Pontifícia Universidade Católica
do Rio de Janeiro

Gabriel Banaggia

Explainable Artificial Intelligence in Machine Learning

Capstone Project Report

Report presented to the Programa de Graduação em Ciência da Computação da PUC-Rio in partial fulfillment of the requirements for the degree of Bachelor in Computer Science.

Advisor: Prof. Simone Diniz Junqueira Barbosa

Rio de Janeiro
March 2024

Pontifícia Universidade Católica
do Rio de Janeiro

Gabriel Banaggia

Explainable Artificial Intelligence in Machine Learning

Report presented to the Programa de Graduação em Ciência da Computação da PUC-Rio in partial fulfillment of the requirements for the degree of Bachelor in Computer Science. Approved by the undersigned Examination Committee.

Prof. Simone Diniz Junqueira Barbosa
Advisor
Departamento de Informática – PUC-Rio

Prof. Clarisse Sieckenius de Souza
Departamento de Informática – PUC-Rio

Dr. Viviane Torres da Silva
IBM Research – IBM

Rio de Janeiro, March 27, 2024

Gabriel Banaggia

Gabriel Banaggia holds an MSc and a PhD in Anthropology, with previous research in Religion and in Science and Technology Studies. His current interests in turning to Computer Science include Human-Computer Interaction, Human-Centered Computing and computational support for scientific discovery.

Bibliographic data

Banaggia, Gabriel

Explainable Artificial Intelligence
in Machine Learning / Gabriel Banaggia; advisor: Simone Diniz Junqueira Barbosa. – Rio de Janeiro: PUC-Rio, Departamento de Informática, 2024.

v., 117 f: il. color. ; 30 cm

Relatório de Projeto Final - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui bibliografia

1. Informática – TCC. 2. Ciência da Computação – TCC. 3. Inteligência Artificial. 4. Aprendizado de Máquina. 5. Explicabilidade. 6. Interpretabilidade. 7. Interação Humano-Computador (IHC). 8. Computação Centrada em Humanos (CCH). 9. Inteligência Artificial Explicável (XAI). 10. Experiência do Usuário (UX). 11. Grandes Modelos de Linguagem. 12. Modelos de Fundação. I. Barbosa, Simone Diniz Junqueira. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

To Leo, always. This is just the beginning.

Acknowledgments

I would like to first thank my advisor, Prof. Simone D. J. Barbosa. Simone has shown me, time and again, first in class and then as an advisor, that it is perfectly possible to combine the quickest and sharpest of wits with the most inspiring and thoughtful guidance. I have always felt truly heard in her presence, and our weekly meetings were a constantly welcome thrill.

I would also like to thank my managers at IBM, Renato Cerqueira and Viviane Torres. Renato brought me back when I could have been lost, and I found in him a kindred spirit. Viviane has shown me constant support, and made me feel an integral part of the team. They both reminded me that sometimes the best of academic thought can also be found elsewhere. My gratitude towards them cannot be overstated.

I would equally like to thank Prof. Clarisse Sieckenius de Souza. Initially through her work, then once again after we met in person, Clarisse has exemplified time and again how brilliance knows no boundaries, disciplinary or otherwise. She is indeed a beacon leading those around her to greatness.

None of this would have been possible without the unfaltering love and support of my entire family, through so many changes and hardships. Helder, Lúcia and Priscilla, first and foremost, my biggest thanks once again.

My friends are too numerous and too important, but a few always deserve special mention. Ana, Fred, Ligia and Pedro, you know you constitute safe haven. André, Laura, Maria and Soldani, there is no way I would have done this without you guys.

My time at IBM was incredibly defining, and it was only possible because Renan and Leonardo believed in my potential. Thank you for this phenomenal ride, indeed this was an internship for the ages.

Having the support of my university was equally fundamental, among many other reasons by offering me a scholarship. Thank you Professors Adair, Augusto, Marcelo, Sonia and Valter, for all these years in the Social Sciences. In the Department of Informatics, my deepest admiration and gratitude go especially to Professors Bruno, Colcher, Hélio, Ivan, Joísa, Kalinowski, Molinaro, Noemi, Tatiana, and Waldemar, as well as Professors Christine and Luana, from the Mathematics Department. You all make learning a true joy.

Abstract

Banaggia, Gabriel; Barbosa, Simone Diniz Junqueira (Advisor). **Explainable Artificial Intelligence in Machine Learning**. Rio de Janeiro, 2024. 117p. Capstone project report – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Explainable Artificial Intelligence aims to assist in opening the black-box of opaque algorithms, especially present in machine learning systems, so they can earn due trustworthiness. This work employs a systematic literature review in order to understand how the concepts of explainability and interpretability, among others, are defined in this context. It offers a structured vocabulary to better comprehend and classify artificial intelligence systems and models, as well as the techniques used to make them more intelligible. Inspired by Human-Centered Computing notions, it becomes crucial to consider the audience to which explanations are owed, emphasizing that it is composed of a community of heterogeneous users.

Keywords

Artificial Intelligence Machine Learning Explainability Interpretability Human-Computer Interaction (HCI) Human-Centered Computing (HCC) Explainable Artificial Intelligence (XAI) User Experience (UX) Large Language Models Foundation Models

Resumo

Banaggia, Gabriel; Barbosa, Simone Diniz Junqueira. **Inteligência Artificial Explicável em Aprendizado de Máquina**. Rio de Janeiro, 2024. 117p. Relatório de Projeto Final de Graduação – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A Inteligência Artificial Explicável pretende auxiliar na abertura da caixa-preta dos algoritmos opacos, especialmente presente em sistemas de aprendizado de máquina, de modo que eles possam conquistar a confiança que for merecida. Este trabalho emprega uma revisão sistemática da literatura para entender como os conceitos de explicabilidade e interpretabilidade, entre outros, são definidos nesse contexto. Ele fornece um vocabulário estruturado para melhor compreender e classificar sistemas e modelos de inteligência artificial, bem como as técnicas utilizadas para torná-los mais inteligíveis. Inspirado por noções de Computação Centrada em Humanos, torna-se crucial considerar a audiência para a qual explicações são devidas, enfatizando que ela é composta por uma comunidade usuária heterogênea.

Palavras-chave

Inteligência Artificial Aprendizado de Máquina Explicabilidade Interpretabilidade Interação Humano-Computador (IHC) Computação Centrada em Humanos (CCH) Inteligência Artificial Explicável (XAI) Experiência do Usuário (UX) Grandes Modelos de Linguagem Modelos de Fundação

Table of Contents

List of Figures

List of Tables

D[aughter]: Daddy, what's a black box?
F[ather]: A "black box" is a conventional agreement between scientists to stop trying to explain things at a certain point. I guess it's usually a temporary agreement.
D: But that doesn't sound like a black box.
F: No—but that's what it's called. Things often don't sound like their names.
D: No.
F: It's a word that comes from the engineers. When they draw a diagram of a complicated machine, they use a sort of shorthand. Instead of drawing all the details, they put a box to stand for a whole bunch of parts and label the box with what that bunch of parts is supposed to do.
D: So a "black box" is a label for what a bunch of things are supposed to do. . . .
F: That's right. But it's not an explanation of how the bunch works.
D: And gravity?
F: Is a label for what gravity is supposed to do. It's not an explanation of how it does it.
D: Oh.
[...]
D: Do chromosomes learn?
F: I don't know.
D: They do sound rather like black boxes.
F: Yes, but if chromosomes or genes can learn, then they are much more complicated black boxes than anybody at present believes. Scientists are always assuming or hoping that things are simple, and then discovering that they are not.

Gregory Bateson,
Steps To An Ecology of Mind (1972).

1
Introduction

Inscrutable, opaque, black boxes, biased, inexplicable. These are just a few of the ways used to describe a set of computational inventions that figure among the most important in recent times, with ever increasing levels of performance and ever more disorienting degrees of complexity (Laato et al., 2022; Tomaino et al., 2022; Chazette, 2023; Dwivedi et al., 2023; Mohammadkhani, 2023; Nauta et al., 2023; Sado et al., 2023). Advancements reached in the last decade in artificial intelligence technologies, especially in machine learning and, more specifically, in deep neural networks, have made them nearly ubiquitous in our lives, being part of numerous automated or semi-automated decision systems in a wide range of application domains (Islam et al., 2022; Guidotti et al., 2018).

The accelerated growth in the adoption of these technological solutions has raised concerns regarding their **explainability**, something considered one of the main challenges - both for development and user communities - to address, which was not part of artificial intelligence ecosystems previously (Laato et al., 2022; Haque et al., 2023). Explainable artificial intelligence emerges thus as a new field within the realm of artificial intelligence studies, with the purpose of designing, assessing, and assisting in the construction of more transparent, interpretable, and less biased models (Guidotti et al., 2018; Alarcon and Willis, 2023; Dwivedi et al., 2023).

The demand for explanations becomes even more essential in the face of these systems' increasing popularity, and the previous emphasis on the accuracy of predictions made by models indifferent to the communication of the reasons behind a particular decision faces increasingly sharp criticism (Nauta et al., 2023). Even possibly due to the achievement of increasingly promising results, at least partly derived from inspiration in the structure of the human brain itself, which enables the learning of sophisticated features by artificial neural networks, the absence of symbolic interpretation capabilities and the explicit explanation of the model's functioning and its responses are becoming increasingly unsustainable (Islam et al., 2022; Ibrahim and Shafiq, 2023; Sado et al., 2023).

In order to facilitate the use and acceptance of the latest artificial

intelligence systems by humans, there has been a recognized need to make them increasingly comprehensible by enhancing transparency in decision-making processes supported by software, often achieved through the integration of explanations as part of the computer applications themselves to mitigate their opacity (Chazette and Schneider, 2020). Explainability becomes a means of achieving desirable characteristics in software that utilizes this type of artificial intelligence, helping artifacts become more transparent, accountable, and reliable, which can be considered essential aspects of a system, filling the gaps left by the absence of information that inhabit opaque models (Chazette et al., 2022; Alarcon and Willis, 2023; Dwivedi et al., 2023).

Part of this call is made by the field of Human-Computer Interaction, whose research indicates the importance of considering the perspectives of the end-user community, collaborating with software engineering teams so that neural networks can provide a range of different explanations regarding their operation and results (Laato et al., 2022; Ibrahim and Shafiq, 2023). Furthermore, this goal becomes even more essential in decision-making processes involving sensitive situations, where decisions will have a direct or indirect impact on people's lives, requiring these systems to not only be robust and accurate but also take into account socially relevant values such as equity, privacy, accountability, reliability, and acceptance, including the possibility of appropriately appealing decisions made based on them, providing reasonable recommendations for adequately altering a particular evaluation (Coeckelbergh, 2020; Karimi et al., 2022; Kaur et al., 2022; Li et al., 2023).

The seriousness required in carefully considering explainability deepens even further when the subject is part of the universe of critical applications, including, for example, the medical and legal fields, an expansion that becomes even more pronounced when one considers the possibility of using autonomous systems, including robotic agents with a high degree of automation working in conjunction with humans on tasks that often carry serious risks (Vouros, 2022; Zini and Awad, 2022). In these cases, any breakdown of trust in the explanations provided by these systems among the parties involved can have serious consequences, even leading to the loss of human lives (Li et al., 2023).

Thus, it can be seen that the absence of explainability constitutes a simultaneously practical and ethical issue, especially when vast amounts of data are involved in model training, which may contain human prejudices that can be inherited by systems that will start making incorrect and unfair decisions, possibly exacerbating problems that a hidden, internally sealed logic may conceal (Guidotti et al., 2018). A series of legislative initiatives around the world aimed at data protection and regulation have also turned their attention

to decision-making systems that use automation, such as profiling individuals, seeking to ensure the individual right to an explanation of the logic involved in processes using these technologies (Guidotti et al., 2018; Sado et al., 2023).

In addition to the already evidenced reasons that point to the need for explainability, the literature also indicates a series of benefits that its adoption can bring in any domain, such as increasing the trust of stakeholders in a particular application, improving the quality of training data and the neural network models themselves, increasing the accuracy and performance of specific tasks, enhancing the communicability of the application as a whole, and even reducing energy consumption since the processes of implementation and use of these types of systems become increasingly computationally costly (Chazette et al., 2022; Habiba et al., 2022; Zini and Awad, 2022; Alarcon and Willis, 2023; Haque et al., 2023; Ibrahim and Shafiq, 2023; Saeed and Omlin, 2023).

As such, first of all this work consists of an investigation on the different meanings that explainability and its related concepts can acquire according to a systematic literature review on the topic. We will deal with a wide range of solutions aimed at achieving greater explainability and elaborate a possible typology to classify the techniques at hand. Finally, we will offer ways in which our findings can equip communities of developers and of users with ways to deal with artificial intelligence systems in a more explainable manner.

This report is structured as follows. Chapter 2 introduces the background of Explainable Artificial Intelligence and organizes the knowledge that was taken into account for proposing the current study. Chapter 3 presents the methodology employed in the systematic literature review, which is further detailed in Appendix A. Chapter 4 gives a rundown of the texts that were discovered in the review and how they were organized for deeper study. Chapter 5 constitutes the bulk of the work, in which the selected texts are considered in depth and the themes found among them are compared, contrasted and discussed. Lastly, Chapter 6 recaps and concludes our findings, indicating avenues for future work.

2
Background and Related Works

To get a better feel of the subject before properly diving into the main idea for this study, initial and non-systematic literature reviews were conducted through searches on Google Scholar and the ACM Computing Surveys journal, using terms such as explainability, explicability, interpretability, AI, XAI, explainable AI, and others. As will be detailed later, the proposal for this work involved conducting a systematic review of the available literature on these topics, following a more well-defined protocol. However, even the procedures outlined for a review of this kind, described in Chapter 3, involve an initial exploratory research that informs the creation of the protocol itself, and this is what is presented in the current chapter, with an initial identification of current trends in the field (Biolchini et al., 2005).

Several possible definitions can be taken into consideration regarding the topic of explainability in the context of artificial intelligence, and since it is still a relatively nascent field, there is not yet a well-established consensus about them. It's worth noting that there is work that points out how the topic was already important in the field of artificial intelligence at least since the 1980s; however, it did not receive the same degree of attention it got in recent years, when efforts began to systematize previously scattered results, providing classifications of methods and taxonomies in a literature that is still clearly in a state of evident effervescence (Guidotti et al., 2018; Chazette et al., 2022; Alarcon and Willis, 2023; Dwivedi et al., 2023).

First and foremost, possibly the most important relationship to establish is between the concepts of explainability and interpretability. While some works equate the two, considering them even as simple synonyms (Kaur et al., 2022; Zini and Awad, 2022; Nauta et al., 2023), others prefer to create a clearer distinction between them, giving them different forms of intelligibility (Kaur et al., 2022; Karimi et al., 2022; Alarcon and Willis, 2023; Saeed and Omlin, 2023). Ultimately, it is evident that each scientific approach is developed to provide a solution to a specific problem, and consequently, the ways in which a context is delineated carry with them, either implicitly or explicitly, their own definitions of what explainability and interpretability would be, with considerable variation within different scientific communities (Guidotti et al.,

2018).

