Everyday Argumentative Explanations for AI

Jowan van Lente
5994608

First supervisor: AnneMarie Borg
Second supervisor: Floris Bex

A thesis presented for the master degree of
Artificial Intelligence

Graduate School of Natural Sciences (GSNS)
Utrecht University
The Netherlands
May 2022

Acknowledgements

First and foremost, my sincere thanks go to AnneMarie Borg for her bright supervision, her encouraging attitude, and her continuous support. Being able to discuss my thoughts on a weekly basis has been a constructive and pleasant routine. Furthermore, I would like to thank Floris Bex for his feedback along the way. Our arguments about explanations and the explanations about arguments have been both informative and fun.

Furthermore, I am grateful to Accenture for the flexible arrangement of the internship and the productive work environment. Special thanks go to Hanning Ma for making me feel welcome and for helping me organize the internship in a way that allowed me to retain focus on this project while getting to know the company.

Finally, I would like to thank my parents and my girlfriend for their optimism and fresh perspectives.

Abstract

There has been an upswing in the research field of explainable artificial intelligence (XAI) of methods aimed at explaining opaque artificial intelligence (AI) systems and their decisions. A recent, promising approach involves the use of formal argumentation to explain machine learning (ML) applications. In this thesis we investigate that approach; we aim to gain understanding of the value of argumentation for XAI. In particular, we explore how well argumentation can produce *everyday explanations*. Everyday explanations describe how humans explain in day-to-day life and are claimed to be important for explaining decisions of AI systems to end-users. First, we conceptually show how argumentative explanations can be posed as everyday explanations. Afterward, we demonstrate that current argumentative explanation methods compute explanations that already contain some, but not all properties of everyday explanations. Finally, we present everyday argumentative explanations, or *EVAX*, which is a model-agnostic method that computes local explanations for ML models. These explanations can be adjusted in their size and retain high fidelity scores (an average of 0.95) on four different datasets and four different ML models. In addition, the explanations incorporate the main characteristics of everyday explanations and help in achieving the objectives of XAI.

Contents

1 Introduction **4**
- 1.1 Problem set-up . . . 4
- 1.2 Research methods . . . 5
- 1.3 Contributions . . . 5
- 1.4 Thesis outline . . . 6

2 Explainable artificial intelligence **7**
- 2.1 Defining explanation . . . 7
- 2.2 Objectives of XAI . . . 7
- 2.3 Everyday explanations . . . 11
- 2.4 Discussion . . . 13

3 Argumentation and everyday explanations **14**
- 3.1 Argumentation . . . 14
- 3.2 Argumentative explanations . . . 15
- 3.3 Everyday argumentative explanations . . . 17
- 3.4 Argumentation and social explanations . . . 19
- 3.5 Arguing and cognition . . . 20
- 3.6 Discussion . . . 20

4 Current methods **22**
- 4.1 Intrinsic approaches . . . 22
- 4.2 Post-hoc integration approaches . . . 24
- 4.3 Discussion . . . 26

5 Preliminaries **27**
- 5.1 Machine learning . . . 27
- 5.2 Classification task . . . 27
- 5.3 Formal argumentation . . . 28

6 Everyday argumentative explanations **29**
- 6.1 Method outline . . . 29
- 6.2 Notions on modelling choices . . . 33
- 6.3 Toy Example . . . 33

7 Quantitative evaluation **35**
- 7.1 Data-sets . . . 35
- 7.2 Black boxes . . . 35
- 7.3 Metrics . . . 36
- 7.4 Results . . . 36

8 Qualitative evaluation **38**
- 8.1 Conforming to the definition . . . 38
- 8.2 Conforming to everyday explanations . . . 38
- 8.3 Discussion . . . 39

9 Conclusion **40**

References **41**

A Appendix **46**
- A.1 Including feature interactivity . . . 46

1 Introduction

Explainable artificial intelligence, often abbreviated by XAI, has gained much attention in the last few years. The increasing complexity of learning-based systems has led to outstanding results in several domains, such as object recognition and natural language processing (Ren, He, Girshick, & Sun, 2015; Devlin, Chang, Lee, & Toutanova, 2019). However, this complexity has simultaneously caused a diminishing of the understandability of these systems (Gunning et al., 2019; Gerlings, Shollo, & Constantiou, 2021). The integration of these increasingly complex systems in high-stake domains, such as healthcare and law enforcement (Lee, 2018), has sparked a societal demand to gain understanding of the underlying decision-making processes of these systems (Gerlings et al., 2021). The recent draft regulations by the European Union for AI-technologies exemplify this demand (European Commision, 2021), and have created an additional motive for developers and scientists to find solutions for XAI.

An upcoming approach toward XAI is integrating knowledge-based AI with learning-based systems. In particular, this approach explores how the transparent nature of knowledge-based AI may complement the opaqueness of learning-based systems (Calegari, Ciatto, & Omicini, 2020). One form of knowledge-based AI that recently gained some attention for this purpose is *formal argumentation* (Čyras, Rago, Albini, Baroni, & Toni, 2021). Argumentation refers to a systematic reasoning process that can handle conflict and is often claimed to be a vital form of human cognition (Mercier & Sperber, 2011; Besnard & Hunter, 2008). Formal argumentation is the study that aims to formally capture the mechanisms of this phenomenon (Bench-Capon & Dunne, 2007; Simari & Rahwan, 2009). It is mainly involved with defining arguments, determining their relations, and establishing the acceptability of (groups of) arguments and counterarguments (Atkinson et al., 2017; Dung, 1995).

Formal argumentation is increasingly used in studies concerned with explanations (Sklar & Azhar, 2018), which are bundled into an XAI-subfield called *argumentative explanations* (Čyras et al., 2021). Relatively few of those studies explore the integration approach described above because most studies focus on *intrinsic* argumentative explanations. Such explanations explain argumentation-based conclusions and thus remain in the domain of knowledge-based AI. Considering that in the broad field of XAI the strongest demand is currently driven by machine learning (ML) applications (Gunning et al., 2019), it is surprising we have not seen more successful attempts to integrate formal argumentation with such learning-based systems.

Besides the fact that these argumentative explanations for ML are relatively unexplored, the potential benefits of utilizing this integration are deliberately mentioned (Čyras et al., 2021; Cocarascu & Toni, 2016; Vassiliades, Bassiliades, & Patkos, 2021). Čyras et al., for instance, envisage a "fruitful interplay" between ML and argumentative explanations which could "pave the way to new developments" (Čyras et al., 2021, p. 6). Others state that the current approaches to argumentation and ML "show promise for future developments" and allow "incorporation of user-friendly explanations and transparency of the output of ML" (Cocarascu & Toni, 2016, p. 229). However, as we see, the authors remain rather implicit about why argumentation has such a promising future for XAI.

1.1 Problem set-up

The largely unexplored territory of argumentative explanations asks for further investigation. Therefore, in this thesis, we carry out such an investigation; we aim to gain understanding of the value of argumentation for XAI. In particular, we wish to assess how well argumentation can explain ML applications. For that purpose, we will focus on one explanation type and one type of target audience, namely *local explanations* and *end-users*. Local explanations refer to explanations that explain a single prediction of an AI system and are opposed to global explanations, which explain an entire model (Arya et al., 2019). An end-user is a user of an AI system that is affected by its decisions. End-users are often differentiated with more knowledgeable target audiences such as experts or scientists (Arrieta et al., 2020).

To uncover how well argumentation can produce local explanations for end-users, it is useful

to know what a 'good' local explanation is. This, however, is largely a subjective matter. It is therefore no surprise that in the research field of XAI the interpretation of a 'good' explanation often relies on the intuition of the practitioner (T. Miller, 2019). To avoid relying on mere instinct, there is a need for a clear and well-grounded frame of reference.

One frame that seems ideal for this purpose is captured by the notion *everyday explanations* (T. Miller, 2019). Everyday explanations describe how humans explain in day-to-day life and are claimed to be important to keep in mind when creating local explanations for end-users (T. Miller, 2019). Based on a large variety of studies from the social sciences, Miller (2019) found that everyday explanations are contrastive, selected, and social. They are *contrastive* when they describe why an event P happened relative to some other event Q, they are *selected* when they consist of a selected subset of causes and they are *social* when they are adjusted to the receiver of the explanation (T. Miller, 2019). These properties are often posed as valuable qualities for XAI (T. Miller, 2019; Gerlings et al., 2021; Adadi & Berrada, 2018; Arrieta et al., 2020; Gunning et al., 2019; Mittelstadt, Russell, & Wachter, 2019), and some explanation methods are even evaluated based on them (Čyras et al., 2019; Prakken & Ratsma, 2021).

Given that everyday explanations are important for XAI, it is natural to ask if argumentative explanations can be posed as everyday explanations. In this thesis we ask that question; we explore how well argumentation can produce everyday explanations. The aim is to gain understanding of what argumentation has to offer for XAI. The main finding is that argumentation can forms an ideal basis for everyday explanations, and, in that way, helps to reach the objectives of XAI.

1.2 Research methods

We apply two main research methods: a literature research and a controlled experiment. Through the *literature research*, we aim to provide a theoretical assessment of the value of argumentation for XAI in the context of everyday explanations. To get a better grasp of what is valuable for XAI, this literature research is first involved with uncovering the underlying objectives of XAI. Afterward, we discuss what constitutes everyday explanations. Subsequently, we conceptually explore argumentative explanations and the extent to which they can be posed as everyday explanations. Finally, we review current argumentative explanation methods and assess if they compute everyday explanations.

In the *controlled experiment* we model, implement, and evaluate an argumentative explanation method. The aim is to uncover the practical value of argumentation for XAI in the context of everyday explanations. The method explains the decisions of ML classifiers. Such classifiers perform classification tasks, which have been a cornerstone of learning-based AI (Russell & Norvig, 2010), and thus serve as a natural starting point. The evaluation of the method will be both quantitative and qualitative. The quantitative evaluation will assess the truthfulness of the method to the model it explains. The qualitative evaluation reviews if the computed explanations incorporate the properties of everyday explanations and whether or not they help in achieving the objectives of XAI.

1.3 Contributions

The contributions of this research are threefold. First, it provides a theoretical assessment of how argumentative explanations can be posed as everyday explanations. Second, it gives an overview of current argumentative explanation methods and the extent to which they compute everyday explanations. Third, it shows how an argumentative explanation method can practically compute truthful explanations that contain the properties of everyday explanations. These theoretical and practical contributions together increase understanding of the value of argumentation for XAI. If we zoom out further, this thesis has a societal contribution by responding to the demand of making decisions of ML models more understandable to end-users.

1.4 Thesis outline

This thesis is structured as follows. In Section 2 we discuss the objectives of XAI and describe everyday explanations. Then, in Section 3, we conceptually determine if argumentation can create everyday explanations. In Section 4 it is reviewed if current methods compute everyday explanations. Afterward, in Section 5, some important concepts of machine learning, classification and formal argumentation are defined. Subsequently, in Section 6, we propose everyday argumentative explanations, or *EVAX*, which is an argumentative explanation method aimed at computing everyday explanations for ML. Finally, in Sections 7 and 8 we evaluate *EVAX* quantitatively and qualitatively.

2 Explainable artificial intelligence

In this section we discuss key concepts in the research field of explainable artificial intelligence (XAI). In particular, we will define what an explanation is, determine what the main objectives of XAI are and explicate what constitutes *everyday explanations*. Also, we shortly discuss the extent to which everyday explanations seem to help in reaching these objectives. Throughout this thesis, we refer to a person or entity that explains as the *explainer* and the receiver of the explanation as the *explainee*.

2.1 Defining explanation

Finding common agreement on what constitutes an explanation has proven to be a difficult task (Josephson & Josephson, 1996; Malle, 2006; T. Miller, 2019). The multitude of definitions that philosophers and social scientists have proposed portrays the versatile nature of the concept. Below we enumerate and discuss three frequently mentioned ones.

- An explanation is an assignment of causal responsibility (Josephson & Josephson, 1996).
- An explanation is a process to find meaning or create shared meaning (Malle, 2006).
- An explanation is an answer to a *why-question* (T. Miller, 2019).

The first definition refers to how an explanation tells a causal story. It points out how finding possible explanations is finding possible causes for the effect. This definition is given in the context of abductive inference. Abduction is a type of inference that encompasses a process of reasoning from effect to cause, and is closely related to 'the inference to the best explanation' (Bex & Walton, 2016). In particular, this process consists of assembling different explanatory hypotheses (causes) and identifying which set of hypotheses has the greatest explanatory power to explain the claim (the effect) (Josephson & Josephson, 1996). Conventionally, the causes are called the *explanans*, and the effect is called the *explanandum* (the thing to be explained).

The second definition is a definition of behavior explanations. Malle defines an explanation as a tool that can either find meaning or manage interactions. An explanation helps to find meaning when it reconciles missing or inconsistent information in our knowledge. An explanation manages interactions by creating shared meaning and by changing others' beliefs and actions (Malle, 2006). These aspects of explanations seem essential to XAI; the act of creating shared meaning between humans and AI systems lies at the core of XAI (Adadi & Berrada, 2018).

The third definition is used by T. Miller. He defines an explanation as an answer to a *why-question*, and claims that we can subdivide this answer into three parts: a cognitive process, a product, and a social process. The cognitive process is the process of causal inference, which is similar to the first definition, the product is the explanation as a result of this cognitive process, and the social process is concerned with transferring that information from explainer to explainee, which is similar to the second definition. T. Miller notes that this social aspect is an essential addition to more traditional, causality-oriented notions of explanation (T. Miller, 2019).

Thus, we see there is an important role for causality in explanation, but we also observe an indispensable social nature. Since the definition by T. Miller captures both senses it will be adhered to in this thesis.

2.2 Objectives of XAI

In this subsection, we first provide some general context on XAI. Then, to get a grasp on the objectives of XAI, we examine the underlying reasons for explaining AI, discuss the main target audiences, and introduce a trade-off between complete and understandable explanations. Afterward, we formulate a comprehensive notion that describes the objectives of XAI.

2.2.1 Context

As mentioned in the introduction, XAI is a rapidly growing research field involved with producing and evaluating methods and techniques that provide insights into AI systems and their outcomes by presenting them in understandable terms (Gerlings et al., 2021). Throughout the literature, one encounters many central terms like explainability, interpretability, understandability, comprehensibility, and transparency. Even though they are often used interchangeably, they all have slightly diverging meanings (Arrieta et al., 2020). To avoid confusion, we adhere to *understandability* as the central term in this thesis, which is claimed to capture the meaning of most terms above (Arrieta et al., 2020). It refers to the ability to present a model and its decision in human-understandable terms. When we speak of an understandable explanation, we refer to an explanation that explains the *explanandum* in understandable terms for the explainee.

XAI methods can be subdivided into two broad groups of methods (Arya et al., 2019):

- **Model-agnostic** explanation methods *do not* require access to the model. They only use the inputs and outputs of the explained model.
- **Model-specific** explanation methods *do* require access to the model. To exemplify, a model-specific explanation method that explains a neural network might require access to the weight values of the neurons.

Model-specific methods generally explain a model more faithfully, because the access to the inner workings of the model allows for a closer approximation. A large advantage of model-agnostic methods, however, is that because there is only a need for input and output, they apply to a wide range of models. As we already mentioned in the introduction, methods compute roughly two types of explanations (Arya et al., 2019):

- **Global explanations** explain an entire black box model.
- **Local explanations** explain a single prediction.

