information

*Article*

# Explanation of Student Attendance AI Prediction with the Isabelle Infrastructure Framework †

**Florian Kammüller *** **and Dimpy Satija**

Department of Computer Science, Middlesex University London, London NW4 4BT, UK; d.satija@live.mdx.ac.uk
* Correspondence: f.kammueller@mdx.ac.uk

† This is a revised and extended version of a paper entitled F. Kammüller. Explanation of black-box AI for GDPR Related Privacy Using Isabelle. Presented at the Data Privacy Workshop DPM 2023, Copenhagen, Denmark, 29 September 2022.

**Abstract:** Right from the beginning, attendance has played an important role in the education systems, not only in student success but in the overall interest of the matter. Although all schools try to accentuate good attendance, still some schools find it hard to achieve the required level (96% in UK) of average attendance. The most productive way of increasing the pupils' attendance rate is to predict when it is going to go down, understand the reasons—why it happened—and act on the affecting factors so as to prevent it. Artificial intelligence (AI) is an automated machine learning solution for different types of problems. Several machine learning (ML) models like logistic regression, decision trees, etc. are easy to understand; however, complicated (Neural Network, BART etc.) ML models are not transparent but are black-boxes for humans. It is not always evident how machine intelligence arrived at a decision. However, not always, but in critical applications it is important that humans can understand the reasons for such decisions. In this paper, we present a methodology on the application example of pupil attendance for constructing explanations for AI classification algorithms. The methodology includes building a model of the application in the Isabelle Insider and Infrastructure framework (IIIf) and an algorithm (PCR) that helps us to obtain a detailed logical rule to specify the performance of the black-box algorithm, hence allowing us to explain it. The explanation is provided within the logical model of the IIIf, thus is suitable for human audiences. It has been shown that the RR-cycle of IIIf can be adapted to provide a method for iteratively extracting an explanation by interleaving attack tree analysis with precondition refinement, which finally yields a general rule that describes the decision taken by a black-box algorithm produced by Artificial intelligence.

**Keywords:** explainable AI; artificial intelligence; machine learning; IIIf (Isabelle Insider and Infrastructure framework); attack trees; precondition refinement; students' attendance

check for updates

**Citation:** Kammüller, F.; Satija, D. Explanation of Student Attendance AI Prediction with the Isabelle Infrastructure Framework. *Information* **2023**, *14*, 453. https://doi.org/10.3390/info14080453

Academic Editor: María N. Moreno García

Received: 1 June 2023
Revised: 24 June 2023
Accepted: 13 July 2023
Published: 10 August 2023

## 1. Introduction

> "Class attendance is a puzzle, buildings are built, rooms reserved, teaching schedules are set, and students enroll with the assumption that faculty-student encounters will occur. Yet quite often many students do not show up. " [1]

Students' absenteeism is an important issue for schools all over the globe because of its negative impacts on overall student performance, especially their grades. Sometimes absences due to illness or some other (religious, bereavement, or tragedies) reasons are unavoidable; however it is known that quite often more students are absent than could possibly be unwell [1,2].

It is surprising that even having strict government policies in place, still students choose not to attend class. Do certain factors encourage absenteeism? Does a particular type of student attend or miss classes regularly? What do students or parents say about why they did not attend school? Will students attend or not attend school next day or next week? We do not have clear information regarding these questions.

Artificial intelligence is an automated machine learning solution for various problems these days but predicting students' attendance, especially at school level, is a topic that has not been explored significantly before (as cited in [3]). Attendance data contain valuable information and predictors about students, analyzing this knowledge and training the data facilitates in predicting students' attendance based on these predictors.

Machine learning models like decision trees or linear regression are compatible with human understanding, whereas neural networks and deep learning models, although successful but complicated machine learning models, are black-boxes for humans. The decisive intelligence of these methods is not clear from the outside. Though it is not always necessary that humans need to know what is going on, their understanding is mandatory for certain critical applications [4].

It is essential to understand and have a clear explanation of the black-box algorithms as these algorithms are trained on datasets which may contain human biases. These biases are expected to be unseen in the opaqueness of the learning algorithm but may become apparent when an explanation is applied. The Isabelle Insider and Infrastructure framework (IIIf) provides explanations of black-box machine learning decisions to humans as it offers rich contexts for actors, infrastructures, and policies. This can be particularly relevant for privacy-critical applications [5]. In this study, we have used this existing method and applied it to the example of pupil attendance in schools to illustrate a nd validate the proposed methodology. Pupils' personal characteristics (gender, location, special needs etc.) were used in the automated decision process of attendance prediction guided by a black-box AI algorithm.

Attendance matters not just because it facilitates the transfer of knowledge about a particular topic/subject area or student activity but being present in class provides pupils with access to other resources, non-content-specific contextual information, and interactions with teachers and peer groups that can positively impact their understanding and sense of connectedness. Classroom learning provides them with far more than subject knowledge—it helps students to understand better and provides orientations that they may not be able to get from outside. Attending classroom sessions helps pupils to stay on track, understand what is expected from them, and nurture essential peer social relationships which promote a sense of belonging [6].

When we think of attendance, the key area of focus is how pupil presence relates to the understanding of classroom learning and how absence affects their knowledge and skills.

As argued by Vissers (2018) [3], it is expected that the students who do not attend class need to put extra efforts into studying at home to catch-up with the students who do attend the class in order to reach the same attainment level. Additionally, absenteeism can make students less well educated because they (absent students) miss the significant part of learning which is dealt with during a class resulting in study delay. Based on students' grades, it is not always possible for learning providers to detect low performers and intervene in time. If teachers know, in advance, that students will or will not attend class, they may be able to intervene earlier and avoid students' bad performance.

> "Instead of a reactive state of mind through evaluating students' grades, a proactive state of mind can be created by increasing class attendance." [3]

The most effective technique for increasing student attendance rate is to predict when it is going to happen, understand the reason behind its occurrence, and act on the causal factors in order to prevent it.

It is difficult to find the exact reason (other than illness or other calamities) for most of the absences. Quite often, environmental factors are used by teachers to assume the pupils' absences. Such clues can be extremely unreliable, so we decided to undertake this empirical investigation to distill reality from myths about students' absenteeism.

### *1.1. Contribution*

This paper extended the work by the first author [5] on conceiving a so-called precondition refinement algorithm to explore infrastructure models given in the IIIf to provide logical explanations for black-box AI decisions [5]. The contribution of this paper is to further validate the previous method with a new case study from the education sector. This application leads to consolidating and further refining the existing PCR algorithm as well as to a deeper understanding of the potentials and limitations of the IIIf explanation approach.

### *1.2. Overview of the Paper*

We first discuss some related work in Section 2 before we give some motivation on explainability and a background on the IIIf in Section 3 to set the scene. In Section 4, we introduce our case study of explaining school attendance predictions given by black-box algorithms, showing how it is represented in the IIIf. Section 5 recalls the definitions of the PCR algorithm, demonstrating and evaluating its application to the case study. Section 6 summarizes, draws conclusions, presents the limitations, and identifies future research.

## **2. Related Work**

Students' absenteeism is an important issue for schools all over the globe because of its negative impacts on overall student performance, especially their grades. It is difficult to find the exact reason (other than illness or other calamities) for most of the absences. Quite often, environmental factors are used to assume the pupils' absences—such clues can be extremely unreliable. The aim of this project was to conduct an empirical investigation to distill reality from myths about students' absenteeism.

