\mathcal{M} \sqsupseteq A$, allowing the existing explanations to adapt to the new information. For S2, we suggest that $\mathcal{K}_\mathcal{M} * r \subseteq A$, where $\mathcal{K}_\mathcal{M} \cup \{r\} \sqsupseteq A$, allowing for the modification of both existing and new information. For S3, we suggest that $\mathcal{K}_\mathcal{M} * r \subseteq \mathcal{K}_\mathcal{M} \cup A$, where $\{r\} \sqsupseteq A$, ensuring that only the feedback is modified.

Consistency conventionally requires that a knowledge base becomes consistent after the revision, even if it is not so beforehand. Formally, $\mathcal{K}_\mathcal{M} * r$ is required to be consistent,[6] which, by Theorem 1, may cover the first two of our desiderata whenever $\mathcal{K}_d$ is coherent with $\mathcal{M}$. Thus, the notion of consistency seems to be desirable across our settings, whenever neither consistency nor coherence is relaxed. In particular, *consistency preservation* (Alchourrón, Gärdenfors, and Makinson, 1985), which requires that a consistent KB be consistent after operating (adding the condition that $\mathcal{K}_\mathcal{M}$ is consistent to the consistency postulate above) seems suitable for all settings, since it requires feedback not introduce such inconsistencies, rather than requiring it fix any which already exist. Note, however, that it may be the case that we are interested in only $\mathcal{K}_d$ remaining/becoming consistent after the revision, given the tentative and approximate nature of $\mathcal{K}_e$. An alternative to be considered is to ensure that the revision does not increase the amount of inconsistency (given a measure for it (Thimm, 2016; Grant and Martinez, 2018)) in $\mathcal{K}_\mathcal{M}$ or $\mathcal{K}_d$.

[6] A singleton set containing r is consistent by Definition 4, so our version of the postulate does not condition on the consistency of the input. Allowing for sets of feedback rules, as in *multiple revision* (Fuhrmann and Hansson, 1994), is future work.

	S1	**S2**	**S3**
Trust	Credulous	Balanced	Sceptical
Setting	Development	Evaluation	Deployment
Users	Single	Small-Scale	Large-Scale
Priority	$r \succ \mathcal{K}_\mathcal{M}$	$r \simeq \mathcal{K}_\mathcal{M}$	$r \prec \mathcal{K}_\mathcal{M}$
Success	✓	$\checkmark^{rs,ws,ps,wps}$	$\checkmark^{ws,ps,wps}$
Inclusion	-	-	$\checkmark^{wi}$
Consistency	$\checkmark^{cp}$	$\checkmark^{cp}$	$\checkmark^{cp}$
Relevance	✓	✓	✓
Uniformity	✓	✓	✓

Table 1: Characteristics of our real world settings and assessment of postulates, where ✓ indicates a postulate is desirable, $\checkmark^x$ indicates that only a weaker postulate is desirable and – indicates novel postulates may be required, with x indicating the following weaker postulates: relative success, weak success, proxy success, weak proxy success, weak inclusion and consistency preservation.

Relevance concerns minimal change of existing knowledge, stating that if $r' \in \mathcal{K}_\mathcal{M}$ and $r' \notin \mathcal{K}_\mathcal{M} * r$, then there is a set of rules $\mathcal{R}$ such that $\mathcal{K}_\mathcal{M} * r \subseteq \mathcal{R} \subseteq \mathcal{K}_\mathcal{M} \cup \{r\}$, $\mathcal{R}$ is consistent and $\mathcal{R} \cup \{r'\}$ is inconsistent. Relevance formalises our third desideratum in terms of only removing information from the data or explanations if it were inconsistent with the feedback being provided by the user, rendering it suitable across our settings. This postulate has important implications for data protection, ensuring that the non-conflicting knowledge is preserved and therefore is desirable in all three settings. However, as defined above, this postulate forces $\mathcal{K}_\mathcal{M} * r$ to be a subset of $\mathcal{K}_\mathcal{M} \cup \{r\}$; in light of our previous discussion, if we want to have the possibility of not only deleting but modifying both the existing knowledge and feedback, we could consider a weaker notion closer to the postulate known as *core-retainment*: in our setting this could be formalised as if $r' \in \mathcal{K}_\mathcal{M}$ and $r' \notin \mathcal{K}_\mathcal{M} * r$, then there is a set of rules $\mathcal{R}$ such that $\mathcal{R} \subseteq \mathcal{K}_\mathcal{M} \cup \{r\}$, $\mathcal{R}$ is consistent but $\mathcal{R} \cup \{r'\}$ is inconsistent.