Whereas global explanations are generally more appropriate for scientific understanding or bias detection, local explanations are usually better suited for understanding and justifying specific decisions (Doshi-Velez & Kim, 2017). As stated before, we will mainly focus on local explanations.

2.2.2 Reasons for explaining

The growing reliance on black box systems in domains like transportation, healthcare, finance, and the military has fueled the need for XAI. This need to explain can be described by four main reasons, assembled in a popular survey on XAI (Adadi & Berrada, 2018). While discussing these reasons, we will integrate findings from other studies as well.

- **Explain to justify.** Recent controversies concerning decisions of black box systems have incited the need to justify these decisions (Bellamy et al., 2019). To accept and act on these decisions, their justification is essential. Explanations can play a justificatory role by increasing the users' understanding, for instance by accentuating the (non-) existence of biases and errors in a system's output (Arrieta et al., 2020). In that way, algorithmic decisions can be defended for being correct, fair, or ethical, which increases trust (Bellamy et al., 2019; Gerlings et al., 2021).
- **Explain to control.** Explanations increase control over a model by uncovering errors and preventing a model from making mistakes (Adadi & Berrada, 2018). The growing need for control, mirrored by legislation in the General Data Protection Regulation (GDPR) (European Commission, 2018), has thus reinforced the need for XAI (Gerlings et al., 2021).
- **Explain to improve.** Practitioners that wish to improve their AI system need to understand the outputs (Adadi & Berrada, 2018). An increased understanding enables the possibility to improve performance, like increasing accuracy or minimizing bias (Arrieta et al., 2020).

- **Explain to discover.** AI systems are often instantiated to discover new patterns in real-world data. When a learned pattern creates a new insight, there is a need for explanation to fully understand that insight; the reasons or causes need to be explicated (Adadi & Berrada, 2018).

We see there is a wide range of motives for XAI, and they naturally have some overlapping components. For instance, when we explain to improve, we often also explain to control or discover. Nevertheless, they provide us with a clear structure of the different senses of the need for XAI. These reasons appeal to both local and global explanations. Intuitively, however, gaining control and improving a system would benefit more from global explanations, whereas justification and discoveries seem to ask for local ones. There is, however, no general rule that describes what type of explanation fits what reason best.

The success of all four reasons relies on the same precondition: explanations should evoke a sense of understanding for the explainees. One can only justify, control, improve and discover once the system and its decision are well understood. A central objective is therefore *to understand*. This may sound obvious, but it grants us an initial finding on the objectives of XAI.

A sole focus on understandability, however, has at least two pitfalls. The first pitfall is that understanding, just as deciding on what is 'good', is a subjective matter; its meaning differs per person. If for a scientist something is understandable, it does not guarantee it is understandable to a high school student. Therefore, there is a need to consciously consider the different types of explainees, or target audiences, one can expect. The second pitfall concerns the fact that a sole focus on understandability may result in a loss of faithfulness. This mechanism is captured by the understandability trade-off. Both pitfalls will be discussed in the next two subsections.

2.2.3 Target audience

The role of the target audience is often neglected in XAI research (Gerlings et al., 2021). Much of the research on XAI is driven by those who develop AI and not by those who need it (Brandão et al., 2019). Besides, there is often no explicit remark about the target audience for XAI methods (Gerlings et al., 2021), which is problematic because different stakeholders have different explainability needs (Arrieta et al., 2020).

When we narrow down the target audience, we first encounter a classical distinction between domain experts and end-users (Arrieta et al., 2020). Whereas domain experts generally search for scientific knowledge, end-users often seek to better understand and verify why a decision was made. Zooming in further, it becomes clear that there is a more versatile span of target audiences. It ranges from data scientists and developers to managers and regulatory agencies, up to product owners and hobbyists (Arrieta et al., 2020). All of these groups have personal explanation needs. To give an example, if an AI system finds a benign tumor in an MRI scan, a doctor may want to explain to the patient how it was found, whereas the manager of the hospital may have legal duties when the tumor turns out to be malignant. Where presenting a single cause would suffice for a doctor, the manager would require a more complete picture. Therefore, when creating an explanation method, a sensible course of action is to narrow down the target audience such that the explanation needs can be more precisely met.

A problem with narrowing down the target audience, however, is that the method would lose universality; the method would apply to a smaller group. A good explanation for a plastic surgeon is not necessarily useful for a hobbyist. On the other hand, however, if one uses a more general description, like *end-users*, there is not one explanation that fits perfectly (Arya et al., 2019). Even though a surgeon and a hobbyist may both be end-users, they have different levels of experience and knowledge, and therefore benefit from different explanations. This makes clear that both narrow and broad definitions of the target audience form problems; it seems as if one has to choose between universality and suitability.

For XAI, a natural solution to this problem is adaptability in explanation methods (Sokol & Flach, 2020). If methods allow explainees to adjust the explanation to their needs, methods become useful for a wider range of audiences. Most current explanation methods, however, do

not incorporate this adaptability (Gerlings et al., 2021). The reason for this absence may be that creating such adaptability becomes rather complex as the range of the target audiences widens. When computed explanations should reach both scientists and inexperienced end-users, the interactivity between the method and the explainee needs to be exceptionally fine-grained.

In this thesis, we regard the target audience as being *end-users*. Because this excludes audiences with higher expertise levels, it lowers the required level of adaptability. It does not, however, allow us to completely ignore adaptability. End-users still have diverging experience levels and therefore different explainability needs. Therefore, a fair amount of focus in this thesis will be spent on the adaptability of explanation methods.

2.2.4 Understandability trade-off

Increasing the understandability of an explanation usually results in the diminishing of the completeness, or faithfulness, of the explanation. This describes the trade-off between two explanation types: complete explanations and understandable explanations. Complete explanations truthfully explain the entire causal chain and necessity of an event, whereas understandable explanations, as we have stated before, are explanations that explain the explanandum in understandable terms for the explainee. Such understandable explanations are often simplifications of complete explanations (Mittelstadt et al., 2019).

The trade-off arises naturally. People explain AI systems because they are too complex to understand (Adadi & Berrada, 2018). To gain understanding, a logical course of action is to simplify; complete explanations are simply too complicated. However, one should be careful because explanations should retain a certain extent of faithfulness to the model (Jacovi & Goldberg, 2020); they should accurately explain the behavior of the model. When one simplifies, there is an increased chance of misrepresenting the model. This is problematic because faithfulness has shown to be crucial for user trust (Papenmeier, Englebienne, & Seifert, 2019). This is where the completeness versus understandability dilemma, or *understandability trade-off*, comes to light: when explaining an AI system one is enforced to find a balance between completeness and understandability, and thus between faithfulness and simplicity (Gerlings et al., 2021).

Both sides of the trade-off are evaluated in different ways. The faithfulness of a model is often measured in *fidelity* (Papenmeier et al., 2019). Fidelity refers to the similarity between the behavior of the explanation method and the model that is being explained. It is claimed to be a core metric of XAI (Mohseni, Zarei, & Ragan, 2021). When explaining classifiers, the fidelity is often the fraction of data points that is assigned to the same output class by both the explanation method *and* the explained model (Mohseni et al., 2021). Evaluating understandability on the other hand, can be done both qualitatively and quantitatively. A qualitative assessment may consist of a qualitative user study or conceptually assessing whether or not the explanations contain understandable properties. Quantitatively it often comes down to user-questionnaires (Arrieta et al., 2020). When one requires an automated metric, however, the size of an explanation can be used to indicate its complexity (van der Waa, Robeer, van Diggelen, Brinkhuis, & Neerincx, 2018).

2.2.5 Comprehensive objective of XAI

The four reasons discussed in Section 2.2.2 made clear that a core part of the objective of XAI involves creating understandable explanations. However, in the subsections thereafter, we showed that this focus on understandability has two pitfalls. First, understanding can vary greatly among different target audiences, and second, understandability can negatively affect the completeness or faithfulness of an explanation, which is unwanted in XAI. Therefore, there is a need to nuance this notion of understandability as the core objective; it should be more comprehensive. First, for an explanation to retain understandability, it should be able to adapt to a broad range of target audiences. Therefore, the underlying objective of XAI should also include the need for adaptability. Second, because it is to be avoided to lose faithfulness in an explanation, the objective should stress the importance of finding a balance between understandability and faithfulness.

A more comprehensive description of the objective of XAI can thus be formulated as *creating understandable, yet faithful explanations for AI systems and their decisions that can adapt to a wide range of target audiences.* Or more simply put, XAI aims for understandable, adapted, and faithful explanations. Below we introduce everyday explanations, which will be used as a central frame of reference in this thesis.

2.3 Everyday explanations

Everyday explanations are explanations that we are used to hearing in day-to-day life. Since they have an ordinary, human-interpretable character and propagate common-sense thinking and talking, they are claimed to be important for XAI (T. Miller, 2019). Everyday explanations explain why particular events occur or decisions were made, and are in contrast with scientific or complete explanations, which describe general scientific relationships or enumerate scientific laws (Mittelstadt et al., 2019).

Three main properties of everyday explanations are assembled in the paper by T. Miller. Based on a large variety of studies from psychology, sociology, and philosophy, it is concluded that everyday explanations are:

- **Contrastive**. Contrastive explanations find their roots in the explainers' tendency to answer why-questions by explaining the cause of an event *relative to* some other event. When an explainee asks "*why P?*", she does not only expect an explanation that describes why that factual event P happened, but also expects information about why a contrasting event Q has not happened (Hilton & Slugoski, 1986). The factual event is often called the fact, whereas the other, contrasting event is called the foil (Lipton, 1990). In this thesis, we adhere to these notions of fact and foil.

  In the XAI literature, the term *contrastive* is sometimes used interchangeably with *counterfactual* (Stepin, Alonso, Catalá, & Pereira-Fariña, 2021). We wish, however, to differentiate between the two. When we talk of counterfactual explanations, we refer to the notion that describes the necessary conditions to *change* the fact to a foil. Such explanation answer the question: "Why P *instead of* Q?". When we refer to contrastive explanations, we refer to a broader notion of explanations that explain an event *relative to* some foil. This includes counterfactual explanations, but also explanations like: "Event P happened *despite of* cause C, which usually implies event Q." We thus regard counterfactual explanations as contrastive explanations, but not vice versa.

  To further clarify the difference, consider the question: "Why is that mushroom poisonous?" A contrastive explanation might be: "*Despite* its white color, which makes it likely it is edible, it is poisonous because of its knobbed cap." An example of a counterfactual explanation is: "The mushroom is poisonous *instead of* edible because the mushroom has a knobbed cap." Both explanations are contrastive, because both explain why the mushroom is poisonous *relative to* some other option (edible in this case). But only the second is a counterfactual statement because it describes the necessary conditions to change the outcome. The wider notion of contrastiveness will be helpful to highlight the contrastive character of argumentative explanations in later sections.

- **Selected**. When explainers explain an event, they tend to provide only a small and biased subset of all possible causes. When looked at closely, we can distinguish between two main components that form this notion of selectedness: minimality and biasedness.

  Minimality refers to the fact that not more than a few causes should be included in an explanation. Rather than presenting the full chain of events, human explainers often only present one or two causes as *the* explanation (T. Miller, 2019). Considering that humans have a limit to the amount of information they can process at once, which is claimed to be around 7 elements (G. A. Miller, 1956), this is not surprising. However, merely minimizing the set of causes is not enough; there is a biased selection process prior to this minimality.

Biasedness describes how the selection process is based on the cognitive biases of an explainer. Even though non-biased selection criteria, like the truth or likelihood of a cause, are important criteria for good explanations (T. Miller, 2019), it is not necessarily how humans select them (Hilton, 1996). Several different cognitive biases determine the selection process. One such bias is *abnormality*, which describes how a cause that is considered unusual is preferred over a 'normal' one (Thagard, 1989). Another criterion is *simplicity*; an explanation that cites fewer causes is preferred over an explanation citing more causes (Thagard, 1989). Two other criteria are *necessity* and *sufficiency*. Necessary causes are generally preferred over sufficient causes, and uniquely sufficient causes tend to be better explanations than cases in which we have multiple sufficient causes (Lipton, 1990). Furthermore, a cause that is considered more *responsible* is generally preferred over less responsible ones (Halpern, 2011). Finally, it is observed that people select causes based on *coherence*, and *generality*. Causes that are consistent with our prior beliefs and able to explain more events tend to be selected for explanations (Thagard, 1989).

- **Social**. Explanations are social when they are adjusted to the explainee (T. Miller, 2019). Recall that an explanation consists of a cognitive and social process (see Section 2.1). Whereas the selected property is more in line with the cognitive process, this property emphasizes the importance of that social process. Explaining is essentially conveying information from explainer to explainee. An essential part is therefore ensuring that the explainee understands the explanation. This amounts to adjusting the complexity, size, or type of the explanation such that the understanding of the receiver is ensured. A common form of this adjustment process is conversation (Hilton, 1990). That process is described by the conversational model by Hilton. It captures the idea that an explanation must be relevant to the question asked, and shows how this can be achieved by a process of exchanging information from explainer to explainee (Hilton, 1990), which usually happens through conversation. In conversation, the explainer can uncover the assumptions and prior knowledge of the explainee and align the explanation accordingly.

These three properties show what constitutes an everyday explanation. Even though they can be viewed as separate aspects, they do overlap and affect each other. For example, when one selects specific causes based on the assumed knowledge of an explainee, the selectedness adds social sense because the chosen subset of causes is then adapted to the explainee. Another example is when an explainer selects a contrastive case out of multiple cases as an explanation; the aim for contrastiveness then also effectuates the selectedness.

2.3.1 The overarching property

If we zoom out, we see that the properties of everyday explanations can be summarized by one overarching theme: they are *cognitive*. They are created, selected, and evaluated by human cognition and describe how an explanation is in line with the reasoning patterns of the explainee.

This cognitive character of everyday explanations is the underlying assumption of the work of T. Miller. Look for instance at the notion that in explanations "probabilities probably do not matter"(T. Miller, 2019, p. 3). This statement resonates with the above claim; thinking in probabilities is not in line with human reasoning. Thus, the underlying assumption is that one should rather produce explanations that *do* follow a human train of thought. And if we look at other XAI studies, we discover that many rely on this assumption too. For example, the rapidly growing amount of counterfactual explanation methods all rely on the idea that counterfactual reasoning is part of human reasoning. In (Byrne, 2019) for instance, six discoveries about ways in which people think counterfactually are described. This study promotes counterfactuals as desirable explanations and thus implicitly assumes that explanations that are congruous to human reasoning are better.

All in all, we see that everyday explanations describe what characteristics cause an explanation to become understandable to humans. Their properties highlight how information becomes

analogous to cognition and delineate how explanations should be adapted to the receiver. In that way, they can be seen as guidelines for understandability.

2.4 Discussion

Recall that the objective of XAI involves creating understandable, adapted, and faithful explanations. After having described everyday explanations, it is interesting to see to what extent they help to reach these objectives. To assess this, we split up the objectives into three parts: understandability, adaptability, and faithfulness, and discuss for every part if everyday explanations help to reach it.

- Everyday explanations assist in creating *understandability*. The contrastive, selected and social properties are, in a sense, guidelines for creating understandable explanations. By describing how explanations become in line with human thought, they help to establish understanding.
- *Adaptability* is also integrated into everyday explanations. The social property describes how explanations should be adjusted to the explainee. Note that this adaptability is a precondition for understandability: adapting an explanation to the explainee enables the explainee to increase her understanding of the explanandum.
- *Faithfulness* does not play a role in everyday explanations. None of their properties describe how causal relations or correlations between the explanandum and the explanans should be retained. On the contrary, actually. Selectedness incorporates the notion of biasedness, which indicates that the causes should not be selected based on their truthfulness, but rather on cognitive biases. This may harm the faithfulness of these explanations.