### *2.1. Students' Attendance*

Students' attendance is an important area of focus for schools, local authorities, and national governments across the globe because of its convincing and positive association with students' attainment, their well-being, and enriched financial stability later in life (Balfanz and Byrnes (2012) [7], Cook et al., (2017) [8], Havik et al., (2015) [9]).

Several studies have emphasized the fact that class attendance is an important predictor for pupils' performance. According to Chen and Lin (2008) [10], class attendance has a significant positive effect on pupils' exam performances. Despite the accessibility of class notes (online or from other students), Nyamapfene's (2010) [11] study emphasized classroom attendance as it is directly associated with academic attainment.

An inverse relationship between performance and absenteeism has been suggested by Westerman et al., (2011) [12]. They revealed that absenteeism has a more negative impact on low performers whereas it has no significant effect on high performers in the class. Negative effects will reflect on students' performance if they do not attend class.

"Missing school matters", as suggested by Balfanz and Byrnes (2012) [7]; in the US, academic performance (irrespective of age) was impacted due to absences, especially for pupils coming from low-income backgrounds because of the inability (lack of essential provisions) to make up for the lost time at home.

In the UK, some absences, e.g., illness or bereavement, may be seen as acceptable. Though acceptable or authorized absences are approved by the schools, they are suspicious about its misuse, especially when these are around end of the term when parents want to avoid fines while trying to book cheaper holidays [2,13].

As opposed to authorized absences, absences without permission (unauthorized absences) are criticized by educational organizations and legislators. In line with national government policies regarding responsible attendance management, many parents have been fined and numerous cases have been taken to court [2,13].

Irrespective of the reason for absence, schools have become strict as authorized and unauthorized absences are equally damaging, destructive, and disruptive to learning. The Department for Education (DfE), UK, has strict policies regarding school attendance with legal responsibilities for both parents (including guardians) and schools. Parents and guardians are legally bound to send their children to school, maintain regular attendance, and communicate truthfully about any absence. Schools follow government policies and guidance and are legally obliged to take important steps for managing student attendance effectively. Schools need to record attendance accurately, be proactive in communication with parents about any absences, and take initiatives to encourage improved attendance [14].

Predicting class attendance is the key to planning better strategies to improve it. Artificial intelligence enables us to improve public and private life. It helps us discover hidden information in huge chunks of data in an automated way which is the key component of data science, and currently motivates its applications in various fields such as law, computational biology, finance, and many more. Bayesian networks are used for EDA (Educational Data Analysis) to gain a better understanding of students' multiplication misconceptions. Saranti et al., (2019) [15] has mentioned that, for highly unbalanced datasets, several algorithms can compute samples from the models with different characteristics. Their work can be a starting point for explaining AI-based decisions in attendance space because the number of present students is predominant in the attendance dataset, which makes it hugely unbalanced.

Artificial intelligence techniques have proved to have superhuman performance, but this performance surge has often been attained by enhanced complexity of AI techniques [6]. Despite having extremely positive results for improving lives, artificial intelligence is coupled with a substantial challenge, the challenge of understanding and hence trust. Machine learning models like decision trees or linear regression are compatible with human understanding whereas neural networks and deep learning models, though successful, they are complicated and machine learning models are not transparent but are like black-boxes to humans. The decisive intelligence of these methods is not clear from the outside.

### *2.2. Inductive Logic Programming and Induction in Automated Reasoning*

Muggleton [16] first introduced the term inductive logic programming (ILP) in 1991. This technique is very useful in bioinformatics and natural language processing. ILP was first implemented as a Prolog program that inductively inferred logic programs from positive and negative examples. The term "inductive" here refers to philosophical (i.e., suggesting a theory to explain observed facts) rather than mathematical (i.e., proving a property for all members of a well-ordered set) induction. The latter mathematical induction is the principle that is used as the machinery in Isabelle (and commonly in other automated reasoning tools) and that we use to define a state-transition relation over a set of states. We call the combination of states and transition relation a Kripke-structure in line with the terminology used in modal and temporal logics. Although the notion of "inductive" used in ILP and Isabelle differ, the PCR algorithm itself has some resemblance to the ILP approach because it also generalizes from negative examples to arrive at a general rule.

Finzel et al. (2022) [17] combined ILP with statistical machine learning methods, in particular Graphical Neural Networks (GNN). Their framework provides verbal explanations for relevant graph substructures as logical rules. By using logical rules for explanations, their approach for adding transparency resembles ours very much in principle, despite using fundamentally different approaches to inductive reasoning (as explained in the previous paragraph). In addition, according to [17], in "ILP systems, learning is usually performed by contrasting examples that belong to a target class with examples that do not belong to the target class (concept learning)". Here, the underlying duality of examples and counterexamples is also similar to our "attack" and precondition refinement approach.

## 3. eXplainable Artificial Intelligence (XAI)

> "Explainable Artificial Intelligence (XAI), a field that is concerned with the development of new methods that explain and interpret machine learning models." (Linardatos et al., (2021) [6]).

### *3.1. Explainability Need*

Explanation for some systems is crucial, e.g., self-driving cars, healthcare systems, or financial decisions. Who will be responsible for an accident if some objects are misinterpreted by self-driving cars resulting in a car crash? The car manufacturing company, software developer, or the passenger? Who will be responsible for a patient's health if healthcare systems for early disease detection misinterpret the healthy tissue as abnormal or vice-versa? A credit scoring system decides if the applicant is eligible for a loan or not. Understanding the reason behind the system's decision can help the loan applicants to obtain credit and become financially stable.

These concerns have inspired deep thoughtfulness within the community about human comprehensibility and generating explanations for decisions is a general issue in cognitive science, social science, and human psychology (Miller (2019) [18]). There is a need to turn these "black-boxes" into "white boxes"; if not entirely, then at least for AI to be explainable.

eXplainable AI is not a new topic, as cited in Schwalbe and Finzel (2023) [19] . In 2004, van Lent et al. [20] used the abbreviation XAI for the term eXplainable artificial intelligence. However, due to the introduction of GDPR rules, the transparency of complex AI approaches geared up.

According to the Defense Advanced Research Projects Agency (DARPA), one of the main goals of XAI is to convey a user-centric approach, i.e., to enable humans to understand their artificial counterparts [21].

In order to help the readers understand xAI terminology, we would like to add some basic XAI related definitions as cited in Saranti et al. (2019) [15].

- *Understanding*: Understanding is a mandatory precondition of explanation and is described as the human ability to understand the problem and to recognize correlations [22]. It can be divided into mechanistic ("How things work?") and functional ("What is the purpose?") understanding [23].
- *Explainability*: Explainability aims to describe the (i) reason, (ii) model, and (iii) decision of AI systems so that humans can understand them [22].
- *Transparency*: If the algorithmic behavior, processes, and decision output of an AI model can be understood by a human mechanistically then the model is said to be transparent [23].

Explainability

> "Explainability is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms. Explainability is being able to explain quite literally what is happening." (Koleñàk (2020) [24])

Chakarborti et al., (2019) [25] and Kulkarni et al., (2019) [26] have another "meta"-view of maintaining a clear model of the users about explainability. Similarly, Miller (2019) suggests that it is expected that causality plays a key role in explanation; however, according to Pearl (2018) [27], a thorough experimental framework and/or expertise is required for several models appearing in the causality literature. Over the recent years, these thoughts have led to tremendous scientific interest in the field of XAI (eXplainable Artificial Intelligence) (Linardatos et al., (2021) [6]).