We can temporarily start from one of the broader definitions of explanation found in the literature to proceed with the description of the current state of the field: "An explanation is a presentation of (aspects of) the reasoning, functioning and/or behavior of a machine learning model in human-understandable terms" (Nauta et al., 2023) This sounds like it could be a good initial definition, as it allows for the inclusion of some of the different senses of explainability: after all, with it, one can both aim to explain details of the general functioning of a system, model, or architecture, as well as the reasons that lead to a specific decision (e.g., classification or prediction), providing this information to the development and/or user communities of applications that make use of these technologies (Kaur et al., 2022; Alarcon and Willis, 2023).

The exploratory analysis indicates that there is a consensus regarding the importance of the explainability principle, considered one of the major current scientific challenges (Guidotti et al., 2018). Since explainable artificial intelligence is still in its embryonic stage, widespread adoption of its techniques and premises in applications currently available to the general public has not yet been observed (Ibrahim and Shafiq, 2023). The very understanding of what explainability means, especially in the context of deep neural networks, is still not something settled (Vouros, 2022). Consequently, there also do not appear to be agreed-upon metrics for evaluating explainable artificial intelligence methods in the literature, leading to a lack of benchmarks for performance comparison, even though some have already been proposed and analyzed (Mohammadkhani, 2023; Nauta et al., 2023).

What is considered desirable is that computational agents using artificial intelligence can be capable of providing informative explanations about their decision-making, even in cases where humans might not be able to, although humans can provide paradigmatic demonstrations of skills that can teach a system how to solve certain tasks (Vouros, 2022). The types of explanations can vary widely, with a considerable range of explanatory techniques available, of varying interest depending on characteristics such as the application domain, the type of explanation itself, the type of data used by the system, the assumed background knowledge of the user community, and the specific question being sought to explain (Zini and Awad, 2022; Nauta et al., 2023). Initially, it is worth noting that the relative importance of a feature emerges as the most common type of explanation, and most explainable artificial intelligence methods focus on explaining individual predictions rather than providing global insights into a model's reasoning (Nauta et al., 2023).

Another aspect of the current situation relates to the widely discussed

need for more systematic studies on explainable artificial intelligence, with the observation of a scarcity of secondary studies (Islam et al., 2022). We would like to exercise caution regarding agreement with diagnoses of this kind, especially when considering that even an exploratory research has found more than one assessment claiming to be the only systematic literature review on a particular topic (e.g., Chazette et al. (2022); Mohammadkhani (2023)), despite the nuances investigated more thoroughly in each of them. Conversely, lists were found that enumerate a significant number of systematic reviews already conducted (cf. references in Zini and Awad (2022); Haque et al. (2023); Nauta et al. (2023)), and it seems reasonable to acknowledge that a systematic review is never truly complete; it reaches a point of saturation when no more new concepts or interpretations can be obtained from new material (Chazette et al., 2022).

There are many distinct ways to provide explainability in artificial intelligence, and they also vary not only in terms of 'how' but also effectively in terms of 'what' is being sought to explain, which are non-trivial questions. Regarding the 'how', it is possible, for example, to invest in visual explanations, rule extraction, providing semantic descriptions (Islam et al., 2022; Ibrahim and Shafiq, 2023; Saeed and Omlin, 2023). As for the 'what', research points to important distinctions between the explainability of the data used for training, the model itself to be trained, and also the results of the model after making specific predictions or decisions (Vouros, 2022; Zini and Awad, 2022; Dwivedi et al., 2023; Ibrahim and Shafiq, 2023).

Furthermore, it is also useful to consider in which phases of an application's life cycle explainable artificial intelligence techniques should be taken into account, which entails different moments and modes of implementation, investing in methods that are inherently explainable to the extent possible (Vouros, 2022; Dwivedi et al., 2023; Li et al., 2023). Similarly, approaches that consider offering explainable artificial intelligence often need to deal with the possibility of a trade-off between explainability and performance, where performance is understood as the degree of accuracy of a predictive model (Guidotti et al., 2018). To address this dilemma, the suggestion is to consider 'for whom' explanations should be provided since the different stakeholders involved in the creation and use of artificial intelligence systems may require very distinct forms of explanations, which need to be properly managed (Orphanou et al., 2022).

Many of the existing studies we have brought to bear in this background exploratory research were aimed at very different application domains, and it is crucial to distinguish between, on one hand, critical domains or those that

involve sensitive data, and on the other hand, the rest. While some of the texts thus found presented arguments that did not take into account different application domains, others were more directly related to one or more of these domains, including economics and finance (like obtaining credit lines), health, education, insurance, transportation, hiring, public safety, the judicial system, scientific research (in various disciplines), robotics, and defense. Meanwhile, applications in sales, advertising, entertainment, message filters, or online search services are examples among non-critical domains (Guidotti et al., 2018; Islam et al., 2022; Kaur et al., 2022; Laato et al., 2022; Dwivedi et al., 2023; Haque et al., 2023; Ibrahim and Shafiq, 2023; Li et al., 2023; Sado et al., 2023).

There are different approaches that have caught our attention in the research on explainable artificial intelligence during the exploratory search, and we highlight two of them. The first one is related to understanding explainability as an emerging non-functional requirement that would be part of the quality assessment of software (Chazette and Schneider, 2020; Chazette et al., 2022; Chazette, 2023), assessable through specific metrics (Ibrahim and Shafiq, 2023; Nauta et al., 2023). In this sense, one of the works takes a stance by asserting the existence of synergy between requirements engineering and explainable artificial intelligence (Habiba et al., 2022). A point that connects this approach with the following one is related to the understanding of requirements geared towards the end-user community, which leads us to the second envisioned perspective.

Focusing on explainability directed towards the user community, various works point out how explainability takes on a different meaning when this group is more directly in mind, considering their specific needs, indicating their relative scarcity compared to other existing ones, and suggesting the need for more primary research involving the perceptions of the humans who use these systems (Alarcon and Willis, 2023; Haque et al., 2023). This literature aims to identify when explanations are necessary or desirable, in what formats and levels of detail (with the possibility of further elaboration when appropriate), and generating the greatest impact on user community trust in the employed technologies and suggested solutions, while keeping the focus on the human agents who will be making critical decisions with the support of these systems, without losing sight of all the stakeholders involved in the process (Orphanou et al., 2022; Vouros, 2022; Alarcon and Willis, 2023; Ibrahim and Shafiq, 2023).

Here are some of the main conclusions highlighted in the initial bibliography we had access to. The available explainable artificial intelligence methods are often more applicable to so-called classical machine learning models (random forests, decision trees, and regression models) than to more complex ones

(such as deep neural networks and generative models, such as transformational architectures) (Mohammadkhani, 2023). The use of collaborative and interdisciplinary research, the value of combining explanatory methods to obtain more powerful explanations (including causal, contrastive, and counterfactual ones), the relevance of explaining different types of data differentially, and the importance of communicating uncertainty (both about the models and the explanations themselves) to the user community are just some of the contributions chosen among the 15 key conclusions synthesized in one of the considered studies (Saeed and Omlin, 2023).

In addition to these, it is worth mentioning among the findings of the texts in question the existence of at least four dimensions that shape explanations in artificial intelligence, namely, format, completeness, accuracy, and timeliness (Haque et al., 2023), while pointing out the lack of research on perceptions of how much trust and reliance there is on machine learning (Alarcon and Willis, 2023). One literature synthesis indicates that there are five main objectives in the communication of artificial intelligence systems with the user community, which are degrees of understanding, trust, transparency, control, and fairness, with corresponding design suggestions, that there is no one-size-fits-all solution that is best for all cases, and that it is necessary to consider the multiple possible trade-offs between them (Laato et al., 2022), potentially creating the possibility of optimizing a system for greater explainability (Nauta et al., 2023).

Finally, it is worth mentioning in this section an apprehension of the conduction of the systematic literature reviews found in the exploratory research. Most of them cite as a reference works authored by researcher Barbara Kitchenham (the most frequent of them being Kitchenham and Charters (2007)): Chazette et al. (2022); Islam et al. (2022); Laato et al. (2022); Chazette (2023); Haque et al. (2023); Mohammadkhani (2023); Saeed and Omlin (2023). However, a closer examination reveals that the proposed systematic review methods (which will be further unveiled in Chapter 3) were not always consistently followed: there were often excessive limitations in the chosen keywords for the searches (cf. Islam et al. (2022); Saeed and Omlin (2023)), improper use of logical operators for conducting searches (cf. Haque et al. (2023)), restrictions on the sources effectively consulted (cf. Haque et al. (2023)), outdated time constraints (cf. Chazette (2023)), a lack of bibliographic reference to the sources that would structure the systematic review (cf. Votto and Liu (2023)), or simply excessive deviations from the systematic research protocol (cf. the so-called “manual search” conducted in Chazette et al. (2022)).

3
Methodology and Execution

A considerable part of this work involved enacting a Systematic Literature Review on the topic of explainability in artificial intelligence, based on the results of the initial searches that were detailed in the previous chapters. This allowed us to determine if the chosen topic was too broad to be covered within the available time, helping to narrow the focus to arrive at the appropriate research questions (Carrera-Rivera et al., 2022). As one of the preliminary steps before conducting the actual review is to ensure its necessity by conducting a search for existing similar reviews (Kitchenham, 2004), if the existence of one or more high-quality reviews was confirmed, providing information on the selected topics (Biolchini et al., 2005), we envisioned two alternatives: either shifting the chosen topics (for example, delving into a specific application context or a specific technology), or conduct a tertiary study by considering the range of existing systematic reviews. Since that was the case, the first option was pursued, as will be detailed further down.

The method used for conducting systematic reviews in computer science is explicitly inspired by guidelines used for similar studies in medical research, with adaptations to be applicable to specific problems in software engineering research (Kitchenham, 2004). Significant differences arose in this transition, including the fact that techniques used in software engineering can potentially impact a part of its lifecycle, with non-trivial effects to isolate, as they interact with other development techniques and procedures (Biolchini et al., 2005). Establishing linear causal links in this way is challenging, especially when the use of the technique and the final outcome of an application are distant from each other and separated by a series of other activities or tasks that can affect the latter (Biolchini et al., 2005). Another fundamental difference compared to medical studies is that in software research protocols, it is not possible to establish double-blind experiments, since in the realm of computing they depend on the participation of humans who necessarily possess specialized knowledge of the procedures they apply (Biolchini et al., 2005). The extensive involvement of humans in this process indicates another interesting approach, with methods related to other fields of knowledge.

That is why a portion of the available bibliography on conducting sys-

tematic literature reviews in computer science indicates that their methods are not only similar to those used in healthcare sciences research but also to those in the social sciences, a field known for conducting observational studies where the definition and measurement of constructs of interest and understanding the impact of context on research outcomes are equally important (Kitchenham, 2004; Carrera-Rivera et al., 2022). One of the consulted texts even mentions that the origin of these research synthesis methods, even in the medical sciences, can be traced back to the field of social sciences, where the first books on these research methodologies were published in the 1980s (Biolchini et al., 2005).

The execution of a systematic review follows a specific scientific method that goes beyond a simple review, through empirical research that allows for generalizations (Biolchini et al., 2005). Among its objectives are the formulation of appropriate research questions, as well as the conduction of quantitative and qualitative analyses, grouping the found studies according to key themes or characteristics through a research protocol that structures the methods to be effectively used in its search, selection, organization, and presentation, in a process that aims to be transparent and potentially replicable (Kitchenham, 2004; Carrera-Rivera et al., 2022).

The texts that detail procedures for conducting systematic literature reviews emphasize that they do not constitute a guaranteed way to find all relevant literature in a specific area (Kofod-Petersen, 2015), a project that would be overly ambitious. For our current purposes, it is worth noting that much more commonly, systematic reviews are carried out by collaborative research teams rather than individual researchers, and more often as extensive academic works such as master's theses or doctoral dissertations (Kitchenham, 2004; Biolchini et al., 2005; Carrera-Rivera et al., 2022). One of the consulted sources even suggests a simplified version of the systematic review protocol for individual researchers (Kitchenham, 2004), which closely aligns with the procedures recommended by the version we have chosen as the main guide for our endeavor (Carrera-Rivera et al., 2022).

The process of constructing a protocol for conducting a systematic literature review is iterative, and even its initial results can help refine it, as was the case in this study. However, it is not advisable to change the data extraction form after starting to fill it out, as this could result in the need for significant rework (Carrera-Rivera et al., 2022). Thus, any potential changes to the protocol after its approval were limited to its initial fields, such as refining research questions or search strings after obtaining some initial results.

The systematic literature review was initially conducted according to a

protocol specified in Carrera-Rivera et al. (2022). A form containing the final established protocol is available in Appendix A, a version in which we arrived at after initial trial and error. While this chapter sums up the review process, the previously referenced section contains a lot more detail about the minutiae of the systematic search, so readers can effectively reproduce the steps taken and critique them as necessary.

After the preparatory study was enacted and the initial proposal drafted and adjusted with continuous help from this work's supervisor, five sets of criteria were chosen to delineate the searches proper. For these choices, the so-called PICOC Criteria were selected, a method mentioned in Carrera-Rivera et al. (2022), that was initially devised as PICO for evidence-based medical research (Richardson et al., 1995) and then expanded upon in the social sciences (Petticrew and Roberts, 2006).

The acronym stands for Population, Intervention, Comparison, Outcome, and Context. In our case, that initially meant roughly Foundation Models, Explainability, Black Boxes, Explanations and Definition, a set of keywords accompanied by synonyms that accompanied the possibly five research questions the systematic review could attempt to answer, being:

RQ1. *How is 'explainability' defined in the domain of foundation models?*

RQ2. *What are the current efforts to explain emergent capabilites in AI?*

RQ3. *To whom are AI explanations owed, and what does that figuration entail?*

RQ4. *How does trying to achieve explainability affect a model's performance?*

RQ5. *What does considering explainability a (non-functional) requirement involve?*

After some deliberation, the first two questions were deemed more feasible for the work given the available time, and the first one was finally selected considering the initial search results. Four digital library sources were considered, and we settled upon two of them to conduct the search queries in the end. A set of different inclusion and exclusion criteria were enacted to arrive from the final search results to the set of texts that were part of the full knowledge extraction process. The finalized search string and corresponding options for querying each repository, as well as the rest of the discussion on how we arrived at it, can be found in Appendix A, as has been mentioned. Notably, we tried our best not to veer too far off the initial intent of contemplating Large Language Models, and even Foundation Models as a whole, but as we will see in the next chapter, our focus ended up having to be Deep Learning more generally instead.

4
Results

The initial search results comprised 213 unique items, including journal articles, conference papers, books, and book chapters. Further details on how these were processed are supplied in Appendix A, which also includes a link to a publicly available spreadsheet recording the application of the inclusion and exclusion criteria mentioned therein.