We see that everyday explanations help to reach two of the three parts of the discussed objective. Their focus on understandability and the emphasis on adaptability resounds with their acclaimed importance for XAI. Nevertheless, one should be mindful of their lack of faithfulness.

Two other implications of everyday explanations, that were already mentioned in the introduction, require some extra attention. First, everyday explanations predominantly apply to local explanations (T. Miller, 2019). They elucidate the salient reasons for a particular event or decision, rather than providing a full causal chain of events. Explaining an entire model in a contrastive, selected and social way is challenging. Therefore, when there is a need for a global explanation, an everyday explanation may be less applicable. Second, everyday explanations seem more suitable for end-users than for domain experts. Everyday explanations are best applied to explainees who lose trust in an AI system when they can not trace the decisions made, rather than for understanding generalized theories (T. Miller, 2019). Therefore, domain experts may benefit more from other, more complete, explanation types.

Nevertheless, when faithfulness is kept in mind, everyday explanations fit the general purpose of XAI. They offer a clear direction for modeling and evaluating explanations. Their value lies in their focus on human understanding and the adaptability of explanations. Being centered around local explanations and end-users is not a reason to reject this direction. On the contrary, it illuminates in what context everyday explanations become most valuable. And since the main focus of this thesis *is* on local explanations and end-users, the properties of everyday explanations will have constructive value; they will form a frame of reference that allows for a structured analysis of argumentative explanations. In that way, they assist in creating a clearer picture of the value of argumentation for XAI. A first step in creating this picture is to assess if argumentation can form a basis for everyday explanations.

3 Argumentation and everyday explanations

In Section 2 we discussed what constitutes everyday explanations and showed how they create understandability and incorporate adaptability. In this section, we assess if argumentation can form the basis for producing everyday explanations. As a start, we sketch some context around the concept of argumentation and introduce formal argumentation. Afterward, we describe two general forms in which an explanation is argumentative. Thereafter, in Section 3.3, we assess for both argumentative explanation forms the extent to which the properties of everyday explanations can be traced back. Subsequently, we zoom in on the cognitive and social nature of argumentation which highlights the suitability of argumentation for creating everyday explanations.

3.1 Argumentation

To get an initial grasp of the meaning of argumentation, we consult a dictionary. The Cambridge online dictionary describes argumentation as: "a set of arguments used to explain something or to persuade people" (Cambridge, 2021b). An argument is defined as "a reason or reasons why you support or oppose an idea or suggestion, or the process of explaining these reasons" (Cambridge, 2021a). When an argument is used to oppose the idea of another argument, it is referred to as a counterargument (Cambridge, 2021c). These definitions show that argumentation has strong links with reasoning and persuasion. But more importantly, it highlights the relation with explanation, which will be further explored in Section 3.2.

In argumentation theory we observe different definitions of an argument, diverging from specific to more abstract ones. Commonly agreed on, however, is the fact that an argument consists of a set of *premises* together with a *conclusion*, in which the premises provide the reasons for the conclusion (Simari & Rahwan, 2009; Besnard & Hunter, 2008). We provide a formal definition of an argument in Section 5.

3.1.1 Formal argumentation

To discuss several formal models of argumentation, we use the conversation below as a running example.

> **Example 3.1.** This is a short discussion on whether Tweety can fly or not between a proponent (P) and an opponent (O).
>
> - P: "Tweety flies because it is a bird." (argument a)
> - O: "Tweety does not fly because it is a penguin." (argument b)
> - P: "The penguin observation was done with faulty instruments." (argument c)

Several attempts to formalize argumentation have been concerned with *deductive arguments* (Besnard & Hunter, 2008). Since people argue to remove doubt about a claim (Walton, 2006), claims that deductively follow from foolproof premises are strong (Modgil & Prakken, 2014). For instance, regarding Example 3.1, if we know for sure that Tweety is a bird and that all birds fly, then argument a would be an infallible claim because it is deductively inferred from true premises. However, such watertight arguments do not exist in the real world. True premises usually do not unfailingly guarantee a true conclusion; they merely add plausibility (Modgil & Prakken, 2014). The fact that Tweety is a bird makes it plausible it flies, however, as argument b shows, this should not necessarily be true. This indicates an incompatibility of deductive arguments with real-world argumentation.

Defeasible arguments are more compatible with how we practically argue. Such arguments also consist of premises that provide support for the conclusion, however, it leaves the possibility for the premises to be true and the conclusion to be false (Modgil & Prakken, 2014). To illustrate, this allows us to accept the claim that Tweety can't fly even though we accept that it is a bird and birds fly. Several formal accounts of defeasible argumentation have been provided. Some major

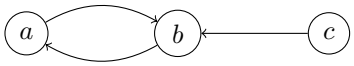

Figure 1: AF_1

ones are abstract argumentation, which refers to arguments as abstract entities (Dung, 1995), ASPIC$^+$, which combines deductive and defeasible arguments (Prakken, 2010) and DeLP, which uses logic programming for defeasible arguments (García & Simari, 2004). In this thesis, we make use of abstract argumentation.

3.1.2 Abstract argumentation

In 1995 Dung presented his pioneering work on abstract argumentation. Since then, a large amount of work on argumentation has been built on these ideas. In abstract argumentation, arguments are abstract entities that have no internal structure, meaning internal relations between premises and conclusions do not play a role. An abstract argumentation framework (AF) is a set of arguments together with a set of attacks. We formally define this notion in Section 5.

We can view an AF as a directed graph in which nodes represent arguments and edges represent attacks between them (Dung, 1995). To illustrate, recall Example 3.1. Consider that arguments a and b attack each other, and that c attacks b. When these arguments and their attacks form an AF, this can be displayed as a graph, which is visualized by Figure 1.

Central to abstract argumentation is the evaluation of sets of arguments that can jointly be accepted. Because the attacks of an AF encode the conflict between arguments, it allows us to identify 'winning' and 'losing' arguments. The properties that are necessary for a set of arguments to be accepted are expressed through different *semantics*. Based on these semantics, we can identify the accepted sets of arguments, called *extensions* (Dung, 1995). In Section 5 we elaborate further on these notions.

3.2 Argumentative explanations

In this subsection, we explore the relation between arguments and explanations. Both are closely related concepts and their distinction can be rather fuzzy (Berland & Reiser, 2009). In some scientific areas, the distinction is not even made at all (Bielaczyc & Blake, 2006; Hogan, Nastasi, & Pressley, 1999). To dive deeper into their relationship, we must discuss an obvious connection between the two. This connection concerns explanatory arguments, which describe when an argument *is* an explanation.

3.2.1 Explanatory arguments

One can differentiate between different senses of an argument. We will be concerned with arguments that have a formal sense, which are arguments with premises entailing a conclusion. In this set of formal arguments, one can find more specific subsets of arguments, or 'sub-senses'. One such subset refers to arguments with an evidentiary sense. Such arguments consist of premises that provide *rational justification* for believing a conclusion. Another subset is concerned with explanatory arguments that answer *why* something is the case (Mayes, 2000). Such explanatory arguments consist of a set of premises, which are the explanans, from which a conclusion, the explanandum, can be inferred (Šešelja & Straßer, 2013). In that way, they are both an explanation *and* an argument. Note that a formal argument can be evidentiary and explanatory at the same time.

An explanation is thus argumentative when we speak of an explanatory argument. An argument, however, seizes to be explanatory when the intention of the speaker changes from informing to persuading the listener (Antaki & Leudar, 1992). To illustrate, consider the following example of an argument.

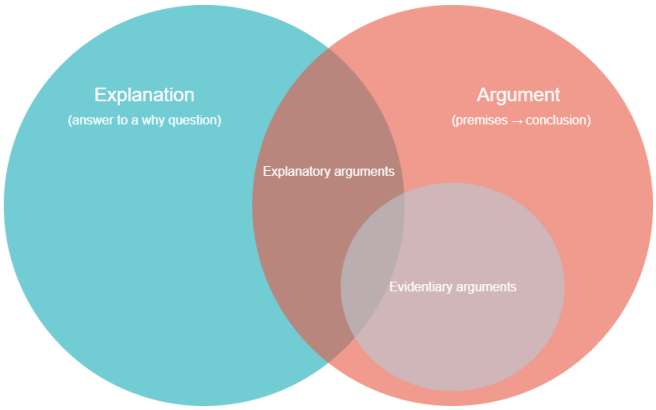

Figure 2: The relation between an argument and an explanation.

Example 3.2. "This mushroom has a foul smell, therefore it is poisonous."

This sentence has the structure of an argument; the premise (stating that the mushroom is foul-smelling), implies the conclusion (it being poisonous). Whether this sentence is also an explanation depends on the underlying assumption of the speaker. If the speaker assumes that the listener thinks the mushroom is indeed poisonous, the speaker wants to inform the listener why it is poisonous. By stating that the mushroom is foul-smelling, the speaker *explains* the reasons for a certain fact. In that way, the utterance is an explanatory argument, and therefore both an argument and an explanation. However, if the speaker assumes the listener thinks the mushroom is *not* poisonous, the speaker uses the sentence to persuade the listener that it *is* poisonous. In that case, the sentence is not an explanation but merely an argument.

To make the difference between persuasion and informing more clear: imagine that the speaker is afraid that the listener, say Alice, will eat the mushroom because she might get sick. From that assumption, the speaker brings up the sentence to persuade rather than inform Alice that the mushroom is poisonous. In that way, the speaker prevents Alice from getting sick by persuading her with an argument. This persuasive character averts the utterance to be an explanation. However, if the speaker would assume that Alice *does* believes the mushroom is poisonous, this persuasion would not be necessary. The utterance would then simply be informative for Alice; she would after all gain an understanding of the reasons behind the toxicity. The sentence then is the intersection of an argument and an explanation. This intersection is what we call an explanatory argument. In Figure 2 the relation between argument and explanation is visualized as a Venn diagram.

This connection between an argument and an explanation indicates how an explanation becomes argumentative in the form of one argument. Argumentation, however, typically involves handling multiple arguments and counterarguments (Besnard & Hunter, 2008). Therefore, it is interesting to see if and how multiple arguments may function as an explanation. Below we show that when arguments together resolve an issue or disagreement, they can function as an explanation for the 'winning' conclusion.

3.2.2 Explanatory discussions

When the dialectics of a discussion are used to resolve an issue, the discussion functions as an explanation for the resolution of that issue. In other words, discussions explain the disagreement they resolve. We will refer to such discussions as *explanatory discussions*. This dialectical way of explaining finds its roots in the idea that a discussion is a competitive interaction in which a propo-

nent and opponent present and defend their claims until one 'wins' and the other 'loses' (Berland & Reiser, 2009). From a more benevolent point of view, this can be seen as a form of a collaborative interaction where two sides are working together to resolve an issue or disagreement (Sawyer, 2005, p. 443). The dialectical process of comparing conflicting information that leads to the resolution naturally explains how that resolution was found. This means that the discussion itself then *is* the explanation.

A special characteristic of this type of explanation is that they, by definition, build up to their conclusion; they describe the series of causes that lead to a particular event, decision, or outcome (Berland & Reiser, 2009). To give an example, consider a debate about lowering taxes between two political parties. To explain why the taxes are being lowered, the dialectics of that discussion can function as an explanation. All posed arguments in favor and against a decision describe how the decision was made. Or consider that someone is weighing off the advantages and disadvantages of working late on a Friday. She is essentially having a discussion with herself; in a dialectical manner, she is resolving a disagreement in her head. When she makes the decision and her colleague asks: "Why are you working late on a Friday?", she could provide a representation of her mental discussion as the answer.

Note that the counterarguments play an interesting role in explanatory discussions: they add a sense of what reasons were against the conclusion, or 'tried to prevent it'. They describe an outcome that has not happened. In that way, they create a nuanced picture of the grounds for a certain claim. They create sentences like: "*even though* cause C occurred, event P still happened." In the next subsection, we see that these counterarguments add a contrastive sense to an explanation.

3.3 Everyday argumentative explanations

In the last subsection, we presented two types of argumentative explanations: explanatory arguments and explanatory discussions. The first describes when an argument is the equivalent of an explanation and the second showed that multiple arguments, when they resolve a disagreement, may function as an explanation too. Both can be defined by the number of arguments (n) they consist of:

- **Explanatory argument ($n = 1$).** The *explanandum* of an explanatory argument is the conclusion of that argument and the *explanans* is the set of premises.
- **Explanatory discussion ($n > 1$).** The *explanandum* of an explanatory discussion is the conclusion of the winning arguments and the *explanans* are the premises of the winning arguments together with all losing arguments (including their conclusions).

If we assume that an argumentative explanation should exclusively consist of one or more arguments, we have captured all argumentative explanations with the two definitions above. We do, however, acknowledge that there may be other explanation forms that can be called argumentative. Nevertheless, we proceed with these two forms because they offer a demarcated concept of argumentative explanations, allowing for structured conceptual analysis.

Recall that we wish to uncover if argumentation can produce everyday explanations. A useful step would be to assess if the two argumentative explanation forms described above can function as everyday explanations. For that reason, we will investigate the extent to which these forms can be contrastive, selected, and social.

3.3.1 An explanatory argument as everyday explanation

For an explanatory argument to be *contrastive* it requires to be an explanation for why an event P happened relative to some event Q. If an argument a explains event P, the premises of a naturally consist of the causes, and the conclusion of a is event P itself. In that way, the premises provide the reasons for the conclusion P. In such arguments, there is no room for another conclusion about a contrastive event Q; an argument solely has one conclusion. To include Q, one would need

another argument where the truth of the premises implies Q. Adding that argument, however, transforms the explanation into a discussion, which is discussed in the next Section (3.3.2). We could, of course, search for contrastiveness in the premises. However, contrastiveness, as defined in Section 2.3, only involves contrasting outcomes (or facts) with other outcomes (or foils) rather than contrasting causes (or premises). Therefore, we claim that contrastive explanatory arguments, in that sense, do not exist.

For an explanatory argument to be *selected*, the premises should be a selected biased subset of all possible causes. If we look at Example 3.2, which explains why a mushroom is poisonous by posing one argument, it is clear that it is a selected subset; just one cause is chosen as being *the* explanans, even though one can think of more reasons why the mushroom is poisonous (say the color or texture). Only in the case that there is just one cause to appoint, it becomes debatable. We can also observe how this subset of causes can be biased. For instance, if we assume that a mushroom with a 'foul smell' is abnormal, we could say that the subset of causes in Example 3.2 is biased because it would be based on the bias that describes the cognitive preference for abnormality.

For an explanatory argument to be *social*, it needs to be adjusted to the explainee. This entails that the way the explanatory argument is presented should align with the needs of the explainee. This, of course, largely depends on the degree to which the explainer is able to adapt to the prior knowledge and underlying assumptions of that explainee. Therefore, up to this point, we can not say much more about the social sense of explanatory arguments. For that, we would need to get a better understanding of the adaptability of the explainer. We elaborate further on this adaptability in Section 3.4.

3.3.2 An explanatory discussion as everyday explanation

For the second form, in which the explanation amounts to an explanatory discussion, we use the following example.