Critical analysis of recent literature about eXplainable Artificial Intelligence (XAI) by Arrieta et al., (2020) [28] raises some challenges, mainly around the post-hoc explanation of black-box machine learning algorithms (e.g., CNN or Deep Learning) and their compatibility with human understanding. An impressive summary of XAI is provided by Gilpin et al., (2018) [29]. The work by Belle and Papantonis (2020) [30] provides a widespread survey of recent explanation techniques including comprehensible illustrations of their use in human-centric examples. Justification and transparency are the foremost benefits of explanation as suggested by Pieters (2011) [31]. According to Pieters (2011) [31], to visualize the relation between explanation goals and their subgoal, an explanation tree—"A tree in which the goals and subgoals of an explanation are ordered systematically"—may be used. Pieters' (2011) [31] work laid a strong foundation to our approach to explainability because attack trees are highly comparable with explanation trees. Attack trees are a risk analysis model used commonly for engineering secure systems [32]. The step-by-step process characterized as the "attack tree refinement" of an attack tree provides a transparent explanation of an attack [33] because the attack tree "explains" a higher level attack step by a sequence of more fine grained sub-attacks. Eventually, "the attack tree refinement" describes a fully grown tree of explanation which can be verified automatically by the model as is presented in the Isabelle Insider framework [33,34]. Therefore, as compared with the explanation tree, the attack, expressed by the parent node, can be "explained" by a subtree of an attack tree. The Isabelle Insider and Infrastructure framework (IIIf) provides explanations of black-box machine learning decisions to humans as it offers rich contexts for actors, infrastructures, and policies. This can be particularly relevant for privacy-critical applications.

Holzinger et al. (2019) [35] establishes the notion of "causability" as "the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use". The authors define "explainability" in a much more technical sense as merely highlighting "decision-relevant parts of the used representations of the algorithms and active parts in the algorithmic model, that either contribute to the model accuracy on the training set, or to a specific prediction for one particular observation". In their view, "explainability" "does not refer to an explicit human model" [35]. Thus, from the perspective of Holzinger et al. (2019) [35], our work lies more in the realm of causability because the use of the infrastructure models with actors and policies in IIIf addresses human aspects and causality, as well as "specified context(s) of use" for explanation.

### *3.2. Isabelle Insider and Infrastructure Framework (IIIf)*

The Isabelle Insider and Infrastructure framework (IIIf) has fully embedded attack trees as "first-class citizens" because of which the provision of a formal semantics for valid attacks based on Kripke structures is possible. Additionally, an efficient decision procedure can be derived using CTL temporal logic. The overall architecture of the IIIf is depicted in Figure 1, emphasizing its modular structure.

To decide the validity of attack trees, Scala (programming language) code is automatically generated from the Isabelle theories for central decision procedures. In the interactive generic theorem prover Isabelle/HOL, the IIIf is embedded as an extension of HOL [36]. With physical and logical components, actors, and policies, the formalization and proof of systems are backed by the Isabelle framework. IIIf's design is perfectly matched for the analysis of insider threats.

According to Kammüller (2022) [5], however, the implemented theory of the temporal logic CTL combined with Kripke structures and its generic notion of state transitions are a perfect match to be combined with attack trees into a process for formal security engineering CHIST-ERA (2019) [37] including an accompanying framework [34].

In our research, we have revealed that the explanation of black-box algorithm-driven decisions are possible using the Isabelle framework. A brief description of the main features of the Isabelle Infrastructure framework is provided here.

#### Attack Trees, CTL and Kripke Structures

For the state-based security analysis with actors and policies, the general framework of IIIf is shaped by numerous case studies. Isabelle's Higher Order Logic (HOL) is extremely rich with Kripke structures and CTL, which facilitates the proofs of meta-theoretical foundations. The general concept of state-transition is provided by this meta-theoretical foundation of IIIf and the properties of applications can be articulated with temporal logic and attack trees.

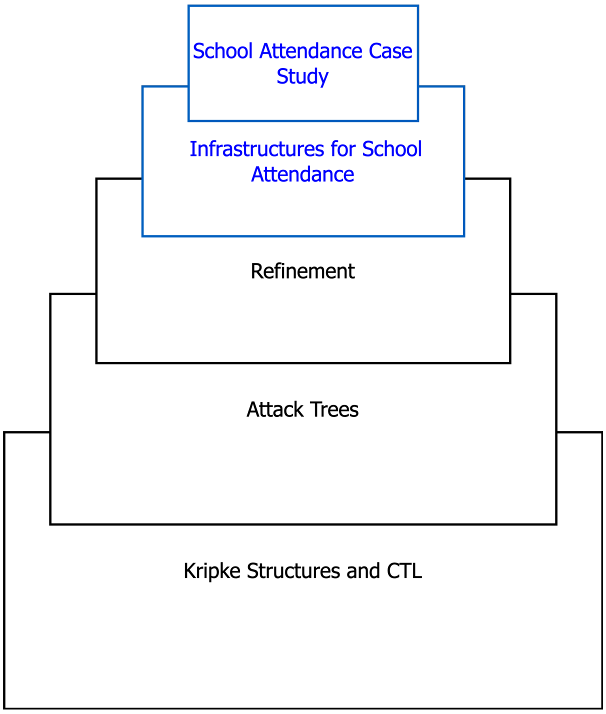
School Attendance Case Study
Infrastructures for School Attendance
Refinement
Attack Trees
Kripke Structures and CTL

**Figure 1.** Isabelle Insider and Infrastructure framework (IIIf).

For a wide range of application modeling, the logical notions and associated concepts are:

- Kripke structures and state transitions:
  A generic state transition relation is $\to_i$; Kripke structures over a set of states $\mathtt{t}$ reachable by $\to_i$ from an initial state set $\mathtt{I}$ can be constructed by the $\mathtt{Kripke}$ constructor as

  $$\mathtt{Kripke}\ \{\mathtt{t}.\ \exists\ \mathtt{i} \in \mathtt{I}.\ \mathtt{i} \to^{*} \mathtt{t}\}\ \mathtt{I}$$

- CTL statements:
  Computation Tree Logic (CTL) can be used to specify dependability properties as

  $$\mathtt{K} \vdash \ \mathsf{EF}\ \mathtt{s},$$

  which means that in Kripke structure $\mathtt{K}$ there is a path ($\mathsf{E}$) upon which the property $\mathtt{s}$ (given as the set of states in which the property is true) will eventually ($\mathsf{F}$) hold.
- Attack trees:
  In Isabelle, attack trees are defined as a recursive datatype having three constructors: $\oplus_{\lor}$ creates or-trees and $\oplus_{\land}$ creates and-trees. And-attack trees $l \oplus_{\land}^{s}$ and or-attack trees $l \oplus_{\lor}^{s}$ consist of a list of sub-attacks which are again recursively given as attack trees. The third constructor inputs a pair of state sets and constructs a base attack step between two state sets. As an illustration, for the sets $\mathtt{I}$ and $\mathtt{s}$ this is written as $\mathcal{N}_{(\mathtt{I},\mathtt{s})}$. To give another example, a two step and-attack leading from state set $\mathtt{I}$ via $\mathtt{si}$ to $\mathtt{s}$ is expressed as

  $$\vdash\ [\mathcal{N}_{(\mathtt{I},\mathtt{si})},\mathcal{N}_{(\mathtt{si},\mathtt{s})}]\oplus_{\land}^{(\mathtt{I},\mathtt{s})}.$$

- Attack tree refinement, validity and adequacy:
  Refinement can also be used to construct attack trees but this process is different from the system refinement described in Kammüller (2021) [36]. A high level attack tree is refined iteratively by adding more detailed attack steps until a valid attack is reached: $\vdash \mathtt{A}\ ::\ \ (\sigma::\ \ \mathtt{state})\ \mathtt{attree})$.
  The definition of validity is constructive so that code can be automatically extracted from it. A formal semantics for attack trees is provided in CTL; adequacy is proved which can enable actual application verification.