Uniformity, formulated in our setting, states that if $\forall \mathcal{R} \subseteq \mathcal{K}_\mathcal{M}$, $\mathcal{R} \cup \{r\}$ is inconsistent if and only if $\mathcal{R} \cup \{r'\}$ is inconsistent, then $\mathcal{K}_\mathcal{M} \cap (\mathcal{K}_\mathcal{M} * r) = \mathcal{K}_\mathcal{M} \cap (\mathcal{K}_\mathcal{M} * r')$. The intuition here is that if r and r' are inconsistent with $\mathcal{K}_\mathcal{M}$ in the exact same way, revising by either retains the same knowledge from $\mathcal{K}_\mathcal{M}$. Once again, uniformity seems to be appropriate across the settings, guaranteeing the regularity of the effects of feedback, which could be useful for ensuring that regulatory guidelines are met.

6 Discussion and Future Work

Table 1 summarises the results of our analysis. Some of the existing postulates are suitable for all of these settings in

their original form, i.e. relevance and uniformity, while the others require alternate versions from the literature. However, across all studied postulates, we believe that there is scope for novel, tailored versions which may be more suitable in the individual settings, as we have indicated. Indeed, even in the cases where there are suitable postulates, others may be preferable, e.g. as we suggested for success. We believe that this highlights many fruitful avenues for future work. Among these, a next step is to characterise the behaviour of each setting with a specific set of postulates and provide the corresponding constructions. Regarding constructions, it seems possible to implement S1 with minimal modifications to traditional belief revision base operators such as *partial meet* and *kernel* (Hansson, 1993). The other two of our envisaged settings lend themselves to non-prioritised revisions that could be implemented through operators such as *credibility limited* (Fermé, Mikalef, and Taboada, 2003) and *screened revision* (Makinson, 1997), in which a portion of the knowledge $\mathcal{K}_p \subseteq \mathcal{K}_d \cup \mathcal{K}_e$ is protected from the revision. For example, it may be the case that unless data points from the dataset $\mathcal{K}_d$ are explicitly mentioned in the feedback, then we protect $\mathcal{K}_d$ from changes, i.e. $\mathcal{K}_p = \mathcal{K}_d \setminus \{r\}$, and modify only explanations. In S2, $\mathcal{K}_e$ may be seen as being modifiable while $\mathcal{K}_d$ is protected (no matter what r is being provided), i.e. $\mathcal{K}_p = \mathcal{K}_d$, for example if the dataset has been curated to be unbiased. Another case could be when a subset of $\mathcal{K}_d \cup \mathcal{K}_e$ needs to be protected from the revision, for example rules representing data points or explanations which have already been delivered to users, *semi-revision* (Hansson, 1997) could be useful here as it allows r to be discarded. Our analysis also suggests that for S2 and S3 it may be desirable to only retain part of the information contained in r. The closest operator in the literature that behaves in this way is *selective revision* (Resina et al., 2020). All these operators are implemented based on classical AGM operators, either checking conditions or modifying the input before applying an AGM revision operator or recurring to other operators such us consolidation (restoring consistency) over $\mathcal{K}_{\mathcal{M}} \cup \{r\}$. For setting S2 and S3 we may need to combine their implementations.

In light of the discussion about consistency, we need to define alternative postulates that better satisfy our proposed desiderata, including tolerating some degree of inconsistency and τ-coherence of $\mathcal{K}_e$ with the model for $\tau \neq 1$. Finally, our analysis assumes independence of interactions and that feedback consists of a single rule. Operators such as those based on *iterative revision* (Darwiche and Pearl, 1994) and *improvement* (Konieczny and Pérez, 2008) are worth studying for continuous feedback, e.g. coming from different users or over time, as well as multiple revision (Fuhrmann and Hansson, 1994) in order to allow arbitrary sets of rules as feedback. We leave exploration of these lines of research to future work.

Acknowledgments

Rago was partially funded by the ERC under the EU's Horizon 2020 research and innovation programme (No. 101020934, ADIX) and by J.P. Morgan and by the Royal Academy of Engineering, UK. Martinez was partially supported by the Spanish project PID2022-139835NB-C21 funded by MCIN/AEI/10.13039/501100011033, PIE 20235AT010 and iTrust (PCI2022-135010-2). The authors thank Musaab Ahmed Mahjoub Ahmed for his feedback.