> **Example 3.3.** This is an explanatory discussion between a proponent (P) and an opponent (O) about why a mushroom is poisonous.
>
> - P: "This mushroom is poisonous because it has a foul smell, and since foul-smelling mushrooms are usually poisonous, this one probably is too."
> - O: "This mushroom is indeed smelly, however, since it has a white color, and since mushrooms with a white color are usually edible, this mushroom is probably edible."
> - P: "It is indeed white, but it also has a knobbed cap. And since mushrooms with a knobbed cap are always poisonous, this one is too."

The proponent's first argument is attacked by the opponent's argument, making it likely the mushroom is edible. However, the first argument is reinstated by the last argument, causing the proponent to 'win'. Hence the textual representation of this discussion could be seen as an explanation for why the mushroom is poisonous.

For this explanatory discussion to be *contrastive*, the discussion should explain why an event P happened *relative to* an event Q. Thus, in this example, it should not only explain why the mushroom is poisonous but also explain why it is something else rather than poisonous. Because the opponent explains why the mushroom is probably edible, which is something else than poisonous, it gains a sense of contrastiveness. This highlights how counterarguments give a discussion a contrastive character; they portray the reasons for the validity of that foil.

Selectedness can be ensured if both minimality and biasedness are incorporated. In the running example, we can identify two causes for the fact that the mushroom is poisonous: the foul smell and the knobbed cap. In addition, there is one reason against it: the white color. When there exist more causes than just the smell, the cap type, and the color, the running example includes just 'a few' of them, hence conforming to the notion of minimality. In addition, explanatory dialogues can, similar to explanatory arguments, include a biased selection of causes. Again, think

of abnormality. Adding an argument stating that the mushroom is poisonous because of a rare purple dot would already add a form of biasedness.

An explanation is *social* if the explainer has adjusted it to the explainee. Just as for explanatory arguments, this is hard to measure for explanatory discussions; one needs to know the extent to which an explainer can adapt to the explanation. In the running example, we could say that the proponent is the explainer and the opponent, since she lost the argument, is the explainee. In that case, it is fair to say that the explanation has a social sense; the explainee is part of a conversation. However, since we regard the whole discussion as *the* explanation, there must be some third party that is the explainee. This means that the proponent and opponent then together operate as the explainer. Their explanation becomes social when that third party gains a sufficient understanding of the reasons for the explanandum. This again depends on the ability of the explainers to adjust their explanation to that explainee. In a context where the explainer is a computational method, which is common in the field of XAI, this entails that a method, to compute social explanations, should incorporate adaptability. In the following subsection, we elaborate on the social character of argumentative explanations.

3.4 Argumentation and social explanations

Whilst discussing argumentative explanations in the last subsection we encountered that one can not identify a social explanation solely based on the product; its social character relies on the adaptability of the explainer. In an everyday setting, this adaptability is usually present because the explainer is a human, and humans naturally adapt to their audience (T. Miller, 2019). For explanations for AI, however, the explainer is often a machine, and adaptability is not guaranteed. A developer should therefore incorporate this consciously. The type of method that a developer uses, however, should allow this incorporation of social explanations. Needless to say, this requirement also applies to argumentative explanation methods.

Argumentation can form an ideal basis for social explanations because it is highly adjustable and can be posed as a form of dialogue. This adjustability of argumentation can best be shown by displaying the wide range of possible explanations that can be instantiated from an abstract argumentation framework (AF) (Dung, 1995). These AF-based explanations amount to different ways of presenting the arguments and their attacks. One leading approach involves identifying and depicting sub-graphs, as exemplified by Figure 1. This approach assumes we can understand the dialectics of argumentation as a graph. In such explanations, arguments are seen as nodes and their relations as edges. These explanations can take on many forms, such as *paths*, *cycles* and *branches* (Šešelja & Straßer, 2013; Cocarascu, Stylianou, Čyras, & Toni, 2020; Espinoza, Tacla, & Jasinski, 2020; Čyras et al., 2019). Another form of AF-based explanations are *extensions* (Fan & Toni, 2015), which are sets of accepted arguments. Similar explanations are created by defining sets of arguments that are *necessary or sufficient subsets* for argument acceptance. In that way, explaining becomes showing which arguments were essential for the acceptance of sets of arguments (Borg & Bex, 2021b).

An AF can also be presented as a conversation or dialogue. Since the adjustment process of social explanations is often in the form of a conversation (Hilton, 1990), this ability of an AF should not be ignored. A frequently used form of an AF-based conversation deployment is a *dispute tree* (Čyras et al., 2019; Fan & Toni, 2015). These dispute trees can act as evidence for the acceptability of arguments. Another good example is the dialectical system for explanatory dialogues, in which a general protocol determines the dialogue process (Arioua & Croitoru, 2015). A further option is a *dialogue game*, which has the form of a game between a proponent and an opponent both aiming to win the game with regard to the topic argument (Raymond, Gunes, & Prorok, 2020). Note that for real-time adjustments with a human explainee, however, there is a need for human-machine interactivity. Enabling such interactivity is one of the core objectives of the research field of human-computer interaction (Tripathi, 2011). For argumentative explanations, we see that interactivity is increasingly being included (Sendi, Abchiche-Mimouni, & Zehraoui, 2019; Rago, Cocarascu, Bechlivanidis, & Toni, 2020; Čyras et al., 2019).

3.5 Arguing and cognition

So far, we have seen that argumentative explanations may function as everyday explanations because they are social and they can be contrastive and selected. In Section 2.3 we claimed that these properties have a common characteristic: they are cognitive in that they follow human reasoning patterns. Therefore we would expect that argumentation, since it can produce such cognitive explanations, follows human reasoning patterns too. In this subsection, we strengthen this idea by showing how arguing is entangled in cognition.

The idea that argumentation is central to cognition resonates back to the major early philosophers Plato and Aristotle, who encountered argumentation in many forms of reasoning and interaction (Kuhn, 1991). And we still do; argumentation is involved in many parts of day-to-day life, both in social context and cognition. Examples are discussions at the dinner table or political debates, or mentally weighing off *pros* and *cons* while deciding on a particular matter. In fact, the sentence "humans argue" is a truism; either you already believe it or you would need to argue against it (Atkinson et al., 2017).

Empirical studies also show that argumentation is central to human reasoning. The work by Mercier and Sperber (2011) for instance, sums up research that shows how people are generally skilled arguers. Studies reveal that the main function of reasoning is an evolutionary product that was not driven by pursuing truth, but by winning arguments. Human reason developed because it had to produce and evaluate arguments to persuade others. This implies we are constantly looking for arguments that can justify our beliefs or actions, causing argumentation to be a core part of human reasoning (Mercier & Sperber, 2011). Additional evidence comes from a study that used Bayesian modeling to test how participants react to argumentative fallacies versus other logical fallacies. Results show participants react more appropriately to argumentative fallacies than logical ones, which suggests that also Bayesian accounts of reasoning match with the idea that argumentation is an essential form of human reasoning (Hahn & Oaksford, 2007).

3.6 Discussion

In this section, we have aimed to investigate whether or not argumentation, at least on a conceptual level, can form a basis for creating everyday explanations. We have distinguished between two forms of argumentative explanations: explanatory arguments and explanatory discussions. We showed that explanatory arguments can be selected and that explanatory discussions can be both contrastive and selected. In addition, we highlighted the social ability and cognitive character of argumentation, which shows its compatibility with everyday explanations.

These results indicate that, on a theoretical basis, argumentation is highly suitable for creating everyday explanations. Especially explanatory discussions and AF-based explanations contribute to this suitability. When both are combined, thus when explanatory discussions are based on an AF, they would allow for contrastive, selected, and social explanations.

To extrapolate these findings to a more general indication of the value of argumentation for XAI, we can say that argumentation, through its ability to create everyday explanations, approximates the value of everyday explanations for XAI. Since that value is widely recognized (T. Miller, 2019; Gerlings et al., 2021; Adadi & Berrada, 2018; Arrieta et al., 2020; Gunning et al., 2019; Mittelstadt et al., 2019; Prakken & Ratsma, 2021; Čyras et al., 2019), the promising words about argumentative explanations do not seem baseless. Do, however, note that since everyday explanations mainly apply to local explanations for end-users, we can not make grounded claims on the theoretical value of argumentation for global explanations and other target audiences. Further research is needed to evaluate such claims.

To further elaborate on this value of argumentation, recall that the objectives of XAI corresponds with a need for understandable, adapted, and faithful explanations. It is now fair to say that argumentation, just as everyday explanations, can assist in reaching these objectives too. Especially the cognitive and social sense of argumentative explanations allow this. The cognitive sense enhances understandability because it describes how explanations become in line with human thinking, and the social sense enables adaptability by stressing how explanations should be

adjusted to the explainee. This means that when argumentative explanations remain faithful to the model they explain, they would touch upon three core parts of the objectives of XAI.

The findings so far remain on a conceptual level. To get a more complete picture of the value of argumentation in the context of everyday explanation, we require practical insights. Therefore, in Section 4, we review current methods and evaluate to what extent they produce everyday explanations. Then, after discussing some important preliminaries in Section 5, we present an argumentative explanation method that incorporates the key findings of this thesis in Section 6.

4 Current methods

In the last section, we conceptually showed how argumentative explanations can function as everyday explanations. In this section, we review current scientific developments in argumentative XAI and assess the presence of the properties of everyday explanations in computed argumentative explanations. The recent survey by Čyras et al. will act as a guideline throughout this section (Čyras et al., 2021). In that survey a distinction is made between *intrinsic* and *post-hoc* argumentative explanations. Intrinsic explanations explain 'native' argumentation methods, that is to say, they explain formal argumentation mechanisms in their own terms. Post-hoc approaches extract information from non-argumentative methods to explain these methods in argumentative terms. Note that not all post-hoc approaches are used to explain learning-based systems, they might as well explain knowledge-based systems.

We will discuss three intrinsic approaches, after which three post-hoc approaches that integrate argumentation with learning-based models are presented. For each of these approaches, we review the extent to which they compute contrastive, selected, and social explanations. To get a clear view of what we mean with such explanations in the context of argumentative XAI, we first present some workable definitions. The guiding survey (Čyras et al., 2021) and the findings so far form the basis for these definitions.

- **Contrastive argumentative explanations** are explanations that provide reasons *pro* and *con* the outcome (Čyras et al., 2020, 2021). Such explanations describe why the fact happened, and, at the same time, provide reasons to believe the foil. In that way, they explain a fact *relative to* some foil and are thus in line with the notion of contrastiveness from Section 2.3. Since in the field of XAI it is common to provide counterfactual statements that describe the necessary conditions to *change* the fact to the foil (Stepin et al., 2021; van der Waa et al., 2018), we will appraise the presence of such statements too.

- **Selected argumentative explanations** have two components: (1) they contain no more than a few arguments that explain an outcome and (2) these arguments should be selected based on at least one cognitive bias. These respectively refer to the two components of selectedness: minimality and biasedness.

- **Social argumentative explanations** are explanations that are created by an argumentative explanation method that allows for adjusting the complexity or size of the explanations. Since an AF allows for multiple different explanation deployments, we regard AF-based explanations as social *by definition*. When methods incorporate a form of interactivity, we also acknowledge the presence of a social sense.

4.1 Intrinsic approaches

Intrinsic argumentative explanations compute explanations *for* formal argumentation methods. Examples are explanations for argumentative recommender systems (Briguez et al., 2014; Rago et al., 2020), decision making based on argument acceptance (Borg & Bex, 2021a; Brarda, Tamargo, & García, 2019) or argument-based planning (Oren, van Deemter, & Vasconcelos, 2020). However, since in this thesis we focus on classifiers, it is perhaps more interesting to look at explanations for argumentative classifiers (Čyras et al., 2019; Cocarascu et al., 2020).

A first argumentative classifier is **DEAr** (Cocarascu et al., 2020). With DEAr, arguments are mined directly from data, and an argumentation debate in an AF is used as a binary classification method. The mining process is executed by a characterization extractor that can identify and select features in datapoints. An argument then consists of a datapoint containing a reduced set of features and an outcome. It can attack another argument if it is *more informative*. Informativeness can be defined in different ways, like through the amount or strength of features. Explanations can be deployed in a dialectical manner in the form of a dispute tree (Cocarascu et al., 2020).

Considering the properties of everyday explanations, DEAr is contrastive in the sense that it shows reasons for and against its prediction. It does not, however, include counterfactual

statements that describe the necessary conditions that change the fact to the foil. The explanations are not selected. Because all features are used in the explanation, there is no selection process, let alone based on a cognitive bias. Moreover, because explanations have a size that corresponds to the number of features in a datapoint, explanations for datasets with hundreds of features become large, complex, and therefore incomprehensible. The explanations can be employed in the form of a dispute tree, which does give it a social character.

Another argumentative classifier is a method that classifies and explains binary outcomes using **Arbitrated Argumentative Disputes** (Čyras et al., 2019). The reasoning in these disputes is influenced by case-based reasoning (CBR) but is driven by abstract argumentation. CBR classifies a new case (or input point) by looking at similar previous cases (or precedents). To determine the output for that case, one looks at the differences between the case and a precedent. The arguments in the AF are cases consisting of features, stages, and an outcome. The features represent all information about a case, the stages represent how that information progresses over time, and the outcome determines if the case is accepted or rejected. To explain the reasoning process of the disputes, arbitrated dispute trees are defined, together with *excess features*. These features belong to the losing's disputant case that caused the winner to win. This helps to clarify why some cases do not hold ground in a dispute, and can thus enhance the understandability (Čyras et al., 2019).

This method computes contrastive explanations: the excess features show what information losing cases have in contrast to the focus case. In that way, they also describe the necessary conditions that would *change* the outcome. Furthermore, it is social since the explanations can be altered interactively in a collaborative environment (Čyras et al., 2019). It is claimed it is selected because only one arbitrated dispute tree is chosen out of multiple possible trees (Čyras et al., 2019). And this does indeed show its minimality. Given that there is no limit to the size of one tree, however, there is no guarantee only 'a few' cases will be included. With complex case studies, the trees may therefore become more difficult to interpret. In addition, the explanations are not selected based on a cognitive bias. Nevertheless, it is mentioned that this may be included in future work. Figure 1 depicts an example of an arbitrated argumentative dispute.

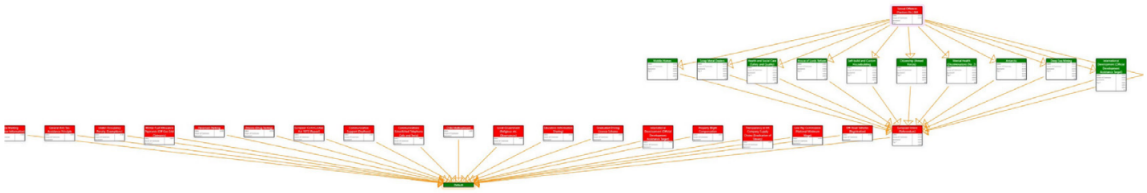

Figure 3: An example of an arbitrated argumentative dispute in a legislation setting (Čyras et al., 2019).

For this proposal, we are looking to explain *learning-based* classifiers. Reviewing intrinsic classifiers such as DEAr (Cocarascu et al., 2020) and the Arbitrated Argumentative Disputes (Čyras et al., 2019) might therefore at first glance seem somewhat senseless. However, if these classifiers can function as a transparent counterpart of a learning-based system, they acquire new explaining capabilities.