As mentioned in Kammüller (2022) [5], several studies suggest that IIIf is highly efficient in explaining automated ML decisions. In this study, we have presented a reasonable methodology to implement explanation within the illustrative IIIf.

## 4. Case Study of Pupils' Attendance in Schools

In this section, we give a brief introduction to pupils' attendance and its relevant factors to motivate the case study that has been used to illustrate how IIIf is applied to it to provide a basis for explanation. Prediction of school attendance is the possibility/probability assigned to pupils to quantify/predict their "attendance". These predictions are used by learning providers as well as local authorities to decide whether a pupil may attend the school or not. It also influences the grades the pupil may receive, which can lead to disadvantaged underperforming pupils. An open question is: How can such attendance predictions be created as they rely on private data? High grades do not automatically lead to a higher attendance rate. It seems rather likely that AI-based decision making procedures are applied within the education department. In order to clarify such opaque relations, logical modeling can help as it remodels the actual context of the original data collection and thus may show up any biases used.

### *4.1. Problem Statement*

This project addressed the limited transparency of black-box AI systems by investigating the effectiveness of eXplainable Artificial Intelligence (XAI) for improving student attendance rate using the IIIf.

### *4.2. Data Collection*

This project was based around students' personal information. To ensure data privacy and security, the data points have been modified to represent the information instead of using the real data. In the next section, we present a simple example to illustrate how the attendance prediction scenario can be represented as an infrastructure model in IIIf.

### *4.3. Model in IIIf*

The IIIf supports the representation of infrastructures as graphs with actors and policies attached to nodes. These infrastructures are the states of a Kripke structure describing the attendance prediction scenario. The behavior is defined by a transition relation on states. This transition between states is triggered by non-parameterized actions `put`, `eval`, `move`, and `delete` executed by actors. Actors are given by an abstract type `actor` and a function `Actor` that creates elements of that type from identities (of type `string` written `''s''` in Isabelle). Actors reside at locations of an infrastructure graph of type `igraph` constructed by its constructor `Lgraph`.

```
datatype igraph = Lgraph
                    (gra: <(location × location)set>)
                    (agra: <location ⇒ identity set>)
                    (dgra: <identity ⇒ dlm × data>)
                    (bb: <data ⇒ bool>)
                    (attendance: <(identity × bool option)set>)

datatype infrastructure =
         Infrastructure (graphI: <igraph>)
                        (delta: <[igraph, location] ⟶ policy set>)
```

For the current application of the attendance prediction scenario, this graph contains the actual location graph of type `(location × location) set` given by a set of location pairs and a function of type `location ⇒identity set` that assigns the actors to their current location. The third component of the datatype `igraph` is of type `identity ⟶ (dlm ×data)`. It assigns security labeled data to actors. The label type is called `dlm`, as a reference to the decentralized label model by Myers and Liskov (1999) [4] who inspired it. It is a pair of type actor × actor set defining the owner and the readers of a data item. The type data contains the private data of users. For our example, we have used the location (UK regions: Coastal, NonCoastal and London), the disadvantage flags (Free School Meal (fsm), education and Health care (ehc), Special Educational Needs (sen), and Child in Social Care (csc)), gender, year group (Phase of education: primary, secondary), season (inferred from the date/time of year), transport, and ethnicity. Note that the field for disadvantages is a set because more than one disadvantage flag could apply to an actor.

```
data = <location×disadvantage set×gender×year×season×transport×ethnicity>
```

The fourth component bb of the igraph data type is of type data ⟶ bool. It is the black-box function: effectively a table that contains the attendance prediction for given data inputs. The final component of type (identity × bool option) set records that a learning provider has requested a pupil's attendance prediction by uploading their identity together with a Boolean field that contains the future decision of the attendance prediction to the set of requests. The second Boolean component containing the answer is lifted by the option type constructor that enables an undefined value None to flag that there has not been any response yet. Each of the components of the type constructor is equipped with a corresponding projection function that allows access to this component in an instance of this type constructor (an element of this type). These projection functions are named gra for the set of pairs of location representing the infrastructure graph, agra for the assignment of actors to locations, dgra for the data at that location, bb for the black-box, and attendance for the pairs of attendance and prediction of actors of requesting attendance prediction. We omitted some standard constructions for infrastructure assembly and the policy definition from local policies. A generic state transition relation over Kripke structures is defined together with logic and decision procedures for IIIf. This is then instantiated to concrete applications of the IIIf—like in the current attendance prediction example—by defining the rules for the state transition relation over a defined infrastructure type, as given by the above igraph. This state transition relation then implements the behavior for attendance prediction systems by explaining how actions executed by actors change the infrastructure state. The execution of actions is conditional on enabledness as defined by the local policies and other conditions of the context. For attendance prediction systems, we considered here the actions put, representing that a learning provider requests an attendance prediction and eval, where an entitled client (presumably an education department) executes a requested attendance application. In the precondition of the rule for a put action, the actor a residing at location l in the infrastructure graph G (given by the predicate Actor a @$_G$l), who is enabled to put an attendance request, uploads their data to the attendance G field into the infrastructure graph G. A potential (education department) Actor c can see a new request since now there is a new pair (a, None) in the attendance request set where the second component of this pair is flagged by the None constructor of the option type as "unprocessed" while the first element is the requesting actor's identity a. The creation of the new element is encoded in the function put_graph_a (omitted here, for details see the source code [38])

```
put : G = graphI I ⟹ a @_G l ⟹ l ∈ nodes G ⟹
          enables I l (Actor a) put ⟹
          I' = Infrastructure (put_graph_a a l G) (delta I) ⟹
          I →_n I'
```

The action eval allows the evaluation of a request filed by actor a by an (presumable) education department c given that c is contained in the readers set of the dlm label lb that is given as the second element of the first element of the data item dgra G a. Also, c needs to be enabled to evaluate attendance requests by the local policy. Given these prerequisites, the actual evaluation is performed by applying the black-box function bb G to the data item d and recording the outcome in the second component of the corresponding pair for a in the attendance set (this technical step is formalized in the function eval_graph_a (again see source code [38])).

```
eval: G = graphI I ⟹ a @_G l ⟹ l ∈ nodes G ⟹
        c ∈ actors_graph G ⟹ (a, None) ∈ attendance G ⟹
        Actor c ∈ readers (dgra G a) ∨ Actor c = owner (dgra G a) ⟹
        enables I l (Actor c) eval ⟶
        I' = Infrastructure (eval_graph_a a G) (delta I) ⟹
        I →_n I'
```

We omitted the state transition rules for the actions `delete` and `move` (see source code [38]). They will be illustrated in the evaluation of the example below. The above infrastructure Kripke model for attendance prediction formalizes attendance prediction scenarios, enabling reasoning in general about all instances. To simulate concrete example scenarios, we can use the generic nature of the IIIf with its polymorphic Kripke structure and state transition. Defining a locale [39] named `SchoolAttendance` allows the fixing of some concrete values for the actors, locations, and local policies and inherits all general definitions and properties of infrastructures from the framework. For simplicity, we considered just three actors—`Alice`, `Bob`, and `Charlie`—and an Education Department `ED`.

```
locale SchoolAttendance =
fixes SchoolAttendance_actors ::<identity set>
defines SchoolAttendance_actors_def:
          <SchoolAttendance_actors ≡ {‘‘Alice’’,‘‘Bob’’,‘‘ED’’}>
```

The locale allows us to initialize a concrete `igraph` with these and other values. Moreover, it serves to illustrate the explanation process that we are going to present next.