For each of the entries we proceeded to a complete reading of the resource's title and abstract, and to a complementary reading of each entry's Introduction and/or Conclusion whenever it was necessary in order to gather the required information. Other than the metadata that was already programmatically available (Title, Type of Source, and Publication Year), that information included full text Availability, Relevance to our RQ1 (with the options being 'Fully Relevant', 'Partially Relevant', or 'Not Relevant'), and the Main Application to which the item made reference, whenever the entry fulfilled the relevance criterion at least partially.

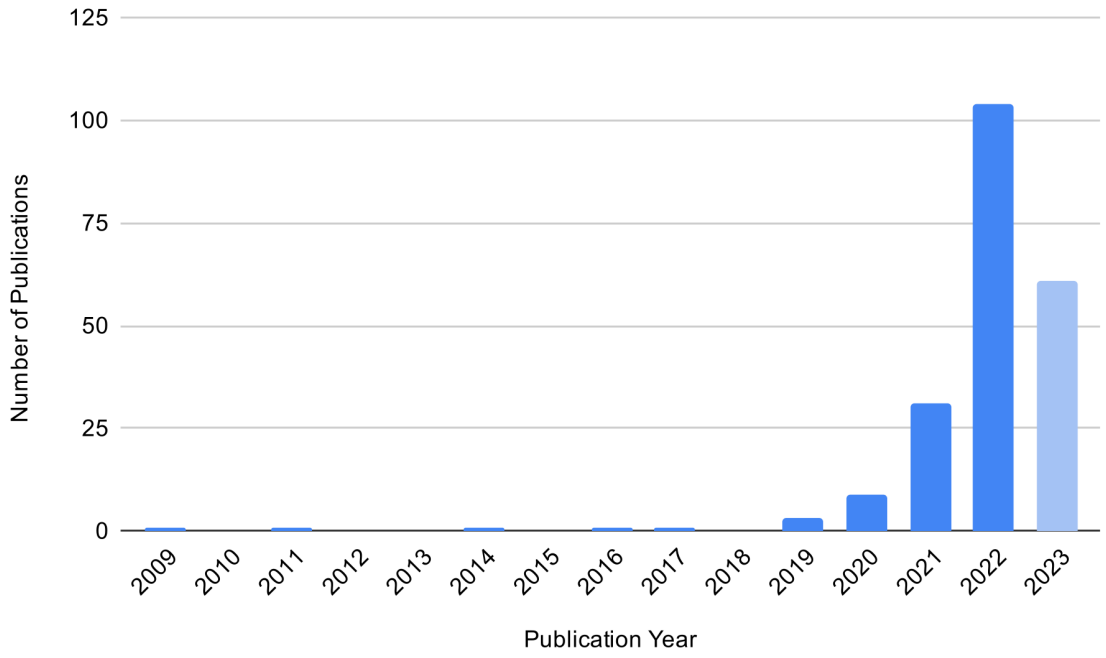

Figure 4.1: Distribution of publications over the years

Figure 4.1 shows an exponentially rising frequency for the texts, with a

clear spike in 2022, the year in which large language models dominated the interest of the public eye in regard to Artificial Intelligence. Since the final queries were run on July 7, 2023, it is very probable the tendency of interest in the theme either confirms itself or tapers out.

The different applications for explainable artificial intelligence techniques found in the initial query are summed up in Figure 4.2. Whenever more specific applications were mentioned, they were preferred to more general characterizations. For instance: if the task was mentioned as 'Drug discovery', that was the label chosen, even if the application also involved 'Property prediction' more broadly. If an item could be classified as having two or more different types of specific applications, the most prevalent one was chosen. When no single application stood out, the item was classified as dealing with 'Various' applications. It was no coincidence that all of the texts that were deemed 'Fully Relevant' for our research question were labeled as dealing with various kinds of applications, since these texts were the ones more interested with more general definitions regarding explainable artificial intelligence as well.

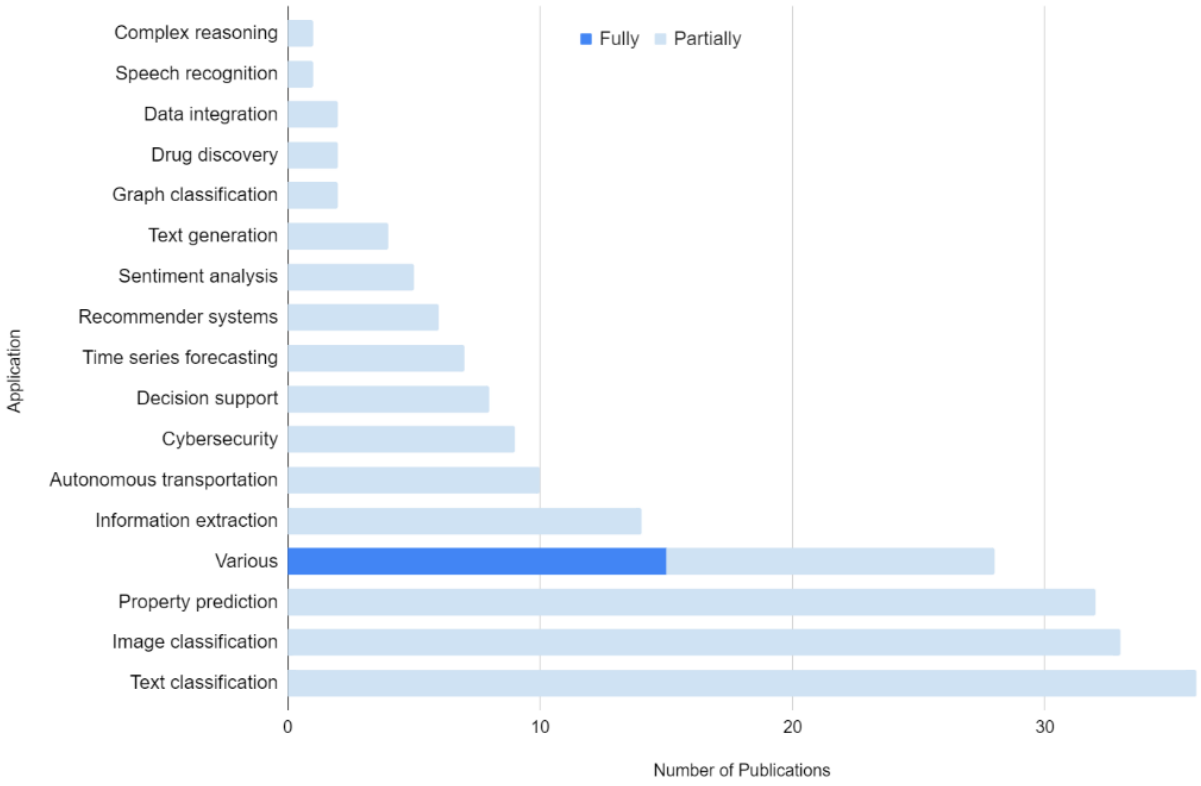

Figure 4.2: Applications frequency

We concluded the steps chosen for inclusion and exclusion of texts by selecting all of the items that were considered 'Fully Relevant' to our research

question with the intent of reading them in full for the information extraction stage. During this process, we realized one of the items did not actually deal directly with the research question, since it did not address a definition of explainable artificial intelligence in the context of deep machine learning. At the same time, from each of the book entries we usually chose one chapter that dealt more directly with the theme of our work. However, there were two instances when two chapters were considered relevant in the same book. That brought us to a final total of 16 texts for the information extraction process, listed in Table 4.1.

Table 4.1: SLR results

First author, Year	**Title**
Barredo Arrieta, 2020	Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI
Chaudhari, 2021	An Attentive Survey of Attention Models
Bruce, 2021	Background: Modeling and the Black-Box Algorithm
Bruce, 2021	Model Interpretability: The What and the Why
Loh, 2022	Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)
Bhattacharya, 2022a	Foundational Concepts of Explainability Techniques
Bhattacharya, 2022b	Model Explainability Methods
Meador, 2022	Explainable AI - Using LIME and SHAP
Correia, 2022	Collaboration in relation to Human-AI Systems: Status, Trends, and Impact
Li, 2022	Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond
Zini, 2022	On the Explainability of Natural Language Processing Deep Models
Liu, 2023	Fundamentals of Deep Learning Explainability
Simon, 2023	Understanding Explainable AI

Continued on next page

Table 4.1: SLR results (Continued)

Nannini, 2023	Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK
Räuker, 2023	Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks
Liu, 2023	Trustworthy AI: A Computational Perspective

The information extraction involved reading the texts in full, highlighting and annotating the most important parts in each one, deriving a set of categories considered significant for our research question, and iterating on them as more items were analyzed, noticing recurrences and gaps along the way. After reading the texts, the passages and annotations from each one were reread and classified in one or more categories. The results of this step are also supplied in full in a publicly available spreadsheet, as detailed in Appendix A, and were organized as thus:

1. Organizational information:
    - Title
    - Year
    - Author(s)
    - Item Type
    - Keywords
2. Summarizing information:
    - Initial context and motivation
    - Application domains and specific applications mentioned
    - Main contribution summary
3. Definitions for key terms:
    - Explainability characterizations
    - Distinction between explainability and interpretability
    - Explainability definition
    - Interpretability definition
    - Relevant additional terms (understandability, comprehensibility, transparency, etc.)
4. Effects and ways of achieving explainability:

- Benefits of explainability/interpretability
- Risks of non-explainability/non-interpretability
- Techniques for explainability/interpretability

5. Key takeaways for additional exploration:

   - Important references (for backward snowballing)
   - Interdisciplinarity
   - Assessment and metrics
   - Future challenges and possible drawbacks

The detailed analysis of these passages constitutes the remainder of this work.

5
Discussion

This chapter discusses all of the themes that we organized during the information extraction procedure, according to five different groups of subjects, as detailed below: organizational information; summarizing information; definitions of key terms; effects and ways of achieving explainability; key takeaways for additional exploration.

5.1
Organizational information

We used the following simple formula in Google Sheets to count the occurrences of keywords in the selected texts:

```
=ArrayFormula(
    query(
        flatten(split(E2:E17,",")),
        "select Col1,count(Col1)
        where Col1 <> "
        group by Col1
        label Col1 'Unique words',
        count(Col1) 'Frequency'" ,
        0)
    )
)
```

There was very little overlap, with only a few keywords appearing more than once. The full list and how many times each keyword occurred are in Table 5.1.

Most of the keywords that appear more than once have to do directly with the main themes of our work, which is to be expected, though there were actually very few terms that appeared more than once in the survey. The great degree of variability in the keywords found indicates considerable thematic dispersion in the domain under review, which will be further explored in the subsequent sections.

Keywords	Count
AI Policy, Altmetrics, Attention, Attention Models, Attention Mechanism, Autonomy, CBR, Comprehensibility, Computer Vision, Data Fusion, EBM, Emerging Topics Detection, Environmental Wellbeing, Expert System, Explaining Decisions, Funding, Generative Models, Governmental Regulations, GradCAM, Healthcare, Human-AI Systems, Interpretable Deep Learning, Interpretation, Law, LIME, LRP, Language Models, Neural Networks, PRISMA, Responsible Artificial Intelligence, Robustness, Rule-based, SHAP, Saliency Map, Scientometrics, Security and Privacy, Social Epistemology, Surveys and Overviews, Transformers, Transparent Embedding Models, Trust, Trustworthiness	1
Accountability, Deep Learning, Fairness, Interpretability, Neural Machine Translation, Privacy	2
Artificial Intelligence, Explainability, Interpretability, Machine Learning, Natural Language Processing/NLP, Transparency	3
Explainable Artificial Intelligence/Explainable AI/ExAI	4

Table 5.1: Keyword Frequency

5.2 Summarizing information

The next subsections discuss the texts concerning initial context and motivation; application domains and specific applications mentioned; and summaries of main contributions.

5.2.1 Initial context and motivation

The texts under review point that Artificial Intelligence has achieved notable momentum in recent years, affirming the existence of a consensus about the paramount importance of machines endowed with capabilities for learning, reasoning and adapting, describing them as pivotal for the future development of the human society (Barredo Arrieta et al., 2020). In comparison with rule-based computer programs, which are said to be limited by the knowledge and intelligence of the programmer, machine learning algorithms have the ability to provide rich insights and accurate predictions even in inordinately complex situations, in some cases surpassing human performance (Bhattacharya, 2022a; Zini and Awad, 2022).

Deep neural networks (artificial neural networks with multiple layers between the input and output layers, systems also referred to as ‘deep learning’)

are being increasingly deployed in real scenarios, used in all sorts of problems, focusing mostly on classification, regression (numeric predictions), and clustering. Reinforcement learning and recommendation systems also deserve special mention due to their rising adoption across different industries, and all of these technologies and models are said to be revolutionizing applications all around (Räuker et al., 2023; Liu and Zaharia, 2022; Zini and Awad, 2022).

At least part of the context the review shows frames the current moment as one ripe for an area that discusses collaboration between human and artificial intelligence systems, even if riddled with growing concerns for the adoption and democratization of the latter (Correia and Lindley, 2022; Bhattacharya, 2022b). Considering their role in the current times, the texts under review deem necessary making these systems trustworthy, in order for humans to rely on them with minimal concern for the potential harm they could do (Liu et al., 2023).

While mentioning the rapid development and sophistication of artificial intelligence systems, several of the texts mention that there are now even low and no-code options available to user with little technical expertise, to the point that almost no human intervention is needed for their design and deployment (Barredo Arrieta et al., 2020; Meador and Goldsmith, 2022; Simon and Barr, 2023).

This ease of use and deployment, and sometimes configuration, could not pose a higher contrast to the way these systems are said to be internally developed and constituted: in this context, the terms that appear in the review include some like black-box, opaque, non-linear, too complicated, overly complex, hard or difficult to understand or even approximate, particularly so for deep neural network architectures (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021a; Li et al., 2022; Chaudhari et al., 2021; Zini and Awad, 2022). A few of the texts under consideration even opted for terms that orbit the semantic universe of occultism to describe these systems, calling them oracles, magical, mystifying, capable of providing silver bullets to automatically decipher life, which makes their unregulated use and the incontestable truths so produced at least somewhat worrying: especially when they are wrong (Bruce and Fleming, 2021b; Bhattacharya, 2022a)

With the so called deep learning revolution that took place accelerated by the breakthrough of the ImageNet dataset in 2012, the Defense Advanced Research Project Agency (DARPA) of the United States sounded a call for Explainable Artificial Intelligence research in 2015, formally launching a program in 2017 intending to develop intelligent systems that end users could understand and trust (Simon and Barr, 2023). The following years

saw explainability as one of the main barriers to practical implementation of these systems, which faced public ire due to misapplication, demands for increased regulation, legal exposure, and also the simple desire for greater transparency for the algorithms which increasingly affect people's everyday lives (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021b).