4.1.1 Twin-systems

The study by Kenny and Keane is ideal to illustrate the idea of a transparent counterpart, also referred to as a *twin-system*. In that study, a twin-system is presented which performs classification tasks alongside a neural network (NN) (Kenny & Keane, 2019). To set up the system, solely inputs and outputs of the NN are used, after which CBR is employed for the actual classification (Kenny & Keane, 2019). The CBR twin-system is used to explain the NN by presenting the most similar case for every classification instance. See Figure 4 for further clarification on the concept. Similarly, we could employ intrinsic classifiers like DEAr as an explanation method.

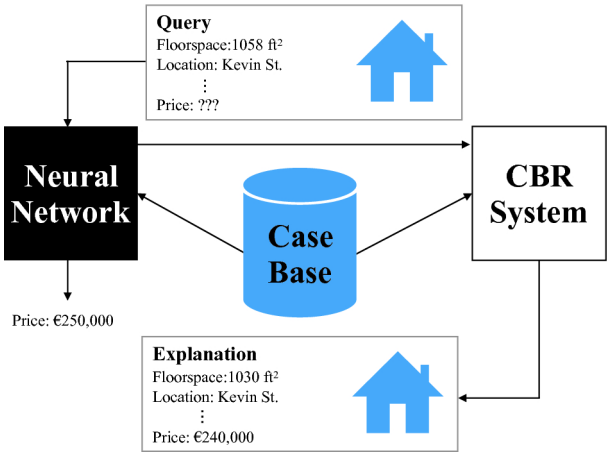

Figure 4: CBR twin-system (Kenny & Keane, 2019).

An example of an argumentative explanation method already behaving as a twin-system is the Top-level Model of Case-based Argumentation for Explanation, or **CBA for explanation** (Prakken & Ratsma, 2021). This model is an instantiation of an argument game defined as an AF and is aimed at explaining ML decisions. It is top-level since it allows more detailed accounts to extend the model. The model represents every input instance as a case, similar to (Kenny & Keane, 2019) and (Čyras et al., 2019), and explains an input instance, or focus case by initiating the argument game in which a proponent presents the most similar precedent. The opponent then provides counterarguments by stating differences between the focus case and the precedent or by citing counterexamples. The proponent then tries to attack these arguments by stating why they are incorrect or irrelevant. The explanation comes down to a presentation of the winning strategy that successfully counterattacks all arguments of the opponent (Prakken & Ratsma, 2021).

This method employs explanations in a dialogical form, which creates prospects for interactive, social explanations. In addition, the fact that cases argue for an alternative outcome causes the explanation to have a contrastive sense. Because it only presents the differences between a precedent and a focus case, it does also have a sense of minimality. However, since the selection process is not based on cognitive biases, we can not regard the explanations as being selected.

We see that when intrinsic classifiers are employed as a twin-system, they have the potential to explain other learning-based classifiers. However, we should remark that they are model-agnostic and can therefore lack faithfulness; the reasoning mechanisms of argumentative classifiers can differ from that of learning-based classifiers. Nevertheless, if one can prove the output is consistently similar, for instance with fidelity scores, the explanations have potential value.

4.2 Post-hoc integration approaches

Post-hoc methods extract information from the model to the explanation method (Čyras et al., 2021). We can divide those methods into *complete* and *approximate* methods. Complete methods provide a complete mapping between the model and the explanations, whereas with approximate methods it is, as we would expect, an approximation of the model (Čyras et al., 2021). We will discuss three approaches that integrate argumentative explanations and learning-based classifiers: one complete and two approximate methods.

The complete integration approach creates Deep Argumentative Explanations for neural networks, abbreviated as **DAX** (Albini, Lertvittayakumjorn, Rago, & Toni, 2020). DAX creates an influence graph based on the NN, which is subsequently used to instantiate a generalized argumentation framework (GAF). In this process neurons or groups of neurons are converted into arguments, and the edges between them represent the attacks. In that way, an AF represents the

NN. Note that if every single neuron is represented by an argument, we have a complete mapping of the network. To evaluate the understandability, a user experiment on DAX for text classification is conducted and the results show it has some beneficial properties compared to other explanation methods.

Since there are no counterfactual statements or counterarguments included that argue for an alternative class, DAX is not contrastive. It neither includes minimality nor biasedness; the explanations can contain up to as many causes as the explained NN contains neurons and there are no cognitive biases involved. Since the created AF allows different ways of presenting the explanation, it does incorporate a social characteristic. Figure 5 visualizes an explanation computed by DAX.

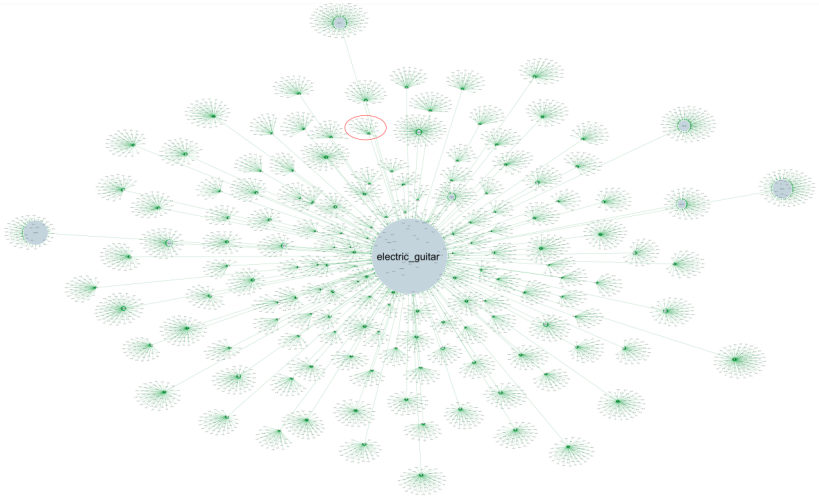

Figure 5: An example of an explanation for a prediction of an image of a electric guitar. The 5145 arguments that are visible are less than 0.4% of all arguments that were originally part of the explanation; the other 1,467,079 arguments are left out (Albini et al., 2020).

The first approximate integration approach combines ML algorithms with argumentation in multi-agent systems (Sendi et al., 2019). This **New Transparent Ensemble Method** extracts arguments from classifiers, assigns sets of arguments to agents, and lets these agents compete against each other according to a multilateral argumentation protocol. An advantage of this method is that it allows a user to add domain knowledge to the classifier, by making use of an 'expert agent' that joins the debate with manually constructed arguments (Sendi et al., 2019).

This method is social because it incorporates the possibility to add domain knowledge. The inclusion of arguments with a foil also gives it a contrastive sense. Because the argumentation protocol terminates when all agents have put forward their arguments (Sendi et al., 2019), there is no guarantee of minimality; agents can not put forward just 'a few' of their arguments. There is also no biased selection of presented arguments involved.

The second approximate method is a **Non-Monotonic Explanation Function** (Amgoud, 2021). This function is an abductive explanation method that creates an AF. The arguments are constructed by determining the minimal set of features that cause an instance to be assigned to a class. The function returns a set of *naive* extensions (Bondarenko, Dung, Kowalski, & Toni, 1997), by computing conflict-free sets of arguments. All arguments in the extension can be used as an explanation. When there are no non-conflicting arguments, the function returns an empty set (Amgoud, 2021).

This method is contrastive in two ways: firstly, a form of contrastive explanation is included in the paper and secondly, it includes counterarguments that argue for an alternative class. The explanations are social because the method computes an AF. Computed explanations can contain just a few arguments, and therefore enable minimality. The extension, however, is as large as the amount of non-conflicting arguments. This has no limit, and explanations can therefore become

large. In addition, because the explanation is a naive extension, it is likely the method returns an empty set. Such output can not be considered a selected explanation; a set containing zero causes implies that no causes have been selected.

4.3 Discussion

The extent to which the properties of everyday explanations are present in the discussed methods is summarized in Table 1. We can see that all argumentative explanations are social, which can largely be attributed to their use of an AF. Most methods are contrastive too: they include counterarguments that explain an event P *relative to* an event Q by posing arguments with a foil as the conclusion. The use of counterfactual statements that describe the conditions to change the fact to a foil, however, is only apparent in two of them (Čyras et al., 2019; Amgoud, 2021).

Establishing the presence of selectedness involved looking at its two components: minimality and biasedness. *Biasedness* is absent in the discussed methods; none of them selects causes based on a cognitive bias. Assessing if the explanations incorporate *minimality* has been less straightforward; the size of the explanation can greatly differ per method. Also, our notion of minimality is rather intuitive; having 'not more than a few arguments' does not indicate a clear limit. Nevertheless, it is fair to say that explanations with hundreds of causes do not conform to this notion. Moreover, such large explanations tend to become difficult to understand for humans, and are thus unwanted anyway. Since most discussed methods compute explanations that grow as the size and complexity of the explained model or dataset increases, there is no guarantee these methods compute 'minimal' explanations. Some can contain thousands (Cocarascu et al., 2020) or even millions (Albini et al., 2020) of causes. As mentioned before, the maximum amount of information pieces a human can process at a time is around 7 (G. A. Miller, 1956). If we would use this number as a restriction that defines minimality in explanations, none of the discussed methods would be able to guarantee it, because none of them includes a restriction on the explanation size. Nevertheless, when the used datasets and the explained models are sufficiently simple, most methods *are* able to incorporate minimality. In those cases, however, *biasedness* still lacks.

The reason methods do not incorporate selectedness may be that including it can diminish the faithfulness of the explanations. Incorporating minimality through a restriction on the explanation size, for instance, may reduce faithfulness because it may force a method to exclude important causes. This can happen when datasets are large and models complex, which means there are generally more important features, and thus salient causes. When the amount of causes in the explanation is restricted, there is an increased chance that a truly salient cause needs to be excluded. This may reduce faithfulness because the explanans then explain the explanandum less truthfully. Biasedness may also negatively affect faithfulness; when the inclusion of biased causes results in biased output, the output may differ from the explained model, which decreases fidelity.

Method/Quality	**Contrastive**	**Selected**	**Social**
DEAr (Cocarascu et al., 2020)	X		X
Arbitrated Argumentative Disputes (Čyras et al., 2019)	X		X
CBA for Explanation (Prakken & Ratsma, 2021)	X		X
DAX (Albini et al., 2020)			X
New Transparant Ensemble Method (Sendi et al., 2019)	X		X
Non-Monotonic Explanation Function (Amgoud, 2021)	X		X

Table 1: The presence of the properties of everyday explanations per argumentative explanation method.

5 Preliminaries

In this section we recall some core concepts in artificial intelligence (AI) and formal argumentation.

5.1 Machine learning

Machine learning (ML) is a subfield of AI that is concerned with programs that use experience to increase performance (Russell & Norvig, 2010). Let us define an ML model.

> **Definition 5.1.** An ML model is a function $f : \mathcal{X} \to \mathcal{Y}$ in which $\mathcal{X}$ is referred to as the *input space*, and $\mathcal{Y}$ the *target space*. The function maps inputs $x_1, x_2, \ldots, x_i \in \mathcal{X}$ to outputs $y_1, y_2, \ldots, y_i \in \mathcal{Y}$, such that $f(\mathcal{X}) = \mathcal{Y}$.

When the elements in the target space $\mathcal{Y}$ take on a finite set of values, we refer to these values as output classes in $\mathcal{C} = \{c_1, \ldots, c_m\}$, with $m > 1$, where m refers to the number of possible output classes. An input point x_i in the input space is equal to a set of features $\mathcal{F}_i$. Below we define a feature.

> **Definition 5.2.** A feature is an attribute-value pair $(a, v) \in \mathcal{F}$, where a indicates the label of the feature and v indicates its corresponding value.

The value of a feature is taken from a set of domains $\mathcal{D}$ all containing potential feature values. A domain $d \in \mathcal{D}$ can be either *discrete* or *continuous* (Mohri, Rostamizadeh, & Talwalkar, 2012). To illustrate, consider the following example in which we want to predict whether or not a student will be accepted into university.

> **Example 5.1.** Let $x_1, \ldots, x_6 \in \mathcal{X}$, where every $x \in \mathcal{X}$ is a student, and let $\mathcal{C} = \{0, 1\}$, where 1 represent a student being accepted to university, and 0 a rejection. The set of features consists of $\{g, t, m\} \in \mathcal{F}$, where g corresponds to the (rounded) average grade of the student, t to whether or not the student passed the entry test and m to whether or not she is motivated. The domain then consist of $\mathcal{D}_g = \{n \in \mathbf{N} \mid n \leq 10\}$, $\mathcal{D}_t = \{0, 1\}$ and $\mathcal{D}_m = \{0, 1\}$. All these domains are examples of discrete domains.

A common distinction in ML tasks is the one between supervised, unsupervised, and reinforcement learning (Russell & Norvig, 2010). The former will be most relevant for this thesis because classification is a form of supervised learning. Nevertheless, to display the full extent of ML, the other two forms will also briefly be discussed. *Supervised learning* is concerned with labeled data, that is to say, for every input $x_i \in \mathcal{X}$, the correct output, or label, y_i is known. The goal is to approximate f such that it accurately mirrors the relationship between the inputs and their class labels. This is done using a loss function, that computes how well f performs, and modifies the model accordingly. This process is referred to as the *training phase*. Then, in the *testing phase*, the function is presented with unseen data (with hidden labels) to assert how well it predicts the corresponding class labels, which is articulated in *accuracy* (Mohri et al., 2012). *Unsupervised learning* works with unlabeled data and is aimed at identifying clusters and exploring underlying structures in the data. With *reinforcement learning*, the model learns by evaluating its outputs, not with correct and incorrect outputs like in supervised learning, but in the form of rewards and punishments. In that way, the model learns from its environment (Mohri et al., 2012).

5.2 Classification task

The task of classification is described as an inference task in which we check if an object belongs to a category (Russell & Norvig, 2010). Such objects can take on many forms like images, sounds, or text files. Any algorithm that carries out a classification task can be considered a *classifier*. If an ML algorithm performs this task, we speak of a supervised learning task in which the elements in the target space $\mathcal{Y}$ can take on one of the values in $\mathcal{C} = \{c_1, \ldots, c_m\}$.

To give an example of a classification task, we have depicted Example 5.1 in Table 2. Observe that the class label of x_6 is unknown. A classification task would be to classify x_6 based on the features and the learned pattern from the other input-output pairs.

$\mathcal{X}$	g	t	m	c
x_1	8	1	0	1
x_2	7	0	0	0
x_3	6	1	1	1
x_4	8	1	1	1
x_5	7	0	1	0
x_6	6	1	1	?

Table 2: Example of a classification task.

5.3 Formal argumentation

Below we recall some definitions of formal argumentation.

> **Definition 5.3.** (Dung, 1995) An abstract argumentation framework (AF) is a pair $(\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a set of arguments, and $R \subseteq \mathcal{A} \times \mathcal{A}$ a set of attacks, such that $\forall a, b \in \mathcal{A}$ the relation $(a, b) \in \mathcal{R}$ means a attacks b.

Recall that central to abstract argumentation is the evaluation of sets of arguments, and that accepted sets of arguments are called *extensions*. We will focus on the *grounded extension*, which is defined below.

> **Definition 5.4.** (Dung, 1995) Let AF $= (\mathcal{A}, \mathcal{R})$ and let $S \subseteq \mathcal{A}$. S is said to be *conflict-free* iff there are no $a, b \in S$ such that a attacks b. S defends $a \in \mathcal{A}$ iff for each b that attacks a there is a $c \in S$ that attacks b. S is *admissible* iff it is conflict-free and defends all its arguments. S is a complete extension of AF iff it is admissible and contains all the arguments it defends. The **grounded extension** of AF $\mathbb{G}$(AF) is the least (w.r.t. $\subseteq$) complete extension.