## 5. Results

In this section, we define a Precondition Refinement Process (PCR Cycle) which is a cyclic method to drive a general logical characterization of what the black-box mechanism decides within any given state of the world. A possible world is described in the IIIf as a Kripke state comprising actors, policies, and infrastructures including any features necessary to specify the context of a human centric scenario. After defining the process, we continue by illustrating its use on the previously introduced school attendance system.

### *5.1. Definition of PCR Cycle*

The Pre-Condition Refinement process (PCR cycle) is a recurring technique, within any given state of the world, which derives a general logical characterization of the decisions made by black-box algorithms. In the IIIf, a Kripke state is described as a possible world which comprises policies, actors, and infrastructures including any essential characteristics to specify the perspective of a human centric state. As compared to the RR-cycle [36], to find the "failure states" we will use attack trees. Failure states can be considered states in which a desired outcome is not given. However, instead of refining system specifications with the use of dynamic behavior of an infrastructure system we will refine the precondition of an explanation. The refinement of the overall preconditions is guided by the general rules of the preconditions and will be iterated until we obtain a general rule for the explanation. This iterative cycle of the precondition refinement will begin with an "attack tree" to prove the temporal property showing that the "failure" states, that do not fulfill the desired outcome, can also be reached. As an additional precondition, "failure states" can be used to propose an alternative route to reach the desired outcome state. The desired outcome can be used to define the termination rule of the PCR cycle. The iterative precondition refinement process derives a broad-spectrum explanation rule which provides a detailed logical description of the path to achieve the desired outcome property, which is a positive classification given by the black-box AI algorithm. The "desired outcome (`DO`)" of the PCR cycle can be compared with the "global policy" of the RR-cycle [36]. The "failure state" can be used to define a *counterfactual* that would have provided an alternative path to a state fulfilling the `DO` property. Besides helping to guide the refinement by an additional precondition, the `DO` also provides the termination condition of the cyclic precondition refinement process. Since the `DO` property is the positive classification of the AI algorithm given as a black-box, the process yields a general explanation of rule that gives a precise logical description how the DO property can be achieved.

This is the brief description of the PCR cycle. We have reconsidered Kammüller (2022) [5] definition of the PCR cycle for this paper. In what follows we provide its high level yet detailed algorithmic description including the formal definitions of the core concepts used. However, before we come to that we need to introduce how we formalize counterfactuals which are the driving concept to refine the precondition in the PCR cycle.

**Counterfactuals**: A counterfactual is best explained by example. We give one that fits into the context of our case study: "if the student has no disadvantages, he would have got a "present" attendance prediction. Intuitively, counterfactuals try to illustrate facts in the current state of the world by showing alternative hypothetical developments of the world that feature the opposite case of the fact. It is not a coincidence that our explicit world model of Kripke structures and state transitions lends itself so naturally to modeling counterfactuals.

However, apart from modeling the different possible worlds and their evolution, we also need a metric for them. As described in Kammüller (2022) [5], the concept of the "closest possible world" or the smallest change to the world that can be made to obtain a desirable outcome, is the key to counterfactuals [40]. We use the step-relation between possible states (worlds) to define a unique notion of "closest" between three states. Intuitively, it formalises the closest predecessor s of the two possible states s' and s'' by stipulating that any other state s0 that is also a predecessor (with respect to $\to^*$) to states s' and s'' must already be a predecessor to s.

$$\begin{aligned}\texttt{closest s s' s''} \equiv\ & \texttt{s} \to^* \texttt{s'} \land \texttt{s} \to^* \texttt{s''} \land \\ & \forall\, \texttt{s}_0.\ \texttt{s}_0 \to^* \texttt{s'} \land \texttt{s}_0 \to^* \texttt{s''} \Rightarrow \texttt{s}_0 \to^* \texttt{s}\end{aligned}$$

This definition is used for defining an additional precondition with respect to desirable outcome DO by simply stating that for a state s with $\neg$DO s there must be an alternative trace leading to another possible world s'' with DO s'' such that they are connected by a closest state s'. Using the definition of closest state, we can define this simply as the set of states for which such a closest predecessor exists.

**Definition 1** (Counterfactuals)**.** *Counterfactuals for a state* s *with respect to a desirable property* DO *are the states* s'' *that fulfill* DO *and have a closest predecessor* s'*.*

$$\begin{aligned}\texttt{counterfactuals s DO} \equiv \{\texttt{s''}.\ & \texttt{DO s''} \land \\ & (\exists\, \texttt{s'}.\ (\texttt{s'} \to^* \texttt{s''}) \land \texttt{closest s' s s''})\}\end{aligned}$$

Our definition of counterfactuals overlaps with the state-of-the-art definition used in the literature but is more abstract. Whereas classical definitions are using a concrete metric to define a closest possible world, our definition uses the state-transition relation to define the notion of closest predecessor state which in turn gives rise to define a set of counterfactuals. This is a set of states in which DO is given and that are equally close to the original state via the unique predecessor. The definition gives *a set* of such possible states because in general there are a number of states that are reachable via the closest predecessor and which all have DO. Our notion of closest is an abstract metric to define all possible counterfactuals. It can be used in implementations to select a specific counterfactual by combining it with more specific metrics, for example, the metric "distance of locations" in our case study.

We will see the application of the concepts of counterfactuals and closest state in the following algorithm.

### PCR Cycle Algorithm

1. Using attack tree analysis in the CTL logic of the IIIf we *find the initial starting condition of the PCR*. The variable B is an element of a datatype for which we seek explanation (in the example it is actors).

$$\texttt{M} \vdash \ \mathsf{EF}\ \{\ \texttt{s} \in \texttt{states M.} \neg\ \texttt{DO(B, s)}\ \}.$$

This formula states that there exists a path (E) on which eventually (F) a state `s` will be reached in which the desirable outcome is not true for `B`. The path corresponds to an attack tree (by adequacy [33]) designating failure states `s`.