References

Alchourrón, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *J. Symb. Log.* 50(2):510–530.

Ali, S.; Abuhmed, T.; El-Sappagh, S. H. A.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Ser, J. D.; Rodríguez, N. D.; and Herrera, F. 2023. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* 99:101805.

Amgoud, L. 2023. Explaining black-box classifiers: Properties and functions. *Int. J. Approx. Reason.* 155:40–65.

Arous, I.; Dolamic, L.; Yang, J.; Bhardwaj, A.; Cuccu, G.; and Cudré-Mauroux, P. 2021. MARTA: leveraging human rationales for explainable text classification. In *AAAI*, 5868–5876.

Coste-Marquis, S., and Marquis, P. 2021. On belief change for multi-label classifier encodings. In *IJCAI*, 1829–1836.

Darwiche, A., and Pearl, J. 1994. On the logic of iterated belief revision. In *TARK*, 5–23.

Falappa, M. A.; Kern-Isberner, G.; and Simari, G. R. 2002. Explanations, belief revision and defeasible reasoning. *Artif. Intell.* 141(1/2):1–28.

Fermé, E. L.; Mikalef, J.; and Taboada, J. 2003. Credibility-limited functions for belief bases. *J. Log. Comput.* 13(1):99–110.

Fuhrmann, A., and Hansson, S. O. 1994. A survey of multiple contractions. *J. Log. Lang. Inf.* 3(1):39–75.

Grant, J., and Martinez, M. 2018. *Measuring Inconsistency in Information*. Studies in logic. College Publications.

Grover, S.; Pulice, C.; Simari, G. I.; and Subrahmanian, V. S. 2019. BEEF: balanced english explanations of forecasts. *IEEE Trans. Comput. Soc. Syst.* 6(2):350–364.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; and Giannotti, F. 2018. Local rule-based explanations of black box decision systems. *CoRR* abs/1805.10820.

Hansson, S. O. 1993. Reversing the Levi identity. *J. Philos. Log.* 22(6):637–669.

Hansson, S. 1997. Semi-revision. *Journal of Applied Non-Classical Logics* 7(1-2):151–175.

Hirsch, T.; Merced, K.; Narayanan, S. S.; Imel, Z. E.; and Atkins, D. C. 2017. Designing contestability: Interaction design, machine learning and mental health. In *DIS*, 95–99.

Holzinger, A.; Malle, B.; Saranti, A.; and Pfeifer, B. 2021. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. *Inf. Fusion* 71:28–37.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. Abduction-based explanations for machine learning models. In *AAAI*, 1511–1519.

Konieczny, S., and Pérez, R. P. 2008. Improvement operators. In *KR*, 177–187.

Lertvittayakumjorn, P.; Specia, L.; and Toni, F. 2020. FIND: human-in-the-loop debugging deep text classifiers. In *EMNLP*, 332–348.

Lyons, H.; Velloso, E.; and Miller, T. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proc. ACM Hum. Comput. Interact.* 5(CSCW1):106:1–106:25.

Makinson, D. 1997. Screened revision. *Theoria* 63(1-2):14–23.

Marques-Silva, J., and Ignatiev, A. 2022. Delivering trustworthy AI through formal XAI. In *AAAI*, 12342–12350.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; Lagnado, D. A.; and Toni, F. 2021. Argumentative explanations for interactive recommendations. *Artif. Intell.* 296:103506.

Resina, F.; Garapa, M.; Wassermann, R.; Fermé, E.; and Reis, M. D. L. 2020. Choosing what to believe - new results in selective revision. In *KR*, 687–691.

Retzlaff, C. O.; Das, S.; Wayllace, C.; Mousavi, P.; Afshari, M.; Yang, T.; Saranti, A.; Angerschmid, A.; Taylor, M. E.; and Holzinger, A. 2024. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. *J. Artif. Intell. Res.* 79:359–415.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI*, 1527–1535.

Schwind, N.; Inoue, K.; and Marquis, P. 2023. Editing boolean classifiers: A belief change perspective. In *AAAI*, 6516–6524.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A symbolic approach to explaining bayesian network classifiers. In *IJCAI*, 5103–5111.

Thimm, M. 2016. On the expressivity of inconsistency measures. *Artif. Intell.* 234:120–151.

Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; and He, L. 2022. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* 135:364–381.

Wu, J.; Huang, Z.; Hu, Z.; and Lv, C. 2023. Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving. *Engineering* 21:75–91.