Recall that arguments in abstract argumentation are seen as abstract entities. In this thesis, however, we require a more structured notion of an argument. For that purpose, we can turn to structured argumentation (Besnard et al., 2014). In structured argumentation, one can make the premises and conclusion of an argument explicit, which allows defining attacks based on the structure of the argument (Besnard et al., 2014). This serves the purpose of the explanation method that we will propose in Section 6; the attacks based on the structure of the arguments will make sure the method computes an AF that will accurately mimic the behavior of an ML model. Below we define our notion of an argument.

> **Definition 5.5.** An argument a is a pair $(\texttt{prem}, \texttt{conc})$, in which `prem` refers to a set of premises, and `conc` refers to the conclusion that the premises in `prem` infer. If an argument has a strength value linked to it, the argument is a triple $(\texttt{prem}, \texttt{conc}, \texttt{str})$ in which `str` $(0 \leq \texttt{str} \leq 1)$ refers to the strength value.

The notion of argument strength requires some extra attention. When modeling argument strength, it is important to be explicit about which type of argument strength is used, otherwise the interpretation of arguments and attacks can become nonsensical (Prakken, 2021). It is claimed there are three types of argument strength: logical, dialectical, and rhetorical (Prakken, 2021). Logical argument strength can be split up into inferential and contextual strength. Inferential strength signifies how well the premises support the conclusion and contextual argument strength indicates how well the conclusion is supported by other arguments. Dialectical strength indicates the strength of an argument in a discussion and rhetorical strength represents the persuasiveness of an argument. Below we define how, in this thesis, attacks rely on argument strength.

> **Definition 5.6.** An attack $(a, b) \in \mathcal{R}$ is defined as follows. Any argument a attacks another argument b iff $\texttt{conc}(a) \neq \texttt{conc}(b)$ and $\texttt{str}(a) \geq \texttt{str}(b)$.

This attack intuitively means that an argument attacks another argument if it is equally strong or stronger and has a different conclusion.

6 Everyday argumentative explanations

In this section, we wish to showcase how argumentation, as a basis for everyday explanations, has practical value for XAI. Recall that in Section 2 we described that XAI asks for understandable, adapted, and faithful explanations, and how everyday explanations help to create both understandability and adaptability but may lack faithfulness. To build on these findings, we present a method that has the following aim: *computing everyday explanations that remain faithful to the model they explain.*

With that aim in mind, we present everyday argumentative explanations, or *EVAX*, which is a model-agnostic, post-hoc method that computes AF-based explanations for decisions of ML classifiers. The explanations have contrastive, selected, and social characteristics; they include contrastive counterarguments, they consist of a fixed amount of arguments that can be selected based on a cognitive bias, and the size can be adjusted. In addition, the results in Section 7 show high and consistent fidelity scores, indicating a stable faithfulness.

6.1 Method outline

EVAX takes as input a labeled dataset, a trained black box model `BB`[1] and a threshold value τ_{select} that controls the size of the output. *EVAX* returns a set of predictions $\mathcal{Y}_{pred}$ and a set of local explanations $\mathcal{E}$. The explanations $e \in \mathcal{E}$ answer the question: "Why did black box `BB` assign class c to input instance x?" These explanations are deployments of an AF that represent the behavior of `BB` around a single datapoint in argumentative terms. This AF thus forms the basis for the explanations, and will, for every classified instance, be referred to as the local_AF. The size of this local_AF can be manually altered by τ_{select}. *EVAX* adopts a model-agnostic approach since it only uses the classifier as an oracle that can be queried for predictions. In that way, it behaves as a twin-system of the `BB`. The high-level mechanisms of *EVAX* are depicted as pseudo-code in Algorithm 1 and explained below.

Algorithm 1: *run_EVAX*(`BB`, labaled_dataset, $\tau_{select} = 20$)

```
X_train, X_test, Y_train, Y_test ← split_dataset(labeled_dataset, test_size = 0.2)
global_arguments ← get_global_arguments(BB, X_train)                    // step 1
for x_i in X_test do
    local_AF ← create_local_AF(x_i, global_arguments, τ_select)         // step 2
    predict(local_AF)                                                   // step 3
    explain(local_AF)                                                   // step 4
    save_results()
get_results(BB, predictions, Y_test)
```

First, *EVAX* divides the labeled dataset into a set of unlabeled datapoints $\mathcal{X}$ (the input space) and a set of labels $\mathcal{Y}$ (the target space), which are then split up into a train set and a test set, respectively $\mathcal{X}_{train}, \mathcal{X}_{test}$ and $\mathcal{Y}_{train}, \mathcal{Y}_{test}$. The default size of the test set is set at 0.2, and the default τ_{select} value is set at 20. Afterward, the method can be divided into four main steps, which are described below. The first step handles all datapoints and is executed just once, whereas the other three steps handle a single datapoint and may be repeated multiple times, up to a maximum of the size of the test set. The mechanisms of these final three steps are visualized in Figure 6.

- **Step 1: Extract a global list of arguments.**
  - This step, represented in Algorithm 2, involves extracting a list of arguments that represent the global behavior of `BB`.

[1] `BB` can be any ML classifier from the scikit-learn library (Pedregosa et al., 2011). Because *EVAX* only requires the input and output of the ML models, however, other ML libraries can be used as well.

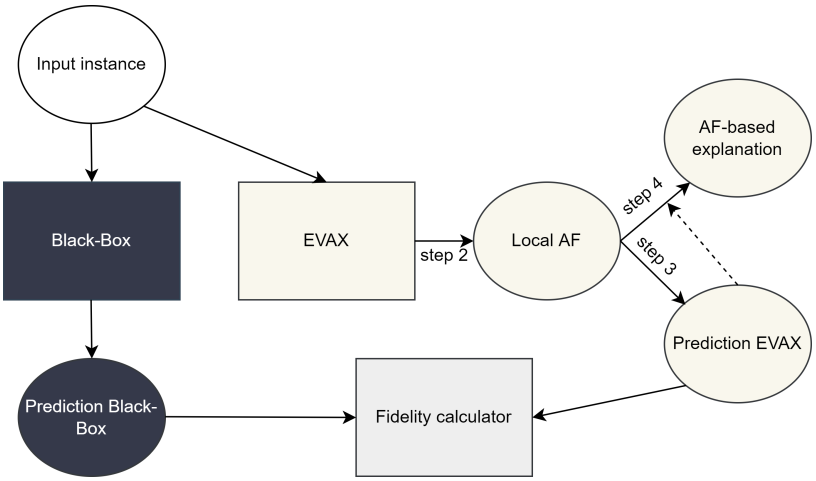

Figure 6: Mechanisms of Steps 2, 3 and 4 of *EVAX*. The dashed arrow from 'prediction EVAX' to the 'step 4' arrow shows how the predicted output class is incorporated in the construction of the AF-based explanation.

- *EVAX* first iterates over all features $(a_n, v_n) \in \mathcal{F}$ of all (unlabeled) datapoints $x_i \in \mathcal{X}_{train}$ and all output classes $c_i \in \mathcal{C}$, and computes for every feature-class pair a decision rule. These rules are accompanied by a precision score $\mathcal{P}$ that articulates the probability that `BB` will assign a datapoint with that particular feature to that particular class. It then saves all $\mathcal{P}$ scores in a triple $((a_n, v_n), c_i, \mathcal{P})$, which is added to a list of triples. To avoid the inclusion of duplicates, *EVAX* first checks if the feature-class pair is not yet in the list.
- Arguments are constructed based on the list of triples. For every triple an argument is constructed in which the feature is the premise `prem`, the output-class c is the conclusion `conc` and the precision score $\mathcal{P}$ is set as the argument strength `str`. Together these arguments form the global list of arguments.

Algorithm 2: get_global_arguments

```
for x_i in X_train do
    for feature (a_n, v_n) in x_i do
        for class c_i in C do
            if ((a_n, v_n), c_i) is new then
                nf ← # of datapoints with this feature
                nfc ← # of datapoints with this feature and class
                P ← nfc / nf
                triple ← ((a_n, v_n), c_i, P)
                prem, conc, str ← triple
                add_argument_to_global_list(prem, conc, str)
```

- **Step 2: Create a local AF.**
  - In the second step, *EVAX* creates a local_AF in every iteration of this step. This is an AF $(= (\mathcal{A}, \mathcal{R}))$ that represents the classifier's behavior around one particular datapoint. Based on the values of that datapoint, it selects a set of relevant arguments $(\mathcal{A})$ from the global list of arguments and defines attacks over them $(\mathcal{R})$. This step is represented with Algorithm 3.

- The argument selection is done by matching the features of the datapoint from $\mathcal{X}_{test}$ with the premises of the arguments. To be exact, given a datapoint x_i and an argument $a \in \mathcal{A}$, if value v_n of a feature $(a_n, v_n) \in \mathcal{F}$ from datapoint x_i is equal to $\texttt{prem}(a)$, a is added to the local AF (meaning $a \in \mathcal{A}$). As a result, all arguments with a premise corresponding to one of the features of the datapoint are selected. To give an example, consider datapoint x_1 has feature (age, old), where $a_1 = age$ and $v_1 = old$. The list of relevant arguments is equal to all arguments a with $\texttt{prem}(a) = (age, old)$. To gain computational efficiency and maintain selectedness, a threshold τ_{select} can be defined, which ensures only the top τ_{select} strongest arguments are included in the list.
- The attacks are defined based on Definition 5.6. It implies that an argument attacks another if it is equally strong or stronger and has a different conclusion.

Algorithm 3: create_local_AF

```
local_AF ← create_empty_AF()
for feature (a_n, v_n) in x_i do
    for argument in global_arguments do
        if prem(argument) == v_n then
            add_argument_to_local_AF(argument, local_AF, τ_select)
for argument a in local_AF do
    for argument b in local_AF do
        if conc(a) ≠ conc(b) and str(a) ≥ str(b) then
            add_attack(local_AF, (a, b))
```

- **Step 3: Predict**
  - The third step, which is shown in Algorithm 4, amounts to predicting the output class c using the local_AF from Step 2. First, the grounded extension of the local_AF ($\mathbb{G}$(local_AF)) is computed, after which the conclusion of the arguments in $\mathbb{G}$(local_AF) is picked as the prediction. Formally, this means that prediction $y_i \in \mathcal{Y}$ is equal to $\texttt{conc}(a)$ such that $a \in \mathbb{G}$(local_AF). Since arguments in the grounded extension are non-conflicting, they always have the same conclusion. Therefore it does not matter what argument in $\mathbb{G}$(local_AF) is picked. When $\mathbb{G}$(local_AF) is empty, *EVAX* will predict the majority class.

Algorithm 4: predict(local_AF)

```
𝔾 ← get_grounded_extension(local_AF)
if 𝔾 = ∅ then
    y ← majority_class
y ← conc(a) | a ∈ 𝔾
```

- **Step 4: Explain**
  - An explanation $e_i \in \mathcal{E}$ is a deployment of local_AF and answers the question: 'Why did BB assign x_i to class c_i?' Since there are various ways in which a user can deploy an AF, as we have shown in Section 3.4, we provide two examples of how the local_AF, which is an AF, can be deployed. The first example incorporates biasedness and the second one has a conversational form.

– A first way in which *EVAX* can deploy an explanation is by selecting an argument from the local_AF based on the cognitive bias of *abnormality*. This bias describes how people tend to choose a cause that is unusual and is claimed to play a key role in human explanations (Thagard, 1989). We have defined the abnormality of an argument as $1-coverage$. The coverage value refers to the fraction of datapoints that the decision rule, out of which the argument is composed, 'rules over'. In other words, the coverage of argument a refers to the fraction of input instances that have a feature equal to $\texttt{prem}(a)$. Since the coverage describes how often a feature is present in a dataset, it essentially describes how 'normal' a feature is. Therefore, a lower coverage means that a feature becomes less normal, thus becomes increasingly abnormal. Therefore we define abnormality as $1-coverage$. The deployment of the local_AF then amounts to selecting the argument with the highest abnormality score that argues for the predicted class. An example of the output is given in Figure 7. It explains why BB assigned x_1 (a mushroom) from the mushroom dataset[2] to class c_1 (poisonous).

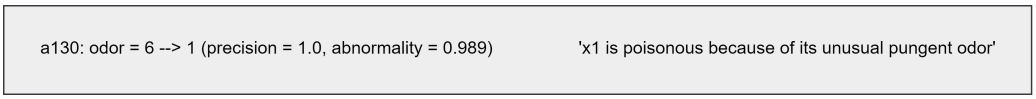

Figure 7: Example output of the most abnormal argument of the local_AF that explains why BB assigned x_1 (a mushroom) to class c_1 (poisonous). On the right, we see the same explanation, but in a more readable form.

– *EVAX* can also provide a dialectical representation of the local_AF, similar to a dispute tree (Fan & Toni, 2015). This representation has the form of a discussion between a proponent (P) and opponent (O) about what class to assign to the datapoint in question. A threshold $\tau_{explain}$ allows the user to choose the number of arguments to include in the explanation. Arguments are split up into *pro* and *con* arguments and are put forward by P and O, who take turns. If the value of threshold $\tau_{explain}$ is even, O starts the dispute, and if it is odd, P starts. After the first argument is put forward, the strongest counterargument is replied.[3] Note that the threshold is different from τ_{select}, because it does not affect the size of the local_AF, but merely the size of the dialectical representation of the local_AF. In Figure 8 we see an example of an explanation for why BB assigned x_1 (a mushroom) of the mushroom dataset to class c_1 (poisonous), with $\tau_{explain} = 4$.

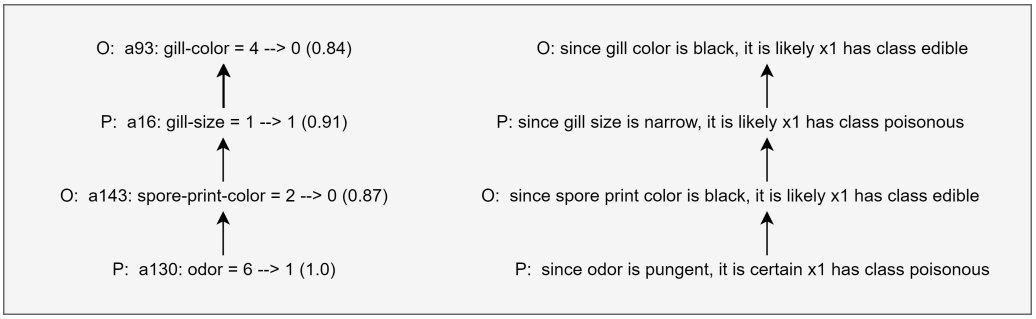

Figure 8: The dialectical explanation of the assignment of x_1 to c_1 by BB. The values between brackets refer to the precision score. On the right, we see the same explanation, but in a more readable form. One must read from top to bottom; the arrows solely indicate the attacks.

[2]This dataset is described in Section 7.1.

[3]The attacks in the explanation indicate a conflicting conclusion, and not necessarily a higher strength value. This is to ensure that counterarguments are included in this explanation form.

6.2 Notions on modelling choices

Since we have based many of the modeling choices on former sections of this thesis and relevant literature, there is a need to clarify some of these choices.