The context to which the review initially points is one in which this burgeoning area of explainability in artificial intelligence has recently gained a lot of traction, not just in the public eye but also in the industry and academia, where there have been comparatively limited works so far (Correia and Lindley, 2022; Loh et al., 2022). The relative paucity of systematic studies on the theme in this still emerging area also contributes at least in part to the difficulty in finding agreement about not only methods but even for the terminology involved in the phenomenon, with growing interest in disambiguating the lexicon and proposing precise definitions for the concepts under use, which should also help with successful regulatory efforts and actionable recommendations (Liu and Zaharia, 2022; Simon and Barr, 2023; Nannini et al., 2023).

Lastly, it should be mentioned, for a characterization of the initial context pointed to by the review, that explainability has also become a prominent research focus in human-computer interaction as well as machine learning communities, fueled by the growth of focus on user interests, social trust, regulatory compliance, model debugging, and business interests, all themes to which we will return in detail later on in the current work (Nannini et al., 2023; Simon and Barr, 2023).

5.2.2 Application domains and specific applications mentioned

Like we mentioned in Chapter 4, we prioritized items in which the focus tended to be more on definitions for explainability than on specific applications. Consequently, it makes sense that, in contrast with the texts that were excluded from a more detailed appraisal, the ones that were read in full deal comparatively less with particular tasks or applications. Nonetheless, several domains of interest are indeed mentioned in the texts under review. They are summarized in Table 5.2.

There were also quite a few more particular tasks mentioned without reference to specific applications and corresponding domains: natural language processing, computer vision, multi-modal tasks, recommender systems, graph systems, dialogue systems, text summarization, machine translation, question answering, sentiment analysis, and information retrieval (Chaudhari et al.,

2021; Li et al., 2022; Liu et al., 2023).

Domains (and example applications)	**Source(s)**
Commerce (retail, supply chain, purchase recommendation)	Bruce and Fleming (2021b); Bhattacharya (2022a); Meador and Goldsmith (2022); Liu et al. (2023)
Education (co-creativity)	Bhattacharya (2022a); Correia and Lindley (2022)
Finance (*insurance*, *credit lending*, *banking*, *auditing*, *risk assessment*)	Barredo Arrieta et al. (2020); Bruce and Fleming (2021a); Bhattacharya (2022a); Meador and Goldsmith (2022); Liu and Zaharia (2022); Liu et al. (2023)
Healthcare (*precision medicine*, *predictive medicine*, *clinical decision support*, *drug discovery*, *disease diagnosis*, *gene analysis*)	Barredo Arrieta et al. (2020); Loh et al. (2022); Bhattacharya (2022a); Correia and Lindley (2022); Li et al. (2022); Liu and Zaharia (2022); Liu et al. (2023)
Hiring (*job application screening*, *human resource management*, *job recommendation*)	Bhattacharya (2022a); Correia and Lindley (2022); Liu et al. (2023)
Leisure (games)	Correia and Lindley (2022)
Military (*patented technologies*)	Meador and Goldsmith (2022); Correia and Lindley (2022)
Security (*prison sentencing*, *recidivism prediction*, *authentication*)	Barredo Arrieta et al. (2020); Meador and Goldsmith (2022); Liu and Zaharia (2022); Liu et al. (2023)
Transportation (*autonomous vehicles*)	Barredo Arrieta et al. (2020); Bhattacharya (2022a); Liu and Zaharia (2022)
Weather (forecasting)	Bhattacharya (2022a)

Table 5.2: Application Domains Mentioned (critical applications *in italics*)

Table 5.2 also makes a distinction between two types of applications, following the literature under review: that of low-stake and high-stake problems, with the latter also called critical or vital tasks (Bhattacharya, 2022a; Liu et al., 2023; Meador and Goldsmith, 2022). Explainability is considered even more crucial for these critical applications, which often include scenarios where experts require more information from the models than simple predictions, where companies or teams operate in more tightly regulated or risky industries, and where affected people's interests warrant even more responsibility in uncovering the reasons for algorithms' outputs (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021b; Meador and Goldsmith, 2022; Liu and Zaharia, 2022; Liu et al., 2023).

5.2.3 Summaries of main contributions

Here we offer brief summaries of the main contributions of each of the texts under review, in light of our research questions. We do this not only to be able to easily reference the place each text occupies in the economy of our argumentation, but also to delineate some of the ways in which we framed them for our purposes. They are described here sorted in chronological date of publication, which also corresponds to the order in which the texts were first read.

Barredo Arrieta et al. (2020) aim to establish a novel definition of explainability in machine learning, having the audience for which explanations are sought as one of their key foci. The authors employ the expression "Responsible Artificial Intelligence" to refer to a series of AI principles that need to be met in real application deployment, with a prominent role dedicated to the idea of 'understandability', to which we will return. Another key aspect of their contribution has to deal with fusing both black-box and transparent methods (such as using knowledge bases with semantic representations) to improve the explainability of exclusive data-driven approaches, combining connectionist and symbolic paradigms.

The same authors indicate that explanations are better when constrictive, by which they mean not only indicating why a model made a specific decision but also why it did not make a different one. They also suggest that referring to probabilities is not as important as offering causal links for supplying a satisfying explanation, translating probabilistic results into qualitative notions. The text mentions the importance of selective explanations, meaning that focusing solely on the main causes of a decision-making process is often sufficient. Finally, the authors put forward the idea that explanation generation

should not be left to end users, since different rationales might be employed depending on existing background knowledge.

Chaudhari et al. (2021) offer a survey with a structured overview on modeling attention based systems, proposing a taxonomy for such techniques. Their main hypothesis is that the magnitudes of the attention weights in a neural network correlate with how relevant a specific region of input is for the prediction of output, something that can be investigated by visualizing the attention values for sets of input and output pairs. One of the conclusions of their survey is that exploring the relationship between attention weights and model interpretability is still an early and active area of research, with future work being capable of looking at attention distributions of models and how they can be modified to offer plausible justifications of predictions.

Bruce and Fleming (2021b) and Bruce and Fleming (2021a) compare intrinsically interpretable methods, which yield output where the analyst can readily determine the role that individual predictors play in predictions, with black-box methods, like random forests or neural nets, which are too complex to allow such determinations. The author points to the existence of a trade-off between the two ways a model can err: it can be too flexible and complex, capturing so much information that it ends overfitting the data and modeling noise as well as real information, or too simple, underfitting the data and failing to capture some useful information. At the same time, the work goes on to say that simple coefficients and decision rules allow for ready interpretation and understanding, while black-box models are not easily interpretable due to how precisely they fit the subtle nonlinear relationships within the data, usually achieving better performance in the process.

The author concludes that ideally a good interpretability method is one that strikes a balance between being robust to unimportant changes in the underlying model while being sensitive to meaningful changes within it. He also mentions that to minimize a project's risk one should stick to using intrinsically interpretable models, which also lowers the time spent learning new interpretability methods and layering them over the original model. On the other hand, to maximize performance it is necessary to spend time developing expertise and toolchains that allow for using black-box models and interpretability methods together.

The work by Loh et al. (2022) directs itself to determining which explainability technique is more appropriate or widely proposed for different types of datasets in the specific domain of healthcare applications. They executed a systematic review arriving at 99 scientific articles, after removing duplicates and journals that were not ranked in the first quartile from an initial

1,194 search results. The authors indicate that the most popular explainability methods were SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (GradCAM). They also indicate that such techniques can be used not only to generate explanations but also to guide the hyperparameter tuning of artificial intelligence models.

Bhattacharya (2022a) and Bhattacharya (2022b) suggest four different dimensions for explaining black-box algorithms, namely understanding the data (by using a robust data curation process, including the analysis of data purity and the impact of adversarial effects), the model (by figuring out how the input data is mapped to the output predictions and being aware of the limitations assumptions and biases of the algorithms), the outcomes (by following why a certain prediction or decision is made by a model) and the end users (since it is vital to choose the right level of abstraction and details to achieve reliable and trustworthy results for the final consumers).

The same author indicates that knowledge extraction methods can be used to deal with essential information about input and output data from which the expected model outcomes are defined, while also noting that result visualization methods can be used to compare model outcomes with previously predicted values, particularly with surrogate models. Two more options mentioned are influence-based methods, which are techniques that help in understanding the role played by certain data features in influencing the model outcome (examples being feature importance, sensitivity analysis, key influencer maps, saliency maps, class activation maps) and example-based methods, which, unlike the other three types, do not require technical knowledge (the main examples are counterfactuals, which try to look at certain single instances of the data to explain a model's decision-making process and suggest which changes would result in a different prediction).

Meador and Goldsmith (2022) focus on the responsibility that is created when automated predictions affect real people's lives. The authors are keen to point that, for problems that deal with smaller datasets, or that are unrelated to natural language processing or computer vision, using more interpretable models might give better or very close results, while being able to be interpreted with much less effort. As such, they remark that simpler algorithms do not mean inferior results, especially when more complex tools like neural networks are used without careful consideration, even in cases where they provide very little benefit. They also point to the fact that there is always great uncertainty around how the logic of day-to-day aspects is applied, noting that even if artificial intelligence methods should not get a free pass, that also does not

mean that there are not many things in the world that we deal with all the time that are black boxes, with no clear way to get better interpretations of them.

In their paper, Correia and Lindley (2022) present the results of a bibliometric evaluation of scientific publications on collaboration between human and intelligent systems. Their findings highlight a significant focus on aspects like trust, explainability, transparency, and autonomy in highly complex scenarios through the use of generative models and hybrid interaction techniques, also pointing to a growth in the number of publications and funding grants, accompanied however by a lack of maturity observable in terms of citation patterns and coherence of thematic clusters.

The focus of Li et al. (2022) is the categorization of algorithms dedicated to interpretation in accordance with three orthogonal dimensions: representations of interpretations, targeting model's types for interpretations, and the relation between interpretation algorithms and models. The first dimension is the most relevant for our case, and it is quite varied, including: feature importance (estimating the contribution of each one with respect to the result given), model response (finding or generating new examples to see how the model behaves when responding to them), model rationale (replacing opaque models with interpretable ones to gain insights on the former's internal processes), and datasets (with algorithms aiming to explain how some data samples in the training set affect the optimization of the models).

As for Zini and Awad (2022), their contribution deals with the specificity of natural language processing tasks, since textual datasets present different challenges like polysemy, sarcasm, slang, cultural effect, and ambiguity, which end up being proliferated when using explainability methods. Since linguistic features are not as straightforward when compared to others like pixels or numerical attributes, and the majority of language models operate with embeddings, which are opaque representations, providing explanations in terms of specific embedding dimensions is not as practical, and requires further processing to dissect the learned knowledge in terms of not only their syntactic but also semantic (including contextual) qualities.

As the authors indicate, it is thus necessary to sparsify the dense term embedding models to be able to map these different dimensions of structure and meaning. That is usually done with the help of attention mechanisms, which the text points out has been accompanied by controversy ever since their inception: while some attention weights are able to provide reliable explanations, they are far from being easily interpretable. They conclude that most available work ends up focusing on understanding the inner workings

of the underlying models rather than understanding particular outputs of classification, which adds to that type of challenge.

The text by Liu and Zaharia (2022) is another one that drives the point that a precise definition of explanation depends on who wants the explanations, for what purpose, and at what time across an artificial intelligence system's life cycle. As they mention, the complexity of the explanations supplied needs to be tailored and selective to the receiving audience, without overwhelming information, including technical jargon. The piece draws on the human-computer interaction and user experience fields to drive forth the argument that there is no one-size-fits-all approach to explanations, so each method needs to be tailored according to different audiences like data scientists, machine learning engineers, business stakeholders, or other types of end users.

Simon and Barr (2023) is another text that reinforces the point that in explainability there are different stakeholder personas with diverse viewpoints, such as auditors, research scientists, industry experts, data scientists and other end users, for whom the definition of what counts as a meaningful explanations, and under which scopes, varies. As such, it is essential to gather direct user feedback and perform user testing to validate acceptance criteria for the explanations given, while not forgetting that users' perspectives on meaningful explanations can also change over time.

The review by Nannini et al. (2023) also centers on the importance of the context of use, but stemming from the point of view of legislation analysis, taking in existing governmental policies in the European Union, the United States of America, and the United Kingdom impacting the fruition of artificial intelligence explanations. Their main argument is that there exists a considerable discrepancy in how technological developers and policymakers tend to differently define these systems and their capabilities, and they suggest that establishing a proactive regulatory approach and regulate explainability of artificial intelligence systems is crucial to ensure their ethical alignment with human values and principles.

The same authors go on to identify misalignment with policies based on the literature from distinct research communities (algorithmic, human-computer interaction, ethics). They indicate that balancing model complexity, end-user expertise, as well as legal and commercial constraints is a considerable challenge, made even more difficult due to pervasive ambiguities, partly stemming from terminological imprecisions in the production of communications and reports across governments and commissioned standardization bodies. As they argue, current documents often lack an informed perspective of explainability as a research object still nascent, novel, and complex, far from being a

solved problem that would be 'easily implementable'. Another important cautionary remark involves the idea that policy trajectories could simply prioritize innovation under a risk management lens.

Räuker et al. (2023) offer a detailed survey on how to explain the inner structures of deep neural networks, arriving at a taxonomy that divides interpretability techniques by what part of a network's computational graph they explain: weights, neurons, subnetworks, or latent representations. The authors drive home the point that these interpretability techniques generate hypotheses and not conclusions, and that producing merely-plausible explanations is insufficient, since evaluating validity and uncertainty are both also key. Their survey brings examples of cases where sometimes very plausible-seeming explanations do not pass simple sanity checks or are very easy to find counterexamples for, claiming for more care since many works in interpretability fail to go beyond simply inspecting results.

As such, their contribution states that a focus on improving interpretability techniques without commensurate increases in capabilities offers the best chance of preventing advancements in artificial intelligence from outpacing our ability for effective oversight. Based on that, the authors argue that improvements in safety rather than capabilities should be the principal goal for future work in interpretability, adding that adequate scaling requires efficient human oversight, since many explanations obtained by state of the art interpretability techniques have involved a degree of human experimentation and creativity in the loop. Ideally, they conclude, humans should be used for screening interpretations instead of generating them, with solutions including using active learning, weak supervision, implicit supervision with proxy models or rigorous statistical analysis of proxies, with a combination of techniques (and the study of their interplay) possibly leading to better results.

Finally, Liu et al. (2023) also offer a survey, this time focusing on what they call six of the most concerning dimensions that have been extensively studied for trustworthy computational artificial intelligence, being: safety & robustness, non-discrimination & fairness, explainability, privacy, auditability & accountability, and environmental well-being. For each one of these dimensions, the authors present an overview of related concepts, finishing with a taxonomy to help understand how each dimension is studied, while also adding a summary of representative technologies in each case.

5.3 Definitions for key terms

The next subsections explore characterizations, definitions, and distinctions between the terms explainability an interpretability.