2. *Find the (initial or refined) precondition using a counterfactual.*
   That is, for a state `s` in the set of failure states identified in the previous step
   (a) Find states `s'` and `s''` such that `closest s' s s''`, that is, `s'` $\rightarrow^*$ `s` and `s'` $\rightarrow^*$ `s''`. In addition, `DO(B,s'')` must hold.
   (b) Identify the precondition $\texttt{pc}_i$ leading to the state `s''` where `DO` holds, that is, find an additional predicate $\Delta_j$ with $\Delta_j$`(B, s')` and use it to extend the previous predicate $\texttt{pc}_i$ to $\texttt{pc}_{i+1}$ `:=` $\texttt{pc}_i \wedge \Delta_j$.
3. *Generalisation.*
   Use again attack tree analysis in the CTL logic of the IIIf to check whether the following formula is true on the entire datatype of `B`: it is globally true (AG) that if the precondition $\texttt{pc}_i$ holds, there is a path on which eventually (EF) the desirable outcome `DO` holds (Note that the interleaving of the CTL-operators AG and EF with logical operators, like implication $\longrightarrow$ is only possible since we use a Higher Order logic embedding of CTL.)

$$\forall\ \texttt{A. M} \vdash \ \mathsf{AG}\ \{\texttt{s} \in \texttt{states M. pc}_i\ \texttt{(A, s)} \longrightarrow \ \mathsf{EF}\ \{\texttt{s. DO(A, s)}\}\}$$

   (a) If the check is negative, we get an attack tree, that is, IIIf provides an explanation tree for

$$\texttt{M} \vdash \ \mathsf{EF}\ \{\ \texttt{s} \in \texttt{states M. pc}_i\texttt{(A,s)} \wedge \neg\ \texttt{DO(A, s)}\ \}$$

   and a set of failure states `s` with $\texttt{pc}_i$`(A,s)` and the desirable outcome is not true: ¬`DO(A,s)`.
   In this case, *go to step 2. and repeat* with the new set of failure states in order to find new counterfactuals and refine the predicate. $\texttt{pc}_{i+1}$ `:=` $\texttt{pc}_i \vee \Delta_j$ where $\Delta_j$ is an additional precondition.
   (b) If the check is positive, we have *reached the termination condition* yielding a precondition $\texttt{pc}_n$ such that for all `A`:

$$\texttt{M} \vdash \ \mathsf{AG}\ \{\ \texttt{s} \in \texttt{states M. pc}_n\ \texttt{(A, s)} \longrightarrow \ \mathsf{EF}\ \{\texttt{s. DO(A, s)}\}\ \}$$

Note that the analysis in Step 3 might potentially reveal a new variable as part of $\Delta_i$ over another datatype (locations in the example). This is not a problem as it will eventually lead to tease out the entire set of parameters that the black-box decision procedure uses. We did not attempt to formalise it explicitly into the above algorithm description to keep the exposition easier understandable.

### *5.2. PCR Cycle Application to School Attendance Case Study*

We now demonstrate the PCR algorithm in our case study introduced in the previous section. We considered the scenario in which Bob receives an evaluation by the Education Department (ED) and it is ‘Absent’. Bob wants to understand why this is the case and requests (attendance) an explanation. The experts in the Education Department cannot give this explanation as they used a black-box machine learning system bb. Now, the IIIf and the PCR algorithm can be used by modeling the scenario and using a `bb` system as a black-box, that is, requesting only its classification output (verdict) for any given inputs. (It is important to note that we request (attendance) really only input output pairs and not a mathematical description of the black-box. This is in contrast to the stronger assumptions made in the literature, for example in Wachter et al., (2018) [40])). The desired outcome `DO` in an infrastructure state `s` is given by the pair that has a field having a `True` as a second component (lifted by `Some`).

```
DO :: <identity ⇒ infrastructure ⇒ bool>
DO(a,s) ≡ (a, Some(True)) ∈ attendance s
```

We show the run of the algorithm by going through its steps 1–3 for the application additionally ornating the numbers with $\alpha, \beta, \ldots$ to indicate the round of the algorithm.

$\alpha$.1 For actor Bob, we use CTL modelchecking in the IIIf to verify the formula

```
Attendance_Kripke ⊢
      EF { s ∈ states Attendance_Kripke. ¬ DO(‘‘Bob’’, s) }.
```

From this proof, the IIIf allows applying Completeness and Correctness results of CTL [33] to derive the following attack tree.

$\vdash \; [\mathcal{N}_{(\mathtt{I,C})}, \mathcal{N}_{(\mathtt{C,CC})}] \oplus_{\wedge}^{(\mathtt{I,CC})}$

The attack tree corresponds to a path leading from the initial state `I` to the failure state `CC` where Bob's approval field in `attendance CC` gets evaluated by the education department `ED` as negative "False". The evaluation steps are:

`I`→`C`: Bob puts in an attendance request; this is represented by a put action. So, the state `C` has `(‘Bob’’, None) ∈ attendance C`.

`C`→`CC`the Education Department `ED` evaluates the attendance request represented as an `eval` action with the result of the evaluation left in `attendance CC`. So, the state `CC` has `(‘Bob’’, Some(False)) ∈ attendance CC`.

To derive the final failure state `CC`, the Education Department has applied the `bb` function as `Some((bb C) d)` which evaluates Bob's request as `Some(False)` (rule `eval`).

$\alpha$.2 Next, the PCR algorithm finds an initial precondition that yields the desirable outcome in a closest state using counterfactuals. The closest state is given as `Ca` which differs from `C` in that Bob has lower disadvantage set (0 elements) as opposed to a 1-element set as in `C`. The precondition derived is

$$\mathtt{pc}_1 \equiv \mathtt{card(disadvantage\_set\ A\ s)} = 1 \wedge \mathtt{A} \;\; @_s \; \mathtt{Non\text{-}Coastal)}$$

The state `Ca` is reachable: Bob first applies for a disadvantage removal via the action `delete`. From the state `Ca`, Bob puts in the attendance application leading to `CCa`, before the Education Department `ED` evaluates leading to `CCCa`. We see that now with the reduced disadvantage set, Bob receives a present attendance prediction.

$\alpha$.3 The next step of the PCR algorithm is generalisation. We want to investigate whether the disadvantage set reduction is a sufficient precondition in general (for all actors) to explain why the `bb` algorithm approves the credit. When we try to prove according to Step 3 that this is the case, the attack tree analysis proves the opposite.

$$\mathtt{M} \vdash \; \mathrm{EF}\ \{\ \mathtt{s} \in \mathtt{states\ M.}\ \ \mathtt{s}_0\mathtt{(‘‘Alice’’,\ s)} \wedge \neg\ \mathtt{DO(‘‘Alice’’,\ s)}\ \}$$

It turns out that Alice who has a disadvantage set of size 2 does not get the "present" attendance scoring either. She lives, however, in the coastal area unlike Bob who lives in the non-coastal area. Following Step 3(a) we need to go to another iteration and go back to Step 2, to refine the precondition.

$\beta$.2 In this $\beta$-run, we now have the state `s` where Alice does not get the approval. According to Step 2(a), we find a counterfactual state as the one in which Alice reduces her disability set to one leading to a new alternative precondition added to the previous one as an alternative (with $\vee$).

$$\mathtt{pc}_2\ \mathtt{(A,\ s)} := \mathtt{card(disadvantage\_set\ A\ s)} \leq 1 \wedge \mathtt{A}\ @_s\ \mathtt{Non\text{-}Coastal}$$

$\gamma$.2 However, there are more alternatives in the counterfactual set. Alice can also first move to London. The new precondition now is created by adding the following $\mathtt{pc}_3$ as an additional alternative with $\vee$ to the overall precondition.

$$\texttt{pc}_3\ \texttt{(A, s) := card(disadvantage\_set A s)} \leq 2 \wedge \texttt{A} \ @_s\ \texttt{London}$$

$\gamma$.3 Going to Step 3 again in this $\gamma$-run, now the proof of the generalisation succeeds.

$$\forall\ \texttt{A. M} \vdash\ \texttt{AG \{s} \in \texttt{states M. pc}_1\texttt{(A, s)} \vee \texttt{pc}_2\ \texttt{(A, s)} \vee \texttt{pc}_3\ \texttt{(A, s)} \longrightarrow\ \texttt{EF \{s. DO(A, s)\}\}}$$