- **Argument construction.** Arguments are constructed out of decision rules, as described in Step 1. The underlying reason is that decision rules reflect the functioning of a model and have an argument-like structure (Vilone & Longo, 2021). Whereas an argument consists of premises that imply the conclusion, a decision rule has an IF-THEN structure in which the truth value of a Boolean expression (the IF-part) determines the outcome (the THEN-part) (Hailesilassie, 2016). Arguments and decision rules are similar because they both describe the conditions that imply a certain outcome. Note that in our case the arguments are accompanied by a strength value (`str`).

- **Argument strength.** As stated in Section 5, it is important to be explicit about which aspect of argument strength is modeled. To be clear, *EVAX* makes use of inferential argument strength. The `str` value of arguments in *EVAX* is computed as the precision score of a decision rule, which specifies the likeliness that the premises imply the conclusion. In that way, a higher strength value entails that the premises more strongly support the conclusion, which indicates inferential strength.

- **Argument selection for local_AF.** The selection process of arguments from the global list of arguments to the local_AF is based on two criteria: (1) at least one feature of the input instance should correspond with the `prem` of the argument, and (2) the argument should be part of the top τ_{select} strongest arguments.

  The first criterion is needed because if the premise of an argument does not correspond to at least one feature of the input instance, the argument does not say anything about that instance and thus remains irrelevant.

  The second criterion assumes that stronger arguments are more relevant. Argument strength describes the precision of a decision rule and thus refers to how accurately it describes the models' statistical behavior. This selection criterion then thus essentially comes down to the likelihood or truthfulness of the argument. Based on the strength, the top τ_{select} arguments are selected. Note that this is not how we have incorporated biasedness in *EVAX*. Explaining through the cognitive bias of abnormality happens *after* creating the local_AF.

- **Feature independence assumption.** Because the argument strength scores are calculated based on the precision scores of decision rules, they essentially describe the correlation between a single feature and the output class. In that correlation, interactions between different features are not included; it assumes independence among them. We have performed an initial experiment in which we include feature interactivity, which is described in Appendix A.1. It investigates the impact of such interactivity on the performance of *EVAX*. Since the naive approach achieves significantly better performance, we have decided not to take into account feature interactivity. Nevertheless, when performance decreases as a result of a strong correlation between features, there may be a need to further investigate this matter.

6.3 Toy Example

Recall the classification task described in Example 5.1. The task is to predict if a student will be accepted into university. In Figure 9 we have presented a similar case to further exemplify the workings of *EVAX*. It represents one iteration of Steps 2, 3, and 4. It thus assumes that the global list of arguments has already been computed.

In this example, a black box predicts that an input instance, 'John' in this case, will be accepted into university. The same input instance is used as input for *EVAX*. Based on that input, *EVAX* creates a local_AF by selecting three relevant arguments, based on the three different features, and

defines attacks over them. It then calculates the grounded extension and predicts that John will be accepted into university. In addition, it computes an AF-based explanation, which in this case is a dialectical representation of the local_AF, as described in Step 4. The threshold $\tau_{explain}$ has a value of 3. The arrows in the representation represent the attacks. Note that the arrows between the different components of *EVAX do not* represent attacks, but indicate the information flow.

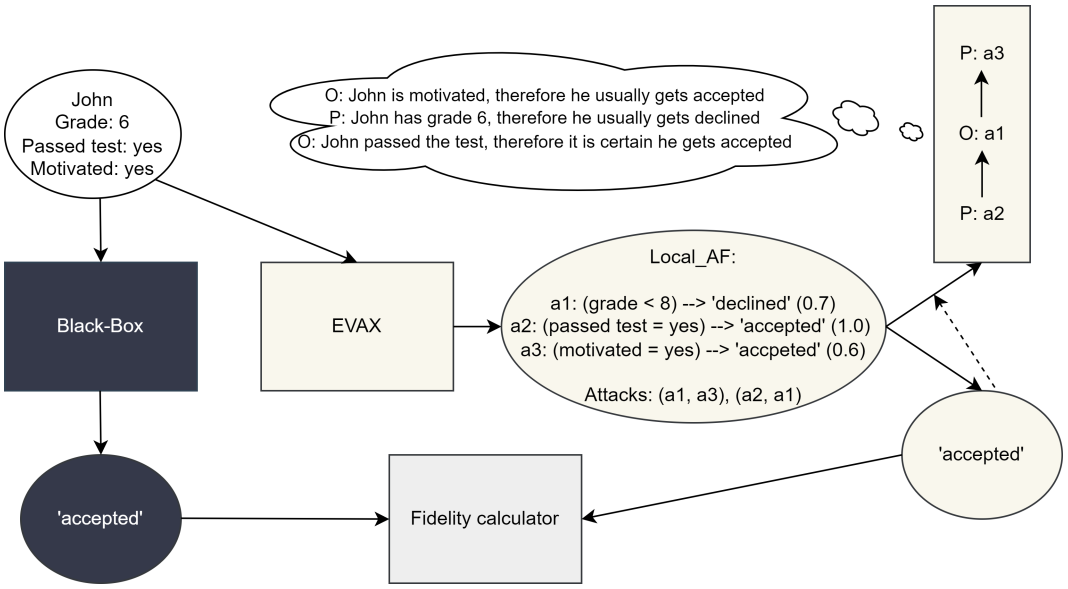

Figure 9: Toy example of *EVAX*.

7 Quantitative evaluation

In this section, we evaluate the performance of *EVAX* using four datasets and five quantitative metrics.

7.1 Data-sets

We use four labeled datasets from the UCI Machine Learning Repository (Dua & Graff, 2017) which are described below and summarized in Table 3.

- With the **Adult** dataset one tries to predict whether or not a person makes more than 50.000 dollars a year. We removed all datapoints with unknown values and discretized the continuous features.
- The **Mushroom** dataset includes instances of 23 different species of mushrooms. The task is to identify whether a mushroom is poisonous or edible. We did not perform any alterations on this dataset.
- The **Iris** dataset is perhaps the best-known dataset in the pattern recognition literature. The task is to predict the type of iris plant. We discretized the continuous values.
- With the **Wine** dataset one wants to predict the type of wine of an input instance. Again we discretized the continuous values.

	***n* of instances**	***n* of attributes**	***n* of classes**	**Attribute type**
Adult	48842	14	2	Mixed
Mushroom	8124	22	2	Categorical
Iris	150	4	3	Numerical
Wine	178	13	3	Numerical

Table 3: Dataset specifications

The discretization of continuous variables is necessary to constrain the number of arguments that are added to the global list of arguments. We have used the *cut* method by pandas (Wes McKinney, 2010) with a bin value of 10. Since higher bin values tend to give better performance but reduce the computational efficiency, we have tuned this value by incrementally increasing the value from 3 up to 20. We found that from a bin value of 10 and upwards, the fidelity did not significantly increase (sometimes it even decreased), while the computational efficiency consistently decreased with higher bin values.

7.2 Black boxes

We have chosen four different ML models with different complexity to test the performance and range of *EVAX*. These models are logistic regression, support vector machines (SVM), random forest, and neural networks. Whereas logistic regression is often considered a relatively understandable model, the other three are generally considered as being black boxes (Adadi & Berrada, 2018). We have included logistic regression to incorporate a wider complexity range. All four models are initialized from the scikit-learn library (Pedregosa et al., 2011).

7.3 Metrics

We use five different evaluation metrics:

- **Fidelity** indicates how well the explanation approximates the prediction of the black box model. It represents the fraction of datapoints that are assigned to the same output class by *EVAX* and `BB`. In a sense, fidelity describes how faithful the explanation is. It provides a quantitative indication of the completeness side of the understandability trade-off. That is to say, a high fidelity signifies a faithful explanation.
- **Accuracy** *(`BB`)* indicates how well our model performs on unseen data. It represents the fraction of correctly classified datapoints. The value between brackets () refers to the original accuracy of `BB`.
- **Size** measures the average minimum amount of arguments necessary to retain the same prediction. In other words, it is the lowest possible τ_{select} score without affecting the accuracy or fidelity. A consistent low size value indicates the method can guarantee to compute small explanations that are consistently faithful.
- **Empty** $\mathbb{G}$ specifies the fraction of datapoints for which the grounded extension $\mathbb{G}$(local_AF) is an empty set. When $\mathbb{G}$(local_AF) is an empty set, *EVAX* relies on a default prediction. A higher 'Empty $\mathbb{G}$' value thus means that accuracy and fidelity scores are increasingly determined by the default prediction, and therefore become less reliable.
- **Time** indicates the number of seconds needed to run the program.

		Fidelity	**Accuracy** *(`BB`)*	**Size**	**Empty** $\mathbb{G}$	**Time (s)**
Adult	*Logistic regression*	0.95	0.73 *(0.72)*	1	0.0	5.06
	SVM	0.93	0.75 *(0.74)*	1	0.0	13.61
	Random forest	0.88	0.77 *(0.78)*	1	0.0	5.06
	Neural network	0.91	0.75 *(0.75)*	1	0.0	5.28
Mushroom	*Logistic regression*	0.98	0.96 *(0.95)*	1	0.0	17.92
	SVM	0.99	0.98 *(0.99)*	1	0.0	13.45
	Random forest	1.0	0.99 *(1.0)*	1	0.0	13.63
	Neural network	1.0	0.95 *(0.95)*	1	0.0	17.87
Iris	*Logistic regression*	0.97	0.97 *(0.9)*	1	0.0	0.24
	SVM	1.0	0.97 *(0.97)*	1	0.0	0.25
	Random forest	1.0	0.97 *(0.97)*	1	0.0	0.26
	Neural network	0.93	0.97 *(0.9)*	1	0.0	0.23
Wine	*Logistic regression*	0.94	1.0 *(0.94)*	1	0.0	2.32
	SVM	0.94	1.0 *(0.94)*	1	0.0	2.60
	Random forest	0.92	1.0 *(0.91)*	1	0.0	2.73
	Neural network	0.86	0.97 *(0.83)*	1	0.0	3.86

Table 4: Quantitative results of *EVAX*.

7.4 Results

Recall that the goal of *EVAX* is to produce everyday explanations while retaining decent faithfulness. The results in Table 4 show high fidelity (an average of 0.95) for all four ML models, which indicates a sufficient degree of faithfulness. Only the adult dataset and the neural network of the wine dataset have relatively low scores. This might be due to the relatively low accuracy

of the BB in those cases. Since the argument with the highest argument strength is always in $\mathbb{G}$(local_AF), the minimum size is always equal to 1. This indicates the model is capable to compute small explanations without losing faithfulness. Assuming that smaller explanations are more understandable, these results indicate that *EVAX* can find a satisfactory balance on the understandability trade-off. Also, we see that the method never computes an empty grounded extension $\mathbb{G}$(local_AF), and hence requires no reliance on a default prediction. These results are obtained on a Windows 64-bit operating system with 16GB RAM and an Intel(R) Core(TM) i5-1145G7 @ 2.60GHz processor.

8 Qualitative evaluation

In this section, we provide a qualitative evaluation of the explanations computed by *EVAX*. The theoretical findings from Section 2 and 3 will form a basis for this evaluation. First, we match the explanations with the definition of an explanation. Afterward, we assess if the explanations sufficiently contain the properties of everyday explanations. Lastly, we discuss if they help to achieve the objectives of XAI.

8.1 Conforming to the definition

The explanations computed by *EVAX* clearly satisfy the definition of an explanation given in Section 2.1. This definition describes how an explanation is an answer to a *why-question*. The computed explanations *do* answer a why-question, which is 'why did `BB` assign this class to datapoint x_i?' The cognitive process, the product, and the social process, which are inherent to the definition, can also be traced back to our explanations. The cognitive process takes place at the algorithmic level and refers to the process of assembling the relevant arguments from the global list of arguments to the local_AF, the product is the deployment of the local_AF, and the social process involves the various ways in which users can adjust the explanation.

8.2 Conforming to everyday explanations

Recall the definitions of the properties of everyday explanations in the context of argumentative XAI, given in Section 4. First, argumentative explanations are **contrastive** when they include arguments *pro* and *con* the conclusion. For explanations computed by *EVAX* this is the case when there is at least one argument with a fact conclusion and a counterargument with a foil conclusion. Such counterarguments enable the explanation to explain the outcome *relative to* an alternative outcome, by showing what features give reason to believe that foil. In Figure 8 we see the use of such arguments.

Note that these counterarguments are *not* counterfactual statements: they do not specify the conditions under which the current output class would *change* to a different output class. An interesting direction for future work, however, is to include such statements. This could be done by searching for an argument in the global list of arguments that would change the conclusions in the grounded extension, hence changing the prediction of *EVAX*. The AF containing that argument would then be the counterfactual situation and the added argument would describe the necessary condition that would change the outcome. Such counterfactual statements can be articulated as: "If this argument would have been included in the local_AF, the prediction would have been Q instead of P."

Second, *EVAX* incorporates **selectedness** by implementing both minimality and biasedness. Minimality amounts to including just a few arguments as the explanation. This is enabled by guaranteeing that the number of arguments in the local_AF does not exceed threshold τ_{select}. In addition, this restricted size has shown not to affect the fidelity score. *EVAX* does also include biasedness because it allows us to compute explanations based on *abnormality*, which is a common cognitive bias in everyday explanations (Thagard, 1989). It amounts to presenting the argument with the highest abnormality score in the local_AF. This size restriction and the biasedness sets *EVAX* apart from the discussed methods in Section 4.

Third, the computed explanations are **social** when they can be adapted to the explainee. Adaptation in *EVAX* can take shape in two manners. First, the number of arguments that are included in the local_AF can be adjusted with τ_{select}. In that way, an inexperienced end-user that requires a single argument to explain the prediction can set τ_{select} at a value of 1. A more experienced user that wants a completer chain of causes can set a higher value. Second, because a computed explanation $e \in \mathcal{E}$ stems from an AF ($= (\mathcal{A}, \mathcal{R})$), the explanation can be presented in various ways. We have shown two examples, visualized in Figures 7 and 8. There are, however, many other ways in which one can deploy an AF, as described in Section 3.4.

8.3 Discussion

Recall that the objectives of XAI involve creating understandable, adapted, and faithful explanations. The results from our quantitative and qualitative evaluation suggest that the explanations computed by *EVAX* assist in reaching these objectives. Since they have the characteristics of everyday explanations, they increase in understandability. Inherent to this fact is that they are social, which describes how adaptability is incorporated. The quantitative results showed consistent fidelity scores, meaning the explanations remain relatively faithful to the model they explain. In that way, they have a focus on all three components of the objective, which highlights their value for XAI.

Nevertheless, there is room for refinements and further investigation. First of all, the contrastive property is now interpreted as the inclusion of reasons *pro* and *con* a conclusion. To better align with more common notions on contrastive or counterfactual explanations (Stepin et al., 2021; van der Waa et al., 2018), there is a need for experimenting with statements that describe the necessary conditions to *change* the fact to the foil. Second, we have only experimented with one cognitive bias. To further enhance selectedness, it may be useful to see to what extent other cognitive biases or combinations of them lend themselves to be formalized and used for argumentative explanations. Regarding the social property, we have only presented one way in which the computed AF is deployed. This social quality will become more apparent when the applicability of other explanation deployments is further investigated. Finally, a user study on the understandability of the explanations computed by *EVAX* can enrich the empirical insights into their quality.