5.3.1 Explainability characterizations

Before we dive into the definitions for explainability and other related terms found in our systematic review, we would like to present a few characterizations around the term to help contextualize it. First of all, some of the texts mention that explainability is still a topic under recent and considerable research, decidedly complex and around which there seems to be no reasonable consensus yet (Liu and Zaharia, 2022; Nannini et al., 2023; Barredo Arrieta et al., 2020). As such, attaining explainability involved dealing with a significant barrier, a problem that is inherent of the latest artificial intelligence techniques brought by sub-symbolism (such as ensembles or deep neural networks), that were actually not present in the previous wave of applications (expert systems and rule-based models) (Barredo Arrieta et al., 2020).

Several of the reviewed works also assert that explainability is something plural, that can be provided in multiple ways depending on variables such as the type problem tackled or the type of data used (and the relative levels of complexity in both cases, which can include feature granularity, provenance tracking, intrinsic characteristics and changes over time) (Bhattacharya, 2022b). In fact, Liu and Zaharia (2022) define eight distinct dimensions from which the complexity of explainability can be understood, namely audience, stage, scope, input format, output format, problem, objective, and method.

Explainability is also characterized as something that is contextual, since it can be expected that the explanations put forth are adequately tailored to specific purposes, target audiences and risk levels, according to distinct sets of stakeholders, and taking into account a plethora of human factors, such as tech literacy and cultural backgrounds (Nannini et al., 2023). There is also mention of the possibility of self-explaining artificial intelligence systems, which entail a strategy of supervising the creation of human understandable explanations for model outputs computed from an model's inner representations, which has been attained for different tasks in computer vision and, natural language processing, even though the extent to which these explanations accurately explain the purported outcomes is still unclear (Räuker et al., 2023).

Another important distinction in how to characterize explainability has to do with it being oriented to more technical or non-technical audiences. Under

technical explainability we can list decisions for systems to be understandable and traceable, including comprehending the purpose of each artifact in relation to the design choices followed. Under lay explainability we can consider emphasizing the kind of expertise that is assumed of the stakeholders involved, if any, while not giving up performance metrics like prediction accuracy (Nannini et al., 2023; Liu et al., 2023). One set of authors (Räuker et al., 2023) offer a compelling analogy to help with the distinction between these two explainability orientations, based on their work about understanding neural networks: they indicate that most methods in the literature used for explaining these kinds of models aim to help a human "open up" the network and study parts of it, which would be similar to, if one wants to understand another human's reasoning, studying their brain directly. In most cases, they conclude, simply asking another human for an explanation of what they are thinking is much more effective; as such, endowing systems with the same self-explaining capabilities would be a promising direction, according to the same source.

The last characterization we would like to address involves considering explainability as a requirement, which is sometimes taken for granted in the texts under review. A few of them, however, do note that achieving explainability is a means and not an end unto itself, however vital, and that it could be taken as a first-classs artifact along with other pillars in the full life cycle of artificial intelligence systems development, along with data, code and models, which would also end up helping to construct more meaningful explanations target at more specific audiences (Barredo Arrieta et al., 2020; Bhattacharya, 2022a; Liu and Zaharia, 2022).

5.3.2 Distinction between explainability and interpretability

One last distinction is important to keep in mind before talking about the specific definitions given for key terms found during the review, and it has to do with a difference, which not all texts under review espouse, between explainability and interpretability, which were the two most common ways to describe very similar concepts. While exactly half of the texts under review make no mention of the possibility of distinguishing between both concepts, the other half do, and address the distinction differently as well.

On the one hand, Bhattacharya (2022b); Zini and Awad (2022) and Li et al. (2022) opt for using the terms explainability and interpretability interchangeably, the last of these mentioning that existing differences are subtle enough that justify that choice, in the process including the term 'attribution' in the same mix as well.

On the other hand are the texts that indicate the existence and the decision to maintain a distinction between the two terms, even when they mention being aware that these differences can be subtle (Liu et al., 2023), or the distinction loose (Nannini et al., 2023). Even when opting to work with the distinction, a couple of texts also note that it is common to see them used interchangeably in the literature due to lack of consensus even in standards communities, with ambiguities encompassing subjects like the type of explanation, the reason for the explanation, the purpose of the explanation, and to whom the explanation is provided (Liu et al., 2023; Nannini et al., 2023). One set of authors diagnose that the absence of a common point of understanding is effectively hindered by the interchangeable misuse of interpretability and explainability in the literature, arguing that there exist notable differences among the concepts (Barredo Arrieta et al., 2020).

There is, however, remarkable consistency in how the texts that argue for the distinction actually differentiate between the concepts (even though which term is chosen for which concept varies slightly), which is something we will keep in mind when presenting the definitions in the following subsections. For now, it will suffice to present how they separate the terms. In the texts that retain the distinction, we can summarize that **explainability** usually refers to understanding the mechanisms and inner workings of artificial intelligence systems and models, focusing on matters like how they operate and why they arrive at specific conclusions, telling us how they behave, and are usually the province of the designers and developers in charge of creating and maintaining them. **Interpretability**, conversely, usually refers to understanding why these systems' and these models' outputs and decisions matter, what do they entail, what kind of response they are supplying, and what are the possible repercussions involved in them, considering real-world practice, often being of more direct interest to policymakers and other users generally interested in artificial intelligence oversight (Barredo Arrieta et al., 2020; Bhattacharya, 2022a; Meador and Goldsmith, 2022; Simon and Barr, 2023; Nannini et al., 2023).

It was slightly surprising to notice that 4 out of the 16 main texts under review do not offer a definition of either explainability nor interpretability. When considering a case for each term, however, more than half of the total offer a specific definition of at least one of them on both accounts. Specific results are listed in Table 5.3.

Text contains a definition for...	Amount
Neither explainability nor interpretability	4
Explainability and other concepts	9
Interpretability and other concepts	10
Only explainability	2
Only interpretability	3
Both explainability and interpretability	7

Table 5.3: Amount of texts that offer definitions for key terms

5.3.3 Explainability definition

Unlike the more contrastive definitions we have just seen, the texts that offer more standalone definitions of explainability employ the term to refer to occasionally distinct concepts, sometimes having to do with the inner workings of systems and models, other times with communicating results and their possible consequences to non-technical end users. Whatever the case, several of the texts under review define explainability as an active, specific capability of artificial intelligence systems, as we shall now detail.

Barredo Arrieta et al. (2020) define explainability as "an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions". Nannini et al. (2023) define the term both as "[t]he ability of artificial intelligence (AI) systems to explain their decision-making processes" and as "the communicative ability to deliver insightful information regarding the inner functioning of complex algorithmic architectures". In turn, Liu and Zaharia (2022) define it as "providing selective human-understandable explanations for a decision provided by an automated system", as well as "the capability to provide an audience-appropriate, human-understandable interpretation of why and how a model provides certain predictions".

Liu et al. (2023) cite and adopt definitions from two other works. In this case, explainability is defined as either simply "the ability to explain or to present in understandable terms to a human" (Doshi-Velez and Kim, 2017), with no explicit reference as to what should be explained; or as a sort of metric, "the degree to which a human can understand the cause of a decision" (Miller, 2019), with no further detailing as to what kind of knowledge this human is expected to have beforehand in order to understand the explanations under evaluation.

There were another three, more indirect, definitions for explainability and closely related terms discovered during the review. When talking about the term, Barredo Arrieta et al. (2020) mention that explainability is "associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans". Loh et al. (2022) define only the expression explainable artificial intelligence, and simply as "the set of features that explain how the AI model constructed its prediction". Finally, Nannini et al. (2023) make reference to the synonymous 'explicability', more widely used in the European context, mentioning its definition as "a principle connected to transparency, crucial to creating and maintaining users' trust in the AI system", and corresponding to "the epistemic condition to comprehend and challenge decisions of AI systems".

Lastly, there are two important definitions of types of explainability that will be important in the rest of this chapter as well, so we will take this opportunity to present them as relayed by Bhattacharya (2022a). The first pair involves the distinction between **local** and **global explainability**, and it depends on whether it deals with "single local instances of the data to understand how a certain range of values or specific categorical value can be related to the final prediction" or if it is used "to explain the behavior of the entire model or certain important features as a whole that contribute toward a specific set of model outcomes". The second pair differentiates between **intrinsic** (also called 'inherent') and **extrinsic** (also called 'post hoc') **explainability**: under the first case, we have models in which "we clearly know the logic or the mathematical mapping of the input and output that the algorithm applies", while the second case "is about first training an ML model on given data and then using certain model explainability techniques separately to understand and interpret the model's outcome".

5.3.4 Interpretability definition

Interpretability is a term that revolves around an equally complex problem that does not have a single problem space (Meador and Goldsmith, 2022). Some of the texts under review stress that there are no formal or well-agreed upon definitions regarding the ways artificial intelligence models should be interpreted, which gives rise to loose definitions about interpretations and interpretability, with motivations around it being diverse and discordant, and the term itself often lacking precise meanings or receiving circular definitions (Li et al., 2022; Räuker et al., 2023).

Whatever the case, here we have grouped some of the recurring, and often disparate, definitions of interpretability, under three different guises. The first of these is closely related to how the term interpretability was defined in Subsection 5.3.2, in contrast with explainability. That means interpretability as dealing more with **comprehending the results** of artificial intelligence systems, for instance predictions, **and what they mean by potential user communities**, than with the inner workings of models themselves. Examples of these include the mention by Nannini et al. (2023), that interpretability can revolve around "model-agnostic approaches that are based on observing systems' inputs and outputs", and the assertion by Meador and Goldsmith (2022) that interpretability is "what makes a good outcome sustainable".

The same authors also state that an "[i]nterpretation is being able to parse information to know why something happened", exemplifying it with an analogy of being able to interpret the error message from your browser to understand that the Wi-Fi is down. One of the meanings given for interpretations by Li et al. (2022) reinforces the idea that, in this sense, interpretability involves domain-specific knowledge, since it deals with discriminative features used for model decisions or the importance of specific training samples as they contribute for inference.

The second meaning for interpretability actually relates it more closely to the way the term explainability was defined in Subsection 5.3.2, having to do more closely with **understanding an artificial intelligence system or model's mechanisms, rules, inside reasoning, inner workings and internal structures and representations**. This was the point of view most prevalent for instance in the works of Barredo Arrieta et al. (2020); Bruce and Fleming (2021b,a). The first set of authors indicate that interpretations should provide features like "an understanding of the model mechanisms and predictions, a visualization of the model's discrimination rules, or hints on what could perturb the model", while the latter two texts focus on yielding "a set of rules that define the predicted class of an outcome variable", calling these "the epitome of interpretability", useful in diagnosing issues and helping debug artificial intelligence systems and models.

In the same vein, Bhattacharya (2022b) asserts that interpretability has to with addressing "the how questions", and Meador and Goldsmith (2022) remind their readers that it is possible to give an interpretation of a hypothetical model even without being able to fully explain how it works.

Finally, Nannini et al. (2023) stress that when they talk about interpretability they refer to the "inner relations" that artificial intelligence systems learn to arrive at their computation functions, which is an understanding

shared by Räuker et al. (2023) when they mention that an interpretability method is "any process by which an AI system's computations can be characterized in human-understandable terms".

The third and last grouping of definitions of **interpretability** focuses on it **as a capability, characteristic or intrinsic property of a model, optionally measuring how predictable or understandable its inferences are** (Li et al., 2022). This sort of definition frequently also has an extensional quality, that is, it is applied to a certain family of models or algorithms that are ascertained as inherently transparent in contrast with others that are inherently opaque. This is where we find for instance a definition from Barredo Arrieta et al. (2020), with interpretability refering to "a passive characteristic of a model referring to the level at which a given model makes sense for a human observer".

Similarly, Bhattacharya (2022b) remarks that "[h]ighly interpretable models, such as linear regression and decision trees, are primarily linear and less complex", while being restricted to "learning only linear or less-complex patterns from the data". Li et al. (2022) mention as examples of models that are "fully interpretable without too much controversy: a set of limited number of rules; a depth-limited decision tree; a sparse linear model", Simon and Barr (2023) cite "decision trees, rule-based algorithms, and linear regression models, [as] intrinsically interpretable, simulatable, and decomposable", while Liu et al. (2023) mention as examples of intrinsically transparent and interpretable models "decision trees and linear regression".

5.3.5 Relevant additional terms

In this subsection we present definitions for important additional terms inhabiting the semantic universe of explainability/interpretability that were also found during the review, albeit with considerably less frequency than the former two. We start with a set of terms defined in Barredo Arrieta et al. (2020), the text that probably has the largest amount of definitions (and probably the most well-thought as well), which is to be expected since, as we already mentioned, that is part of the authors' explicit objectives.

The first of these terms is **transparency**, which is considered the opposite of black-boxness and defined as a quality of an artificial intelligence system that, by itself, already supplies "a direct understanding of the mechanism by which a model works". The same authors subdivide transparent systems into three categories, namely simulatable models, decomposable models, and algorithmically transparent models, although diving into the specificities of each

one is out of the scope of the present work.

The second import term from Barredo Arrieta et al. (2020) we would like to retain is **understandability** (which the same text also equivalently dubs 'intelligibility'), which denotes the capacity of an artificial intelligence system for enabling a human to understand its purpose "without any need for explaining its internal structure or the algorithmic means by which the model processes data internally". It is also a way of measuring "the degree to which a human can understand a decision made by a model", and surfaces as "the most essential concept in [Explainable Artificial Intelligence]". Understandability also depends upon the concept of **audience**, which is a way of taking into account "the cognitive skills and pursued goal of the users of the model" jointly with the term we will consider next.

The last term we would like to retain from the mentioned authors is thus **comprehensibility**, which, in the context of machine learning models, refers to "the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion[, ... resulting in] symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities". It is a concept connected to understandability since it also relies on the capability of specific audiences to understand the knowledge contained in the model, but deals more directly with the substantive information that the system is processing, by means of single 'chunks' of information that should be "directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion".

A trio of interesting and tightly related concepts arise from Bruce and Fleming (2021a)'s proposal of using so-called 'interpretability methods' to generate explanations for the predictions of underlying models. As the author mentions, "[j]ust as models themselves can underfit or overfit their data, so, too, can interpretability methods underfit or overfit their underlying models". Thus, what is known as bias (underfitting) and variance (overfitting) in the relation between a model and the data it is trained on, finds as correlates here the ideas of robustness and fidelity. **Robustness** is defined as "the sensitivity of explanations generated by an interpretability method to varying characteristics of the underlying model", with highly robust explanations being less sensitive to minor changes in a model's characteristics as it is tweaked. **Fidelity**, accordingly, is defined as "how well an explanation approximates the behavior of the underlying model", with high-fidelity explanations differing more easily after the underlying model changes due to how closely they approximate it. The same text also remarks on the trade-off between robustness and fidelity, with

its ideal management resulting in an interpretability method's **reliability**, to which having a wide range of interpretability methods at disposal contributes.