### 5.3. Discussion

With respect to the explanation, the algorithm finishes with the precondition

$$\begin{aligned}\texttt{pc}_n\ \texttt{(A, s) :=}\ & \texttt{card(disadvantage\_set A s) = 0} \wedge \texttt{A}\ @_s\ \texttt{Coastal} \vee \\ & \texttt{card(disadvantage\_set A s)} \leq \texttt{1} \wedge \texttt{A}\ @_s\ \texttt{NonCoastal} \vee \\ & \texttt{card(disadvantage\_set A s)} \leq \texttt{2} \wedge \texttt{A}\ @_s\ \texttt{London}\end{aligned}$$

for any A of type actor. Effectively, for each of the three partial preconditions $\texttt{pc}_1$, $\texttt{pc}_2$ and $\texttt{pc}_3$, the goal DO (A, s) is reachable (EF). In terms of CTL logic, the PCR algorithm combines each of these single implications into the overall result

$$\forall\ \texttt{A. M} \vdash\ \texttt{AG \{s} \in \texttt{states M. pc}_n\ \texttt{(A, s)} \longrightarrow\ \texttt{EF \{s. DO(A, s)\}\}}$$

This combination of single precondition-implications into one implication with all pre-condition accumulated into a disjunction is a result of Step 3 of generalisation of the PCR-cycle. A great advantage of the approach using IIIf and hence Isabelle is that such meta-theoretical steps can be fully verified. The following theorem expresses this summarizing exactly the described accumulation step.

$$\begin{aligned}&\texttt{M} \vdash \texttt{AG \{s. P1 s} \longrightarrow \texttt{s} \in \texttt{S\}} \quad \Longrightarrow \\ &\texttt{M} \vdash \texttt{AG \{s. P2 s} \longrightarrow \texttt{s} \in \texttt{S\}} \quad \Longrightarrow \\ &\texttt{M} \vdash \texttt{AG \{s. P3 s} \longrightarrow \texttt{s} \in \texttt{S\}} \quad \Longrightarrow \\ &\texttt{M} \vdash \texttt{AG \{s. P1 s} \vee \texttt{P2 s} \vee \texttt{P3 s} \longrightarrow \texttt{s} \in \texttt{S\}}\end{aligned}$$

This theorem is proved in the IIIf and can be immediately applied as part of Step 3 of the PCR algorithm. Its proof and application present an additional refinement of the PCR-algorithm compared to the earlier publication [5] that emerged through the present case study.

One might ask whether the re-engineering we apply in our method affects the prediction performance. However, we re-engineer the black-box function as a logical rule with preconditions which does not affect the original black-box function nor its prediction performance. Additionally, if the PCR-algorithm terminates with a complete exploration of the state space, that is, identifying all possible preconditions, then the re-engineered logical rule is equivalent to the black-box function and thus also has the same predictions but improved transparency.

It turns out that often there is a bias in the data that has been used to train the black-box algorithm. For our case study, we deliberately used such an example to show its potential use for the logical explanation we provide. Since we give a general rule that formally describes and explains the decision process based on actual features of the context of the world. Here, the full run of the PCR algorithm would reveal that for different UK locations where students with more disadvantages lives, well established commutations system (like London) is needed for students to be present in the school. While our example is synthesized (based on the real data points), biases like these are known to be implicit in data sets because of the data workers who provided the training data classifications. A bias may be built into the black-box classification function by the data-workers who manually value decisions on the training data set for the black-box function. If they are themselves biased their decisions influence the valuation. This bias is revealed by the explanation. A very important contribution of our explicit logical model is thus to reveal such biases that are implicit in black-box AI algorithms for data evaluation.

As a general remark on limitations, the samples collected from the target population for the XAI Educational data were limited to student's personal parameters and did not include activity or prior performance parameters.

## 6. Conclusions

In this paper, we have shown how the RR-cycle of the IIIf can be adapted to provide a method for iteratively extracting an explanation by interleaving attack tree analysis with precondition refinement. This precondition refinement (PCR) cycle finally yields a general rule that describes the decision taken by a black-box algorithm produced by AI. Since it is a logical rule within a rich context of an infrastructure model of the application scenario, it provides transparency and explanation. The purposes of the right of explanation for an AI decision procedure are supported by the PCR cycle. The PCR cycle only needs to be slightly adapted to the RR-cycle by implementing an algorithm to define a methodology for interleaving attack tree analysis with a step-by step refinement of a precondition using additional preconditions. Responsible for the ease of this adaptation is the first-class representation of attack trees in the IIIf. That is, the existing correctness and completeness results of attack trees with respect to the CTL logic defined over Kripke structures allows changing between attack trees and CTL EF formulas. Thus, attack trees can be reused as explanation trees because they explain how failure states are reached. This in turn allows for the construction of additional preconditions that guide the refinement of the precondition. This project has validated the algorithm of the PCR cycle by a case study of students' attendance. Further work should address to what extent finding the additional preconditions and thereby the refined precondition can be automated. This study has illustrated that IIIf's RR-cycle can be adapted to specify a method for generalizing an "Explanation" using attack tree analysis with cyclic precondition refinement, with which we can derive a general rule that defines the AI decisions produced by a black-box algorithm of machine learning.

### *Future Work*

As discussed at the end of Section 5.1, our Definition 1 of counterfactuals is abstract. It identifies a set of counterfactual states for a given desirable outcome `DO`. Although this definition is mathematically precise it leaves a number of potential candidates for refining the precondition. This is again not a problem for the definition of the algorithm because we can iterate over all the possible states step by step each time refining the precondition. However, for a practical implementation, we would propose to look for a deterministic implementation as a future avenue of research. This can be done in Isabelle by providing a constructive function as a refinement of the definition, that is, it iterates over the set of counterfactuals proving that it conforms to Definition 1 and reaches all the elements of the counterfactual set. Such a constructive function definition then also allows code generation into programming languages like Scala.

The PCR cycle is mathematically precise and allows to generate an abstract definition as a logical rule for explanation of black-box AI functions. Although the current paper represents a proof of concept for our methodology by presenting a case study of applying the PCR cycle, we use the working hypothesis that descriptions of a black-box function as a logical rule are transparent for humans. Although in many other works, for example [17] similar working hypotheses are used, clearly this point can only be proved by empirical research with user groups and systematic evaluation of experiments. This is beyond our study but a necessary step towards fully exploring and validating our approach.

**Author Contributions:** Conceptualization, F.K.; software, F.K.; validation, F.K. and D.S.; formal analysis, F.K.; investigation, D.S.; resources, D.S.; data curation. D.S., writing—original draft preparation, D.S. and F.K., writing—review and editing, F.K.; visualization, F.K. and D.S.; supervision, F.K.; project administration, F.K.; funding acquisition, D.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The Isabelle resources are available at [38].