9 Conclusion

The central goal of this thesis has been to gain understanding of the value of argumentation for XAI. We have approached this by using everyday explanations as a frame of reference. First, we showed how XAI asks for understandable, adaptable, and faithful explanations and discussed the contrastive, selected and social character of everyday explanations. Thereafter, we conceptually showed that argumentative explanations can be posed as everyday explanations and that argumentation has a social and cognitive character. Whilst reviewing current methods, we showed that computed explanations already contain contrastive and social characteristics, but tend to lack selectedness. By presenting *EVAX*, we showed how argumentative explanations can show all characteristics of everyday explanations while remaining high fidelity scores.

The key finding is that argumentation has value for XAI through its ability to produce faithful, everyday explanations. The cognitive character of everyday argumentative explanations describes how they become understandable to humans, their strong social sense enables adaptability and the fidelity scores of *EVAX* show that faithfulness can be guaranteed. This shows how argumentation helps in achieving the objectives of XAI, which indicates its value.

Note, however, that this thesis does not provide a full overview of that value; it can be extended in multiple ways. First, the properties of everyday explanations can be further worked out by including counterfactual statements, incorporating more cognitive biases, and testing more explanation deployments. Second, experimental evaluations with human users would more closely assess the quality of argumentative explanations. Finally, the understanding of the value of argumentation may be enlarged by investigating it from another frame of reference (other than everyday explanations). Given that everyday explanations mainly apply to local explanations for end-users, having a framework that allows for evaluating global explanations and explanations for target audiences with more expertise may uncover the existence of a broader reach of the value of argumentation.

Nevertheless, the optimistic words about argumentative explanations turn out to be legitimate; argumentation has shown its usefulness in multiple ways. Given the immature state of this approach, it has the potential to grow into playing a vital role in the mission for understandable AI. We hope the findings in this thesis motivate scientists and practitioners to continue to explore the value of argumentation for XAI.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Albini, E., Lertvittayakumjorn, P., Rago, A., & Toni, F. (2020). DAX: deep argumentative explanation for neural networks. *arXiv*, *2012.05766*.

Amgoud, L. (2021). Non-monotonic explanation functions. In J. Vejnarová & N. Wilson (Eds.), *Proceedings of the 16th European Conference of Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU, 2021* (Vol. 12897, pp. 19–31). Springer.

Antaki, C., & Leudar, I. (1992). Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, *22*(2), 181–194.

Arioua, A., & Croitoru, M. (2015). Formalizing explanatory dialogues. In C. Beierle & A. Dekhtyar (Eds.), *Proceedings of the 9th International Conference of Scalable Uncertainty Management, SUM, 2015* (Vol. 9310, pp. 282–297). Springer.

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Arya, V., Bellamy, R. K. E., Chen, P., Dhurandhar, A., Hind, M., Hoffman, S. C., ... Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv*, *1909.03012*.

Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., ... Villata, S. (2017). Towards artificial argumentation. *AI Magazine*, *38*(3), 25–36.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... Zhang, Y. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15.

Bench-Capon, T. J. M., & Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial Intelligence*, *171*(10-15), 619–641.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science education*, *93*(1), 26–55.

Besnard, P., García, A. J., Hunter, A., Modgil, S., Prakken, H., Simari, G. R., & Toni, F. (2014). Introduction to structured argumentation. *Argument & Computation*, *5*(1), 1–4.

Besnard, P., & Hunter, A. (2008). *Elements of argumentation*. MIT Press.

Bex, F., & Walton, D. (2016). Combining explanation and argumentation in dialogue. *Argument & Computation*, *7*(1), 55–68.

Bielaczyc, K., & Blake, P. (2006). Shifting epistemologies: Examining student understanding of new models of knowledge and learning. In *Proceedings of the International Conference of the Learning Sciences, ICLS, 2006* (Vol. 1, pp. 50–56). International Society of the Learning Sciences.

Bondarenko, A., Dung, P. M., Kowalski, R. A., & Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, *93*, 63–101.

Borg, A., & Bex, F. (2021a). A basic framework for explanations in argumentation. *IEEE Intelligent Systems*, *36*(2), 25–35.

Borg, A., & Bex, F. (2021b). Necessary and sufficient explanations for argumentation-based conclusions. In J. Vejnarová & N. Wilson (Eds.), *Symbolic and quantitative approaches to reasoning with uncertainty* (pp. 45–58). Springer.

Brandão, R., Carbonera, J., de Souza, C. S., Ferreira, J. J., Gonçalves, B., & Leitão, C. F. (2019). Mediation challenges and socio-technical gaps for explainable deep learning applications. *arXiv*, *1907.07178*.

Brarda, M. E. B., Tamargo, L. H., & García, A. J. (2019). An approach to enhance argument-based multi-criteria decision systems with conditional preferences and explainable answers. *Expert Systems with Applications*, *126*, 171–186.

Briguez, C. E., Budán, M. C., Deagustini, C. A. D., Maguitman, A. G., Capobianco, M., & Simari, G. R. (2014). Argument-based mixed recommenders and their application to movie suggestion. *Expert Systems with Applications*, *41*(14), 6467–6482.

Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In S. Kraus (Ed.), *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI, 2019* (pp. 6276–6282).

Calegari, R., Ciatto, G., & Omicini, A. (2020). On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale*, *14*(1), 7–32.

Cambridge. (2021a). *Argument.* Cambridge University Press. Retrieved from `https://dictionary.cambridge.org/dictionary/english/argument`

Cambridge. (2021b). *Argumentation.* Cambridge University Press. Retrieved from `https://dictionary.cambridge.org/dictionary/english/argumentation`

Cambridge. (2021c). *Counterargument.* Cambridge University Press. Retrieved from `https://dictionary.cambridge.org/dictionary/english/counterargument`

Cocarascu, O., Stylianou, A., Čyras, K., & Toni, F. (2020). Data-empowered argumentation for dialectically explainable predictions. In G. D. Giacomo et al. (Eds.), *Proceedings of the 24th European Conference on Artificial Intelligence, ECAI, 2020* (Vol. 325, pp. 2449–2456). IOS Press.

Cocarascu, O., & Toni, F. (2016). Argumentation for machine learning: A survey. In P. Baroni, T. F. Gordon, T. Scheffler, & M. Stede (Eds.), *Proceedings of Computational Models of Argument, COMMA, 2016* (Vol. 287, pp. 219–230). IOS Press.

Čyras, K., Badrinath, R., Mohalik, S. K., Mujumdar, A., Nikou, A., Previti, A., ... Feljan, A. V. (2020). Machine reasoning explainability. *arXiv*, *2009.00418*.

Čyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., ... Hapuarachchi, T. (2019). Explanations by arbitrated argumentative dispute. *Expert Systems with Applications*, *127*, 141–156.

Čyras, K., Rago, A., Albini, E., Baroni, P., & Toni, F. (2021). Argumentative XAI: A survey. In Z. Zhou (Ed.), *Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI, 2021* (pp. 4392–4399).

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, 2019* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*, *1702.08608*.

Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, *77*(2), 321–358.

Espinoza, M. M., Tacla, C. A., & Jasinski, H. M. R. (2020). An argumentation-based approach for explaining goals selection in intelligent agents. In R. Cerri & R. C. Prati (Eds.), *Proceedings of the 9th Brazilian Conference of Intelligent Systems, BRACIS, 2020* (Vol. 12320, pp. 47–62). Springer.

European Commision. (2021). Europe fit for the digital age: Commission proposes new rules and actions for excellence and trust in artificial intelligence. *Europan Commision: Geneva, Switzerland*.

European Commission. (2018). *2018 reform of EU data protection rules.* Retrieved from `https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf`

Fan, X., & Toni, F. (2015). On computing explanations in argumentation. In B. Bonet & S. Koenig (Eds.), *Proceedings of the 29th Conference on Artificial Intelligence, AAAI, 2015* (pp. 1496–1502). AAAI Press.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

García, A. J., & Simari, G. R. (2004). Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming*, *4*(1-2), 95–138.

Gerlings, J., Shollo, A., & Constantiou, I. D. (2021). Reviewing the need for explainable artificial intelligence (xai). In *Proceedings of the 54th Hawaii International Conference on System Sciences, HICSS, 2021* (pp. 1–10). ScholarSpace.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. (2019). XAI - explainable artificial intelligence. *Science Robotics*, *4*(37).

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological review*, *114*(3), 704.

Hailesilassie, T. (2016). Rule extraction algorithm for deep neural networks: A review. *arXiv*, *1610.05267*.

Halpern, J. Y. (2011). Causality, responsibility, and blame: A structural-model approach. In S. Benferhat & J. Grant (Eds.), *Proceedings of the 5th International Conference of Scalable Uncertainty Management, SUM, 2011* (Vol. 6929, p. 1). Springer.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65.

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, *2*(4), 273–308.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, *93*(1), 75.

Hogan, K., Nastasi, B. K., & Pressley, M. (1999). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition and instruction*, *17*(4), 379–432.

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020* (pp. 4198–4205). Association for Computational Linguistics.

Josephson, J. R., & Josephson, S. G. (1996). *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Kenny, E. M., & Keane, M. T. (2019). Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ANN-CBR twins for XAI. In S. Kraus (Ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI, 2019* (pp. 2708–2715).

Kuhn, D. (1991). *The skills of argument*. Cambridge University Press.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1).

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247–266.

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press.

Mayes, G. R. (2000). Resisting explanation. *Argumentation*, *14*(4), 361–380.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, *34*(2), 57–74.

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

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

Mittelstadt, B. D., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In D. Boyd & J. H. Morgenstern (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT, 2019* (pp. 279–288). ACM.

Modgil, S., & Prakken, H. (2014). The $ASPIC^+$ framework for structured argumentation: a tutorial. *Argument & Computation*, *5*(1), 31–62.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. MIT Press.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3-4), 1–45.

Oren, N., van Deemter, K., & Vasconcelos, W. W. (2020). Argument-based plan explanation. In M. Vallati & D. E. Kitchin (Eds.), *Knowledge Engineering Tools and Techniques for AI Planning* (pp. 173–188). Springer.

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *arXiv*, *1907.12652*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argument & Computation*, *1*(2), 93–124.

Prakken, H. (2021). Philosophical reflections on argument strength and gradual acceptability. In J. Vejnarová & N. Wilson (Eds.), *Proceedings of the 16th European Conference of Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU, 2021* (Vol. 12897, pp. 144–158). Springer.

Prakken, H., & Ratsma, R. (2021). A top-level model of case-based argumentation for explanation: Formalisation and experiments. *Argument & Computation*, 1–36.

Rago, A., Cocarascu, O., Bechlivanidis, C., & Toni, F. (2020). Argumentation as a framework for interactive explanations for recommendations. In D. Calvanese, E. Erdem, & M. Thielscher (Eds.), *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR, 2020* (pp. 805–815).

Raymond, A., Gunes, H., & Prorok, A. (2020). Culture-based explainable human-agent deconfliction. In A. E. F. Seghrouchni, G. Sukthankar, B. An, & N. Yorke-Smith (Eds.), *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS, 2020* (pp. 1107–1115). International Foundation for Autonomous Agents and Multiagent Systems.

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Proceedings of the 28th Annual Conference on Neural Information Processing Systems, 2015* (pp. 91–99).

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence - A modern approach, third international edition.* Pearson Education.

Sawyer, R. K. (2005). *The cambridge handbook of the learning sciences.* Cambridge University Press.

Sendi, N., Abchiche-Mimouni, N., & Zehraoui, F. (2019). A new transparent ensemble method based on deep learning. In I. J. Rudas, J. Csirik, C. Toro, J. Botzheim, R. J. Howlett, & L. C. Jain (Eds.), *Proceedings of the 23rd International Conference of Knowledge-Based and Intelligent Information & Engineering Systems, KES, 2019* (Vol. 159, pp. 271–280). Elsevier.

Šešelja, D., & Straßer, C. (2013). Abstract argumentation and explanation applied to scientific debates. *Synthese*, *190*(12), 2195–2217.

Simari, G. R., & Rahwan, I. (Eds.). (2009). *Argumentation in artificial intelligence.* Springer.

Sklar, E. I., & Azhar, M. Q. (2018). Explanation through argumentation. In M. Imai, T. Norman, E. Sklar, & T. Komatsu (Eds.), *Proceedings of the 6th International Conference on Human-Agent Interaction, HAI, 2018* (pp. 277–285). ACM.

Sokol, K., & Flach, P. A. (2020). One explanation does not fit all. *Künstliche Intelligenz*, *34*(2), 235–250.

Stepin, I., Alonso, J. M., Catalá, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, *9*, 11974–12001.

Thagard, P. (1989). Explanatory coherence. *Behavioral and brain sciences*, *12*(3), 435–467.

Tripathi, K. (2011). A study of interactivity in human computer interaction. *International Journal of Computer Applications*, *16*(6), 1–3.

van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., & Neerincx, M. A. (2018). Contrastive explanations with local foil trees. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, WHI, 2018* (Vol. 32).

Vassiliades, A., Bassiliades, N., & Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, *36*.

Vilone, G., & Longo, L. (2021). A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. *Frontiers in Artificial Intelligence*, *4*, 717899.

Walton, D. N. (2006). *Fundamentals of critical argumentation*. Cambridge University Press.

Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56 – 61).

A Appendix

A.1 Including feature interactivity

Since the argument strength values of *EVAX* assume independence among features, we have experimented with taking into account interactivity among features by altering the strength value according to their interactions. Even though this is not included in *EVAX*, it is interesting to elaborate on the used approach and the results of this experiment.

We have experimented with feature interactivity by altering the argument strength score based on the permutation-based feature importance of the feature that is included in the argument. We have used the permutation_importance function from the scikit-learn library (Pedregosa et al., 2011). The permutation importance of a feature is determined by calculating the increase in the model's prediction error after permuting the feature. A feature is 'important' if shuffling its values increases the model error because it indicates that the model relied on the feature for the prediction. Because permuting a feature also cancels out all interactions with other features, this importance measure directly takes into account feature interactions (Fisher, Rudin, & Dominici, 2019).

This type of feature importance refers to a different definition of a feature than the one used in this thesis. We refer to a feature as an attribute-value pair, whereas permutation-based feature importance techniques refer to features as the attribute itself (as consisting of the full range of values). Therefore, the computed importance scores indicate the impact of all possible values of a feature together for the output, regardless of the output class. The importance of separate attribute-value pairs for separate output classes is therefore ignored. Arguments in *EVAX*, however, are composed of specific attribute-value pairs arguing for a particular output class, and therefore do not have a one-to-one correspondence to the permutation-based feature importance scores.

To overcome this problem, we have created dummy variables for every different feature value and one-hot encoded them. The result is that we can now, for every single feature-value pair, compute a corresponding permutation feature importance score. To incorporate this score into the arguments, we have multiplied the strength value `str` of every argument with this permutation score. In that way, arguments with more important features get a higher score. We have set τ_{select} at a value of 20.

The results depicted in Table 5 show the impact of these altered argument strength scores. We observe a clear drop in fidelity when the permutation score is included. This may be due to low feature interactivity in the model, or a mismatch between the strength values and the permutation importance scores. Because the central aim for *EVAX* was computing everyday explanations with high fidelity, we have chosen to leave out this feature importance. Further research is needed to assess if the inclusion of feature interactivity in the argument strength scores is needed when there is a stronger correlation among features.

Argument strength	**Mean fidelity**
No feature interactivity	0.95
With feature interactivity	0.77

Table 5: Mean fidelity scores with and without feature interactivity.