The final term we would like to delve into here is **trustworthiness**, starting with the way it appears in Li et al. (2022). Referred to as a 'notion', trustworthiness is defined as the reliability or faithfulness of the interpretations produced by algorithms compared to a model's behavior. As the authors say, "[a]n interpretation algorithm is trustworthy if it properly reveals the underlying rationale of a model making decisions". By that expression, the authors have in mind "all categories of information that help to understand the model", adding that trustfully recovering the rationale of a model is independent from the model's correctness, being one of the ad hoc requirements to help verify if "the information provided by the interpretation algorithm can be trusted".

The same term appears, with a slightly different connotation, in the survey conducted by Liu et al. (2023), defined as "programs and systems built to solve problems like a human, which bring benefits and convenience to people with no threat or risk of harm", with a focus on their transparency, by which the authors mean being able to understand their internal mechanisms. On their review focused on legislation, Nannini et al. (2023) mention that artificial intelligence trustworthiness is composed of both explainability and interpretability, adding that, especially when concerning European regulations at the time of the study (which according to these authors employs only a coarsely defined terminology), these concepts were actually weakened in favor of "ensuring the traceability and oversight of high-risk AI systems rather than empowering end-users' explanation demands".

5.4 Effects and ways of achieving explainability/interpretability

The next subsections describe the benefits of explainability/interpretability, the risks of the lack thereof, and techniques for achieving them. In accordance with the distinction between explainability and interpretability that we adopted in the previous section, from now on we will refer to each term with their specific meaning, even when the texts under review use them more haphazardly. In the following subsections, that means we will always attempt to present the benefits of explainability first, and then the benefits of interpretability later. The only exceptions for the term usage defined here will be direct citations, where we will retain the original terms used by the authors whenever that does not risk comprehension, and indicate necessary alterations otherwise.

As a reminder of the definitions we will be working with, by explainability we refer to ways of understanding the mechanisms and *inner workings* of artificial intelligence systems and models, including how they operate and why they arrive at specific conclusions, and are usually of more direct interest to their developers. By interpretability, on the other hand, we refer to ways of understanding what systems' and models' outputs are, what they mean and why do they matter, including what kind of response they are supplying and what do they entail in real-world practice, being of special concern to policymakers and other stakeholders affected by them.

5.4.1 Benefits of explainability/interpretability

We have clustered the benefits of aiming for explainability into four overarching groups, roughly accordingly with the following themes: informativeness or transparency, interactivity or accessibility, performance improvement, and discovery making.

The first set of benefits of trying to improve the explainability of artificial intelligence systems has to do with **informativeness** or transparency, an objective that is very tightly coupled with the stated goal of the first concept itself: having more information helps users in their decision-making processes, supporting them in their activity of relating their choices with solutions given by automated models (Barredo Arrieta et al., 2020). That is usually obtained by having transparency, allowing the inspection of the internal workings of models and architectures, including being able to understand how the model might behave when introduced to new data (Chaudhari et al., 2021; Bruce and Fleming, 2021a; Loh et al., 2022; Meador and Goldsmith, 2022). Informed users gain higher situational awareness and end up being empowered to make better decisions, so explainability helps to ensure that individuals have meaningful information about the logic involved in predictions (Simon and Barr, 2023; Nannini et al., 2023).

On a related note, some of reviewed texts argue that being able to extract information about the inner relations of a model also improves the overall mental model of the audience, allowing for more knowledge transference (Barredo Arrieta et al., 2020). That includes some forms of mechanistic comprehension as well, with the added benefits of being able to reverse engineer successful models and supply counterfactual examples, both of which are considered "ambitious but potentially very valuable goals" (Räuker et al., 2023).

In addition, one of the texts also mentions how greater informativeness

aids in system transferability, since it eases the task of elucidating the boundaries that might affect a specific model, supporting the reuse of acquired knowledge in related types of problems, as well as in privacy awareness, since being unable to understand what kind of information has been captured by the model and stored in its internal representation may entail privacy breaches that could otherwise be avoided (Barredo Arrieta et al., 2020).

A different set of benefits has to do more with **interactivity** or accessibility, more connected to the ways that end users, especially those with less technical expertise, can get more involved in the process of dealing with and possibly tinkering with existing models and algorithms that would seem incomprehensible at first sight. Explainability then eases the burden that non-technical users can feel when employing these systems, possibly becoming more involved over time in their development and improvement (Barredo Arrieta et al., 2020).

One of the major benefits of explainability, commented on by several of the texts under review, involves **improving the performance** of artificial intelligence systems and models. Later on we will see how aiming for explainability can sometimes be at odds with achieving higher scores in metrics like accuracy, but for now we will focus on the opposite case. As a considerable number of the reviewed literature states, achieving greater explainability can positively impact a system's performance, enhance its development process, and even shine light on opportunities for debugging, especially in deep neural networks, thanks to being able to get closer to the root of incorrect predictions when they happen (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021b; Loh et al., 2022; Bhattacharya, 2022a; Räuker et al., 2023).

More directly, when higher explainability allows developers to learn why a model is performing poorly, that can give rise to better feature engineering, hyperparameters fine-tuning, allow for better training strategies, investigate failure cases to design better prompts in the case of generative systems, as well as deal more swiftly and satisfactorily with model drift through appropriate monitoring practices (Meador and Goldsmith, 2022; Li et al., 2022; Zini and Awad, 2022; Liu and Zaharia, 2022; Räuker et al., 2023).

Not only that, explainability methods can also confer better insight into additional quality features of a system or model, going beyond simple performance metrics, and facilitating channeling human experience and intuition to improve them, correcting their potential deficiencies, including detecting deception when that is the case, which is a particularly hard task to do with generative natural language applications, increasing the number of ways humans are able to evaluate artificial intelligence systems (Bruce and Fleming,

2021a; Bhattacharya, 2022a; Simon and Barr, 2023; Räuker et al., 2023).

Lastly, and just as importantly, explainability is also a way of **improving discovery**, whether that means eliciting new domain knowledge that involves a specific task at hand, which is sometimes picked up automatically by artificial intelligence systems and models, or learning fresh capacities, reaching new findings and even understanding novel limitations about artificial intelligence systems themselves, possibly giving rise to a plethora of scientific discoveries in the process, eliciting latent knowledge and allowing for reverse engineering more understandable or verifiable solutions (Bhattacharya, 2022a; Li et al., 2022; Zini and Awad, 2022; Liu and Zaharia, 2022; Simon and Barr, 2023; Räuker et al., 2023).

Directing our attention now to the benefits of aiming for interpretability, these were also grouped according to the following four major themes: fairness and impartiality, robustness, causality, and trustworthiness.

The first major kind of benefit interpretability can confer to artificial intelligence systems involves **fairness** and impartiality, both with regard to the input data and the predictions outputted. In this case, that usually means detecting, highlighting, and consequently allowing the correction of different types of hidden bias (usually defined from a social standpoint) in the training data, in order to avoid unfair results and uses, and support more impartial decision-making (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021a; Liu and Zaharia, 2022; Simon and Barr, 2023).

Several of the texts under review mention that aiming for fair results not only helps a model perform better, but is also an important ethical consideration, especially when dealing with applications that influence human lives, with potentially severe consequences, and which becomes an even more crucial endeavor when people, particularly from disadvantaged groups, who may be the most adversely affected by these technologies, are involved (Chaudhari et al., 2021; Liu and Zaharia, 2022; Räuker et al., 2023; Barredo Arrieta et al., 2020)

Another benefit mentioned has to do with **robustness**, which involves how stable and reliable results are, especially in the context of adversarial perturbations. The relationship between interpretability and robustness can go lots of different ways: while designing more interpretable models helps combat vulnerabilities against adversarial perturbations, investing in more robust systems also helps them become more interpretable as well, which becomes especially relevant when confidence in general, conceived as a generalization of both robustness and stability, is expected from a reliable critical application. Another case involves creating more adversarial examples that can challenge,

help validate, and improve an existing system, since interpretability tools are particularly well-suited to creating adversaries (Barredo Arrieta et al., 2020; Li et al., 2022; Simon and Barr, 2023; Liu et al., 2023; Räuker et al., 2023).

Interpretability thus helps with another aspect of robustness: dealing with drifting, which includes monitoring data, including outliers, and tracking how both data and models can drift not only during development but also after deployment. That involves comprehending not just surface issues with data, but also its deeper characteristics and how datasets that are used in offline model training differ from the ones encountered in online operation (Liu and Zaharia, 2022; Meador and Goldsmith, 2022).

Finding **causality** among data variables is another considerable benefit of reaching for interpretability, even more so in the context of machine learning applications that only discover correlations among the data they learn from, which is not sufficient for unveiling cause-effect relationships. However, since even if correlation does not imply causation, the former is involved in the latter, so interpretable models are more apt at validating results provided by causality inference techniques or providing a first intuition of possible causal relationships within the available data. As such, interpretability can help guarantee truthful reasoning, acting as insurance that only meaningful variables are responsible for inferring an output (Barredo Arrieta et al., 2020; Liu and Zaharia, 2022).

Lastly, and just as relevantly, interpretability helps with improving **trustworthiness**, a benefit that has several different and related meanings in the texts under review, and which was by far the one that was more consistently mentioned by them. As one set of authors mentions, the search for trustworthiness is probably the primary aim of interpretability in artificial intelligence systems, and can be "considered as the confidence of whether a model will act as intended when facing a given problem" (Barredo Arrieta et al., 2020). Before detailing the term further, we should mention three other important characteristics related to it: it is not an easy property to quantify, not every trustworthy model can be considered interpretable on its own, and it is just as necessary to allow humans to know when a model may be untrustworthy (Barredo Arrieta et al., 2020; Räuker et al., 2023). We will come back to this last point soon.

Interpretability in artificial intelligence systems helps with combating mistrust from distinct audiences, which can include even people from inside the organization deploying them, including but not limited to their direct users, clients or operators, but also auditors, which can support increased adoption rates (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021b,a; Bhattacharya,

2022a; Loh et al., 2022; Zini and Awad, 2022; Liu and Zaharia, 2022; Simon and Barr, 2023; Liu et al., 2023). In that process, understanding and contextualizing individual stakeholder needs are essential actions to cultivate trust in artificial intelligence systems (Simon and Barr, 2023).

A couple of texts under review mentioned that trustworthiness can also be a direct source of increasing an application's ability to generate value, predominantly because, since interpretations are artifacts derived from artificial intelligence systems that can guide users in deciding what to do next based on model output, it has been discovered in research that that users informed on both model results and reasoning generally take a shorter decision time than those who are only given the model results (Meador and Goldsmith, 2022; Simon and Barr, 2023).

Finally, having more trustworthiness also entails being able to better defend a system or model in a legal context, for example with regulators, legislators, or civil litigators, meeting regulatory requirements, algorithmic transparency guidelines, evidencing risk mitigation strategies and fallback plans. That is increasingly more crucial, since regulations have empowered individuals with the right to demand trustworthy clarifications regarding decisions that can affect them and were based on automated systems (Bruce and Fleming, 2021b; Bhattacharya, 2022a). Here we come back to how essential it is to be able to also adequately elucidate wrong predictions and determine proper accountability for them, establishing responsibility in the case of misuse or deployment failures (Räuker et al., 2023).

5.4.2 Risks of non-explainability/non-interpretability

In this subsection, we deal with explicit mentions in the review of the risks associated with employing systems without aiming for explainability and interpretability. While several risks can be inferred from simply not putting to use the benefits detailed in the previous subsection, we found it nonetheless relevant to specify here the issues that received special interest in the texts under review. They can be grouped into four categories: low acceptance, biased predictors, inscrutable predictions, and harmful outputs.

We begin with the risk of artificial intelligence systems or models **getting less acceptance** than they otherwise could if they were more explainable and interpretable. This is particularly the case with less technically savvy users, and even more importantly in domains that already rely (or should rely) on evidence-based practices to guide decision-making, like healthcare, justice, and autonomous driving, to name a few. As the research under review

shows, predictions with low explainability and interpretability often have their credibility under question, which is even more significant when they are not forthright with their capabilities and above all their limitations, so the negative public perception can lead directly to lower acceptance of potentially useful technology (Loh et al., 2022; Li et al., 2022; Zini and Awad, 2022; Nannini et al., 2023; Liu et al., 2023)

Next we encounter the risk of **basing decisions on unjustifiable predictors**. Without explainability and interpretability, it becomes impossible to tell if biased or otherwise inappropriate predictors have been included as part of the system or model, especially when recourse is made to a voluminous amount of low-level predictors, like pixel values or language tokens. The risk of employing, even inadvertently, 'red-flag' predictors (for instance, using race for evaluating loan-approvals), that ignore algorithmic fairness guidelines, involves dealing with an insurmountable ethical issue (Bruce and Fleming, 2021a). Failing to detect undue human bias, particularly in high-stake domains, risks not only generating unfair predictions but also perpetuating, engendering and amplifying said biases. Developing and utilizing systems and models that do not account for those possibilities (and for how explainability and interpretability can help tackle these issues), pose a major risk, which can be alleviated by remembering that these are human-produced artifacts that very possibly reflect prior sociotechnical challenges in the choices that inform their design, and acting on this sort of reflection (Bhattacharya, 2022b; Zini and Awad, 2022; Simon and Barr, 2023; Nannini et al., 2023).

A different, even if closely related risk, is that of, whether they are accurate or not, **using predictions that cannot be justified**. As we have already seen, these can lead to severe issues of deployment, communication, error-correction, completeness, knowledge acquisition, and legitimate usage, all of which can be avoided or at least mitigated with recourse to explainability and interpretability methods that help unlock information that would otherwise be hidden inside opaque systems or models (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021a,b).

Just as importantly, some of the texts under review explicitly mention a risk closely associated with the previous two, which involves systems lacking in explainability and interpretability **actively causing harm** through their predictions. That has led to the creation of far-reaching regulatory efforts aimed at ensuring that no decision is made solely on the basis of data processing (Barredo Arrieta et al., 2020). Faulty and biased model outputs evidently can have prejudicial consequences at times, but even when they are working as intended, unchecked artificial intelligence systems, when deployed

with the goal of reinforcing user engagement based on profiling, can also be responsible for fostering hate communities, leading to social unrest, for instance based on gender or race discrimination. That is a risk that explainability and interpretability methods can help combat, making sure that algorithms are not blindly operated on auto-pilot without taking into consideration their potentially damaging unintended side-effects (Bruce and Fleming, 2021a; Bhattacharya, 2022b).

5.4.3 Techniques for explainability/interpretability

There were several taxonomies for explainability or interpretability methods found during the systematic review, and since none of them conflicted with one another, we will present here a potentially unified version that allows for each specific technique to be properly distinguished as needed. In other words, the different criteria that can be used to categorize the methods, except when explicitly noted, are orthogonal to one another.

5.4.3.1 Model or algorithm type

The first and probably most significant distinction between systems, models and algorithms that aim for explainability and interpretability has to do with their specific type, and was significantly prevalent in the literature under review. On the one hand, we have systems that are considered to be transparent, while on the other we find the so-called black-boxes. We'll devote our attention here to one type at a time.