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Friedman, P.; Rodriguez, F.; McComb, J. Why Students Do and Do Not Attend Classes. *Coll. Teach.* **2014**, *49*, 124–133. [CrossRef]
2. Moodley, R.; Chiclana, F.; Carter, J.; Caraffini, F. Using Data Mining in Educational Adminstration: A Case Study on Improving School Attendance. *Appl. Sci.* **2020**, *10*, 3116. [CrossRef]
3. Vissers, M. Predicting Students' Class Attendance. Master's Thesis, Tilburg University School of Humanities, Tilburg, The Netherlands, 2018. Available online: http://arno.uvt.nl/show.cgi?fid=147795 (accessed on 17 July 2023).
4. Myers, A.C.; Liskov, B. Complete, Safe Information Flow with Decentralized Labels. In Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, USA, 6 May 1998; IEEE: Piscataway, NJ, USA, 1999.
5. Kammüller, F. Explanation of Black Box AI for GDPR related Privacy using Isabelle. In Proceedings of the Data Privacy Management DPM '22, Co-Located with ESORICS 22, Copenhagen, Denmark, 26–30 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; Volume 13619.
6. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef] [PubMed]
7. Balfanz, R.; Byrnes, V. The Importance of Being There: A Report on Absenteeism in the Nation's Public Schools. 2012. Available online: https://www.attendanceworks.org/wp-content/uploads/2017/06/FINALChronicAbsenteeismReport_May16.pdf (accessed on 13 July 2023).
8. Cook, P.; Dodge, K.; Gifford, E.; Schulting, A. A new program to prevent primary school absenteeism: Results of a pilot study in five schools. *Child. Youth Serv. Rev.* **2017**, *82*, 262–270. [CrossRef]
9. Havik, T.; Ingul, J.M. How to Understand School Refusal. *Front. Educ.* **2021**, *6*. . [CrossRef]
10. Chen, J.; Lin, T.F. Class Attendance and Exam Performance: A Randomized Experiment. *J. Econ. Educ.* **2008**, *39*, 213–227. [CrossRef]
11. Nyamapfene, A. Does class attendance still matter? *Eng. Educ.* **2010**, *5*, 67–74. [CrossRef]
12. Westerman, J.W.; Perez-Batres, L.A.; Coffey, B.S.; Pouder, R.W. The relationship between undergraduate attendance and performance revisited: Alignment of student and instructor goals. *Decis. Sci. J. Innov. Educ.* **2011**, *9*, 49–67. [CrossRef]
13. The Department for Education. Can You Take Kids on Term-Time Holidays without Being Fined? 2023. Available online: https://www.moneysavingexpert.com/family/school-holiday-fines/ (accessed on 17 July 2023).
14. GOV.UK. Department for Education. 2023. Available online: https://www.gov.uk/government/organisations/department-for-education (accessed on 17 July 2023).
15. Saranti, A.; Taraghi, B.; Ebner, M.; Holzinger, A. Insights into Learning Competence through Probabilistic Graphical Models. In *Machine Learning and Knowledge Extraction, Proceedings of the 2019 International Cross-Domain Conference, CD-MAKE 2019, Canterbury, UK, 26–29 August 2019*; Lecture Notes in Computer Science; Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E., Eds.; Springer: Cham, Switzerland, 2019; pp. 250–271. [CrossRef]
16. Muggleton, S.H. Inductive logic programming. *New Gener. Comput.* **1991**, *8*, 295–318. [CrossRef]
17. Finzel, B.; Saranti, A.; Angerschmid, A.; Tafler, D.; Pfeifer, B.; Holzinger, A. Generating Explanations for Conceptual Validation of Graph Neural Networks. *KI—Künstliche Intelligenz* **2022**, *36*, 271–285. [CrossRef]
18. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
19. Schwalbe, G.; Finzel, B. XAI Method Properties: A (Meta-)Study. Available online: http://arxiv.org/abs/2105.07190 (accessed on 2 August 2023).
20. van Lent, M.; Fisher, W.; Mancuso, M. An Explainable Artificial Intelligence System for Small-Unit Tactical Behavior. In Proceedings of the IAAI'04—16th Conference on Innovative Applications of Artifical Intelligence, San Jose, CA, USA, 27–29 July 2004; pp. 900–907.
21. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
22. Bruckert, S.; Finzel, B.; Schmid, U. The Next Generation of Medical Decision Support: A Roadmap toward Transparent Expert Companions. *Front. Artif. Intell.* **2020**, *3*, 507973. [CrossRef] [PubMed]
23. Páez, A. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds Mach.* **2019**, *29*, 441–459. [CrossRef]
24. Koleňák, F. Explainable Artificial Intelligence. Master's Thesis, Department of Computer Science and Engineering, University of West Bohemia, Plzeň, Czech Republic, 2020.
25. Chakraborti, T.; Sreedharan, S.; Grover, S.; Kambhampati, S. Plan Explanations as Model Reconciliation: An Empirical Study. In Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Republic of Korea, 11–14 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 258–266.
26. Kulkarni, A.; Zha, Y.; Chakraborti, T.; Vadlamudi, S.G.; Zhang, Y.; Kambhampati, S. Explicable Planning as Minimizing Distance from Expected Behavior. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, Montreal, QC, Canada, 13–17 May 2019; International Foundation for Autonomous Agents and Multiagent Systems: Pullman, WA, USA, 2019; pp. 2075–2077. Available online: https://www.ifaamas.org (accessed on 2 August 2023).
27. Pearl, J. Theoretical Impediments to Machine Learning with Seven sparks from the Causal Revolution. *arXiv* **2018**, arXiv:1801.04016.
28. Arrieta, A.B.; Díaz-Rodríguez, N.; Javier Del Ser, A.B.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molinaa, D.; Benjamins, R.; Chatila, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]

29. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.A.; Kagal, L. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv* **2018**, arXiv:1806.00069.
30. Belle, V.; Papantonis, I. Principles and practice of explainable machine learning. *arXiv* **2020**, arXiv:2009.11698.
31. Pieters, W. Explanation and Trust: What to Tell the User in Security and AI? *Ethics Inf. Technol.* **2011**, *13*, 53–64. [CrossRef]
32. Schneier, B. *Secrets and Lies: Digital Security in a Networked World*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
33. Kammüller, F. Attack Trees in Isabelle. In Proceedings of the 20th International Conference on Information and Communications Security, ICICS 2018, Lille, France, 29–31 October 2018; LNCS; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11149.
34. Kammüller, F. Combining Secure System Design with Risk Assessment for IoT Healthcare Systems. In Proceedings of the Workshop on Security, Privacy, and Trust in the IoT, SPTIoT'19, Kyoto, Japan, 11–15 March 2019.
35. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining Knowl. Discov.* **2019**, *9*, e1312. [CrossRef]
36. Kammüller, F. Dependability engineering in Isabelle. *arXiv* **2021**, arXiv:2112.04374.
37. CHIST-ERA. SUCCESS: SecUre aCCESSibility for the Internet of Things. 2016. Available online: http://www.chistera.eu/projects/success (accessed on 2 August 2023).
38. Kammüller, F. Isabelle Insider and Infrastructure Framework with Explainability Applied to Attendance Monitoring. 2023. Available online: https://github.com/flokam/Dimpy (accessed on 2 August 2023).
39. Kammüller, F.; Wenzel, M.; Paulson, L.C. Locales—A Sectioning Concept for Isabelle. In *Theorem Proving in Higher Order Logics, Proceedings of the 12th International Conference, TPHOLs'99, Nice, France, 14–17 September 1999*; Bertot, Y., Dowek, G., Hirchowitz, A., Paulin, C., Thery, L., Eds.; LNCS; Springer: Berlin/Heidelberg, Germany, 1999; Volume 1690.
40. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanantions without Opening the Black Box: Automated Decisions and the GDPR. *Harv. J. Law Technol.* **2018**, *31*, 841.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.