Transparent systems or models, also called intrinsically or inherently interpretable, or having ante-hoc explainability, are those that, by their own characteristics, are considered easier to study, and to generate explanations and interpretations from. They generally have simpler structures than their counterparts, and mainly consist of systems that are based on either rules, regression analysis, or decision trees (Barredo Arrieta et al., 2020; Zini and Awad, 2022; Liu and Zaharia, 2022).

Black-box systems or models, also called extrinsically or externally interpretable, or having post-hoc explainability, conversely, are those that can only be adequately comprehended 'after the fact', through attempting to dissect layers of hidden knowledge using both syntactic and semantic lenses, through the use of additional specific techniques that can produce explanations or interpretations that might not be reliable or can even be misleading. At the same time, there is practically no other way to go about it, since the systems

and models to which these methods are applied are by themselves, and almost by design, too complicated for any human to fully understand (Barredo Arrieta et al., 2020; Meador and Goldsmith, 2022; Zini and Awad, 2022; Liu et al., 2023).

We think it is necessary to stress that this distinction does not concern a lack of technical knowledge from the audience interested in comprehending these systems or models, since even the experts that conceive them consider that parts of them are, at least to the best of currently available knowledge, inherently opaque. There are specific guidelines that attempt to provide explainability or interpretability for these systems and models 'by design', making them 'self-explaining/self-interpretable'. However, that is usually done through two means: either a model's outputs are explained or interpreted with supervision by human experts, and some processing is done to relate those understandings to the model's workings, or one generates a transparent model based on the results achieved by a black-box model and proceeds to explain or interpret the former instead of the latter (Liu and Zaharia, 2022; Räuker et al., 2023). We will revisit these ideas over subsequent parts of this subsection.

5.4.3.2 Comprehension subject

The following division is very closely tied with the distinction we accompany in this work between explainability and interpretability, as defined in subsection 5.3.2. It deals with what the main subject of understanding for a specific technique is, or in other words, what is being targeted for explanation or interpretation: the inner workings of a system or model, or its 'outer' elements, as we shall now see.

When a system or model aims at explaining their internal structures and representations, the method employed to do so can be considered an **inner explainability** technique. These deal chiefly with understanding how different parts of a model work, for instance its architecture, or the attention weights of a transformer, or the possibly latent knowledge representations in a machine learning artifact. Here we could group any method used to generate explanations that aim to help humans 'open up' black-boxes and study their constituent parts (Bhattacharya, 2022a; Zini and Awad, 2022; Räuker et al., 2023).

On the other end we have **outer**[1] **interpretability** techniques, those that deal primarily with giving interpretations for a system or model's func-

[1]This adjective was not found directly in the literature under review, but is proposed here as a parallel to the preceding one.

tion based on their inputs and outputs, data and predictions, or even the computational artifact as a whole, abstracting at least a good part of its innermost mechanisms. These are techniques that focus on highlighting the evidence pertaining to a system's so-called 'decisions', giving interpretations on what supports them. To use an analogy found in the literature under review, while inner explainability techniques would be akin to, in order to understand a human's reasoning, studying their brain directly, outer interpretability techniques would be more similar to simply asking the person what they are thinking and receiving an adequate response (Zini and Awad, 2022; Räuker et al., 2023).

5.4.3.3 Model specificity

Not all explainability or interpretability techniques can be used for all kinds of systems or models, so they can be divided according to how they deal with model specificity. On the one hand we have **model-specific** techniques, which are usually but not always tied to aforementioned intrinsically comprehensible models or algorithms, like the already cited linear/logistic regressions, decision trees, and rule-based models, but also to others like Generalized Additive Models, and Bayesian networks (Bhattacharya, 2022a; Zini and Awad, 2022).

On the other hand, **model-agnostic** techniques, which tend to be preferred over over model-dependent approaches, as with the former even complex machine learning algorithms can be explained, to some degree, in principle can work no matter what type of system is attempting to be explained or interpreted. More often than not, these are aimed at dealing with individual predictions instead of systems as a whole, making them more adequate to deal with complex, differentiable architectures like the ones involved in artificial neural networks (Liu et al., 2023).

5.4.3.4 Input data conformation

When taking into account the type of input that an artificial intelligence system works with, their conformation is also something that can deeply affect what kind of explainability or interpretability technique can be employed. Data is usually categorized as being either structured or unstructured, with some typologies also including semi-structured as a third kind or as a subset of structured data.

The most typical example of **structured data** are tabular datasets which store relational elements, although graph-based databases are also fairly

common. These were initially the bread and butter of input data for artificial intelligence systems of all kinds, and are still fairly prevalent. Structured data is usually very machine actionable, due to, among other reasons, fairly often being purely quantitative (Bhattacharya, 2022a).

On the opposite end of the spectrum we have **unstructured data**, such as images and text, which are usually harder for machine learning systems to explain or interpret, since they usually try to identify granular-level features that are not straightforwardly intelligible to humans. Different input data modalities are thus specifically taken into account when choosing particular explainability or interpretability techniques, according to their suitability to the task at hand (Bhattacharya, 2022a; Zini and Awad, 2022; Simon and Barr, 2023)

5.4.3.5 Scope

Another majorly present distinction for explainability or interpretability techniques has to do with the scope of what they target, which generates at least two possibilities: local or global methods, depending on whether they are aimed at single instances of data, on one hand, or systems or models considered as wholes, on the other.

Techniques with a **local scope** work by segmenting the solution space for the task to which their models are applied, and provide explanations or interpretations for specific output instances, like predictions, simplifying the problem in less complex subspaces that can be more readily comprehended. As such, they tackle how individual model predictions are locally influenced by their individual feature values and weights, describing how each contributes (positively, negatively, or negligibly) to the model's outcome for a singular observation of interest (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021b; Liu et al., 2023; Nannini et al., 2023; Simon and Barr, 2023).

For example, one could learn that the prediction of a high value for a particular home was due primarily to its view in a particular direction (a feature that might operate similarly for certain other homes but not generally to all homes). These methods tend to have higher fidelity and lower robustness, since they are closer to a specific way of functioning that is ideally but not necessarily successfully generalized to new data (Bruce and Fleming, 2021b).

Examples of specific local methods found in the literature review with more than one mention include ConceptSHAP, SHapley Additive exPlanations, Local Interpretable Model-agnostic Explanations, Layer-wise relevance propagation, Saliency Map, Guided Backpropagation, and GradCAM, while

methods with a single mention include Case-based reasoning, Fuzzy classifier, Explainable Boosting Machine, Testing with Concept Activation Vectors, and Visual Attention Maps (see Figures 5.2 and 5.3).

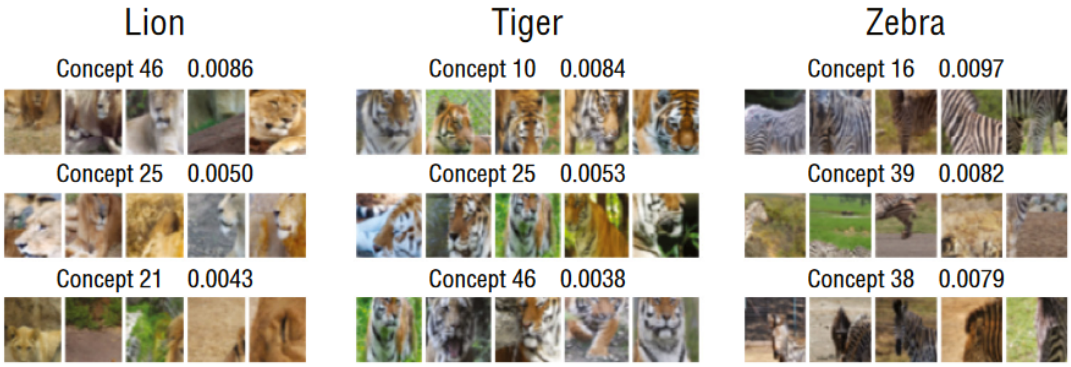

Figure 5.1: Depiction of a ConceptSHAP output for a convolutional neural network classification task. Each series of thumbnails shows the 'concepts' most associated with certain image classes. Source: Yeh et al. (2022)

Figure 5.2: An example of using Visual Attention Maps, images that give an interpretation of the most relevant areas in an automated caption generation task by illuminating them. Source: Bhattacharya (2022b)

When techniques are aimed at explaining or interpreting more than a single result of an artificial intelligence system, but are still not aimed at explaining it as a whole, the literature suggests the term **cohort scope**. That is the case when what is to be comprehended is a particular batch of data instances, grouped according to some specific criteria, to try and understand whether, for instance, predictions are being skewed according to some predetermined bias (Bhattacharya, 2022a; Liu and Zaharia, 2022).

On the opposite end we find techniques directed at explaining or interpreting artificial intelligence systems as wholes, in which case we talk about them having a **global scope**. They can do so in various manners, like measuring how individual features impact predictions globally (on average) across the

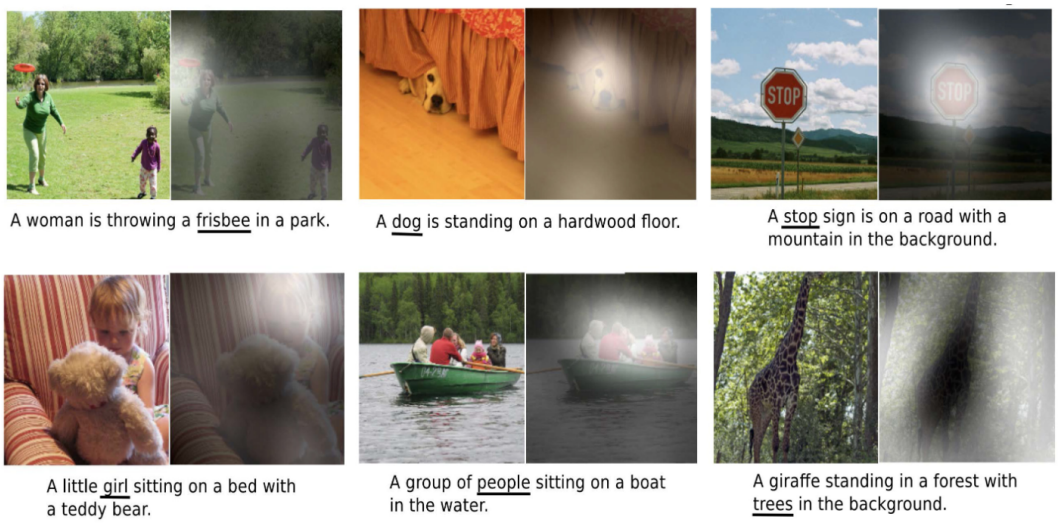

Figure 5.3: A different presentation of Visual Attention Maps, underlining the word that corresponds to the illuminated image. Source: Xu et al. (2015)

model, giving an overall vision of how it assigns weights to each one, usually based on employing local methods and (differently) aggregating their results. As such, global methods provide a holistic view of how a system or model behaves, possibly also including in its summary the choice of data and algorithms, being the closest analog in the case of black-box systems and models to the coefficients and decision rules provided by intrinsically intelligible ones (Bruce and Fleming, 2021b; Liu et al., 2023; Simon and Barr, 2023).

Examples of specific global methods found in the literature review with more than one mention include Permutation Feature Importance, Partial Dependence Plots, Individual Conditional Expectation plots, SHAP (which can also work in a local scope), and SmoothGrad, while the only global method the only received a single mention were Representation-Based Explanations.

5.4.3.6 Method of analysis

There are a few different methods used in explainability or interpretability techniques to analyze the most relevant attributes of the data and the parameters used in specific artificial intelligence systems. There were four such methods found during the systematic review: perturbation-based, gradient-based, proxy models and relevance propagation.

The idea behind **perturbation-based methods** is considered relatively simple: to investigate important features in the data, a straightforward way is to measure how a model's outcome changes as random perturbations are applied to different parts of the input, since larger changes would be observed for more important features. In other words, perturbation-based methods

leverage artificial variation on individual instances to construct comprehensible local approximations using linear models to explain resulting predictions or behavior. By far the two most popular perturbation-based methods include Local Interpretable Model-Agnostic Explanations (abbreviated as LIME, see Figure 5.4) and SHapley Additive exPlanations (SHAP, see Figure 5.5) (Li et al., 2022; Liu and Zaharia, 2022).

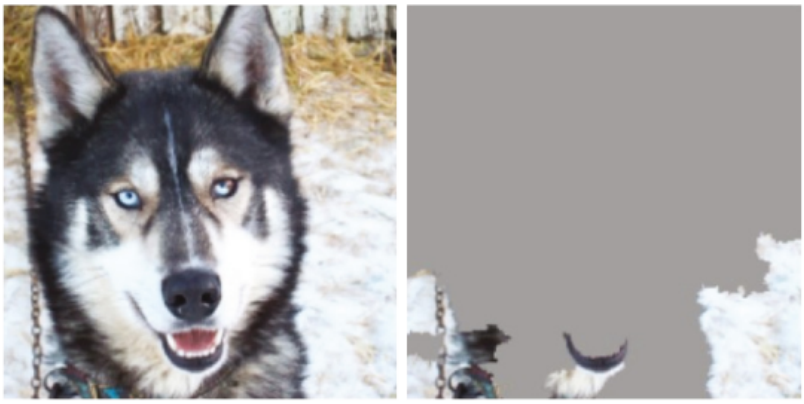

Figure 5.4: LIME showing the rationale behind a wrong prediction, with a husky misclassified as a wolf due to snow present in the background. Source: Ribeiro et al. (2016)

There are lots of ways that perturbations can be applied on a feature level, for instance replacing certain values with zero or random instances; picking one or a group of pixels (super-pixels) for an image explanation; blurring, shifting, or masking operations. In artificial neural network architectures, this can help characterize even the role of individual neurons, using datasets to comprehend which types of inputs they respond to or are maximally excited by, much like an electroencephalography does to human brain activity. Artificial neurons can equivalently be ablated, or even parts of the network dissected to try and establish causal rather than simply correlational relationship between the activations and the overall behavior and outcomes of the neural network (Liu et al., 2023; Räuker et al., 2023).

Gradient-based methods, in turn, employ partial derivatives on input instances to differently attribute importance to features in the data. A gradient is a concept from multivariable calculus that constitutes of a vector field – a function that can give the direction and the rate of fastest change at any given point in the problem space under consideration. Gradients can be applied to different modalities of data, with images and text being particularly relevant due to their usually high dimensionality (see Figure 5.6). Gradient-based methods tend to generate robust results, though risking model fidelity when

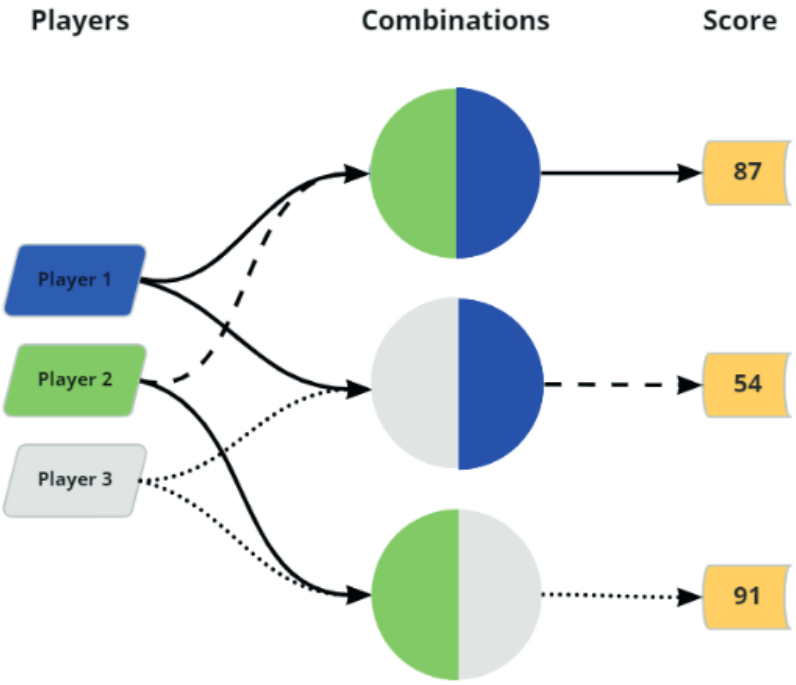

Figure 5.5: A simple diagram illustrating the main idea behind Shapley Values. Player 2 is the most impactful overall. Source: Meador and Goldsmith (2022)

doing so. The literature also indicates that recent work has started to unify these with perturbation-based approaches (Liu et al., 2023; Liu and Zaharia, 2022; Simon and Barr, 2023).

Proxy models, sometimes also called surrogate explainers in machine learning, are methods that work by simplifying the original computational artifact to be interpreted, and they work by rebuilding a whole new system based on the trained model under analysis. The new construct usually attempts to optimize its resemblance to its antecedent's functioning, while reducing its complexity and keeping a similar performance score. The simpler model chosen is invariably an intrinsically understandable one, easy to interpret while attempting to approximate the predictions of the original model as closely as possible (Barredo Arrieta et al., 2020; Bhattacharya, 2022a).

Popular choices for specific interpretable algorithms are linear regressions, decision trees and rule-based systems, and each one usually attempts to analyze three types of relationships between input features and target outcome: linearity (whether and how strongly each input feature is straightforwardly related to the predicted result), monotonicity (whether an overall increase to input feature values leads to an unequivocally directed change in the target outcome), and interaction (to try and gauge how individual features interact with each other to impact a model's conclusions) (Bhattacharya, 2022b).

Lastly we have **relevance propagation**, most often found in its version

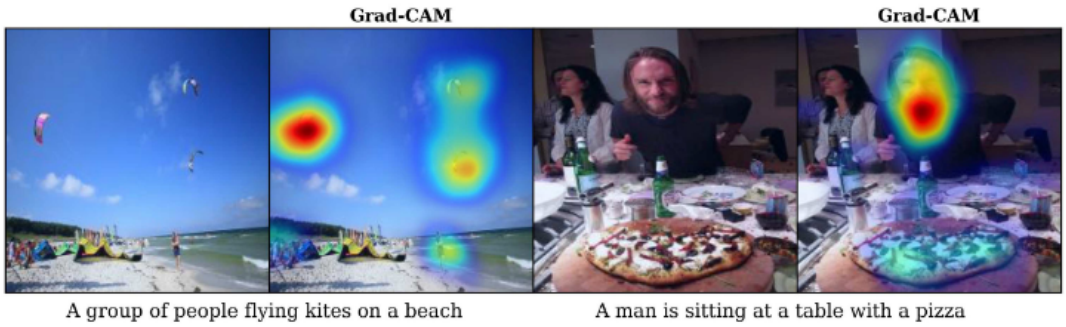

Figure 5.6: An example of Grad-CAM in a captioning task. The overlaid heatmaps indicate the most relevant areas for interpreting what influenced the generated text. Source: Selvaraju et al. (2017)

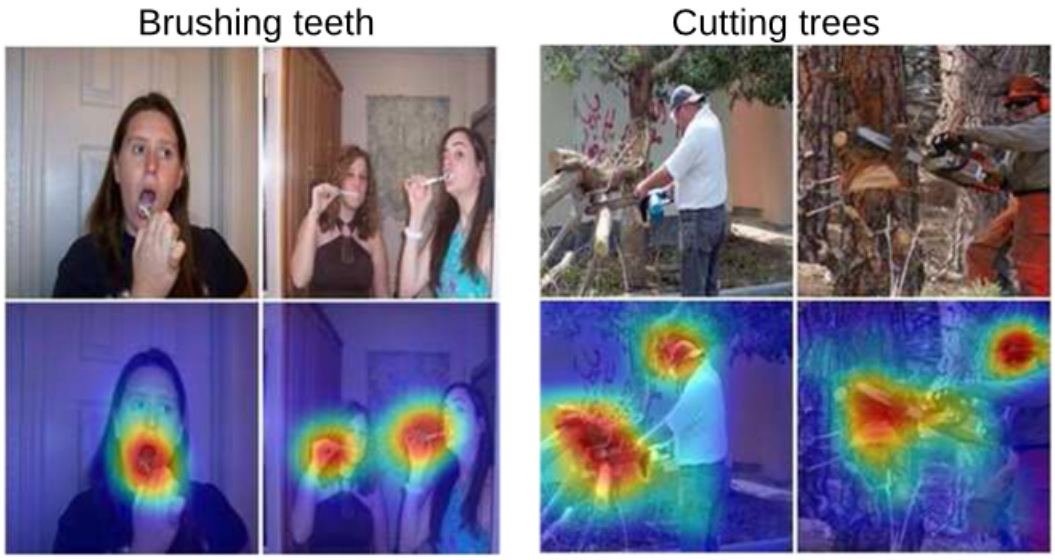

Figure 5.7: Another gradient-based example, this time in a classification task. Here, CAM's heatmap shows the areas relevant for the predicted class. Source: Zhou et al. (2016)

applied to artificial neural networks across model layers in image classification tasks, called Layer-wise Relevance Propagation, though it can also be applied to Support Vector Machine models. This method recursively computes a relevance score for each layer of computations, proceeding backwards from the last one, to try and understand the contribution of each feature (for instance, each pixel on an image, or word tokens in text) to the prediction function that classifies an input, generating for instance a heatmap superimposed on the original image, or differently coloring and weighted underlining the original text, that highlights the critical regions that most affected the prediction (Li et al., 2022; Loh et al., 2022; Bhattacharya, 2022b).

5.4.3.7 Output type

The output type refers to what is generally considered to be the explanation or interpretation effectively produced by explainability or interpretability techniques. There were four primary types found in the literature review, being: feature importance, example-based, model rationale, and dataset impact.

Feature importance, also variously known as feature relevance, influence-based outcomes, or sensitivity analysis, is the result of computing a relevance score for each variable managed by a model, which quantifies the relative effect a feature has on its outputs, indirectly explaining the model and interpreting its predictions. It is an attempt at showcasing direct associations between the input and output of an artificial intelligence system. This is most often done through some variation of permutation feature importance, which is a perturbation-based method that compares how much the model error increases when all the values of a feature are randomized, ranking them in descending order of importance according to the increase in error (Barredo Arrieta et al., 2020; Nannini et al., 2023; Bruce and Fleming, 2021a).

In other words, feature importance is a way to identify the dominant features in the data used to train a model or calibrate a system, and analyzing their effect on the predictions comparing the outcomes with known scenarios or exemplary situations. Feature importance is probably the most popular choice for root cause analysis that helps with detecting failures, understanding and debugging machine learning systems, and other than permutation importance scores there are also statistical correlation scores, decision tree-base scores, and others. It is a method often also employed for feature selection and dimensionality reduction steps in machine learning performance improving pipelines, though in those cases domain expert validation is always crucial before drawing any conclusions (Bhattacharya, 2022b).

Additionally, feature importance is an explainability or interpretability result from the analysis of structured datasets, where the features are clearly defined, being less relevant for unstructured data like text or images as the features or patterns used by the models in those cases can be significantly more complex and a lot less intelligible, if at all, to humans (Bhattacharya, 2022b).

Example-based responses work similar to one of the ways through which humans try to explain a new concept someone else: often, we try to make use of examples that our audience can relate to. What example-based explainability and interpretability techniques do is attempt to select certain instances of the original data to explain the behavior of the system, assuming

that observing similarities between the current observation and a historic one can be used to explain black-box models. This tends to be more challenging to do with high-dimensional structured data since, for these approaches, not all the features can be included to explain the model, so applying some feature selection method in advance helps (Bhattacharya, 2022b).

The generated example-based outcomes should ideally be user-centered, which here means comparing the result with end users' predictions based on assumed knowledge. Supposing the model forecasting does not match the user community's prediction, a good example should function to justify what changes could have happened in the input observations to get a different outcome, comparing system outcomes with known scenarios or situations (Bhattacharya, 2022a). Example-based methods' outcomes can be of three different subtypes: prototype examples, counterfactual examples, or adversarial examples.

Prototypes, in the context of example-based explanations or interpretations, are supposed to be the typical exemplars of their categories, for instance unequivocal classes in classification tasks, or extreme values in regression tasks. Since they exemplify paramount instances of the available data used for learning, they tend to be effectively and directly understood by humans, which helps in understanding the type of prediction a system makes. To create this set of exemplary data points, that allow explaining or interpreting a model's result by how the target data resembles one of them, activation maximization is one of the employed techniques, since prototypes can be found or even generated through an optimization process (Li et al., 2022; Räuker et al., 2023).

Counterfactuals, in turn, which are the most popular example-based explainability or interpretability output types, tend to indicate to what extent a particular feature has to change to significantly change the predicted outcome. Counterfactual examples work by supplying 'what if' scenarios, answering hypothetical questions about an instance of data through contrasting it with different, possibly existing, observations. For classification tasks, counterfactuals are generated by trying to predict the opposite class for a binary problem, or the most similar class for a multiclass problem. For regression tasks, the counterfactual is generated with recourse to different parts of the response spectrum, usually symmetrical to the targeted input, or close to the ends if the original observation is close to the average (Bhattacharya, 2022b; Liu et al., 2023).

Especially in critical applications where human elements are directly involved, it is crucial to provide concrete insights and recommendations that people can act upon if they wish to be classed otherwise (supposing, of course,

that they are in accordance with how the classification system works in the first place): for that reason, the counterfactual examples should have changes in the input that are as small as possible, but that would still significantly alter the prediction made by the model. Additional care must be taken to not only generate actionable counterfactuals, but to also avoid the so-called Rashomon Effect, since for any real-world problem it is possible to find multiple counterfactual examples that can contradict each other, creating confusion instead of comprehension. As usual, human intervention and the application of domain knowledge to pick up the most relevant example can help in mitigating this hardship (Li et al., 2022; Bhattacharya, 2022b).

Adversarials, lastly, are examples very closely related to counterfactual ones, generated with similar optimization methods, but with the additional requirement of altering data inputs in ways that are usually imperceptible to humans: for instance, changing the value of a single specific pixel (or set of pixels) in an image that results in it being labeled with a different class altogether, misleading the prediction. Adversarial examples are often used to reveal vulnerabilities in deep models and even to attack artificial intelligence systems, so their analysis can help understand the machine learning process and improve the robustness of the trained model (Li et al., 2022).

A different type of output from explainability and interpretability techniques involves attempting to give intelligibility to the **system's rationale**, for instance supplying information about the data or the model through automated knowledge extraction methods. This becomes especially relevant in deep models, since their full reasoning process is incredibly complex due to their nonlinearity and the enormous amount of computations involved in their creation and usage. To give the underlying system rationale human intelligible semantics, this process can be approximated with the use of graph models or decision trees, using the aforementioned proxy models. These rule-based systems can then, for instance, be queried or otherwise analyzed to generate semantically adequate explanations or interpretations to stand in for the original system's rationale (Bhattacharya, 2022b; Li et al., 2022).

Finally, one more possible output of explainability or interpretability techniques is analyzing and displaying the **dataset impact**, which deals with how specific sets affect machine learning dynamics, including a system's design. Here we find data-centric approaches like the ubiquitous Exploratory Data Analysis (whose importance, we would like to add, cannot be overstated), used to understand whether the data is consistent, well curated, and well suited for solving the underlying problem, as well as data profiling, which can help detect data and concept drifts, for instance. These are ways of extracting key

insights from the input data (and also from a model's predictions) that result in detailed reports that increase a system's explainability or interpretability (Bhattacharya, 2022a,b).

5.4.3.8 Output format

The results of explainability and interpretability techniques can not only be of various types, as we have just seen, but these can also come in different formats or modalities. The most prevalent are images, but textual accounts are just as important, with the more direct tabular resources also making an appearance. All of these formats can be coupled with each other to produce more complete responses, but we keep them separated here to focus on the specificities of each mode of presentation.

Visual results are usually the preferred outputs of several explainability and interpretability methods, since they are considered to be information-rich and more readily intelligible. They are considered the most suitable way to introduce complex interactions within the variables involved in a system or model to users not necessarily acquainted with machine learning modeling, and feature selection as well as dimensionality reduction can be employed to help produce visualizations, though the latter tends to incur a loss in comprehensibility (Barredo Arrieta et al., 2020; Bhattacharya, 2022b; Zini and Awad, 2022).

To visualize results like ranges of feature values, or the relative importance of some of a model's attributes, bar charts and diagnostic plots are the most common choices. To look at parameters like attention weights or pixel relevance, formats like saliency maps and highlight overlays and usually employed, indicating the spatial support of particular classes or how regions differently contributed to a specific prediction. Colors and intensities can also be used to indicate how each feature or area is associated with the system or model's outcome Bruce and Fleming (2021b); Chaudhari et al. (2021); Liu and Zaharia (2022).

Textual results are less prevalent than visual ones, and include every method that generates symbols that explain the functioning of a system or model or that help interpret the associations between their data and predictions. Textual outputs for explainability or interpretability techniques can be natural language sentences that convey how a model works internally, in general or for a specific prediction, or how different training data is related to a classification or regression result, in both cases ideally in ways that are understandable to the appropriate audiences. This kind of result can even

attempt to be automatically generated if explainability or interpretability are baked in the system by design, for instance using specific large language models for this task (Barredo Arrieta et al., 2020; Liu and Zaharia, 2022).

One last distinct format that merits attention are **data instances**, which can be almost purely numerical, but also visual, textual or even a combination of all of these. These output formats are usually tied to example-based interpretability techniques, since they involve providing pure instances of data as part of (or even as the entirety of) the explanations or interpretations for the system or model under consideration. These data instances can be real observations that come from training or testing datasets (if no privileged information will be leaked this way), or even synthetic constructions that help to make the system or model more intelligible by comparison (Barredo Arrieta et al., 2020).

Additionally, all of the previously mentioned outputs, regardless of format, can also be made interactive, with options that help the techniques' user communities further explore the explanations or interpretations supplied with additional information that can once again be of different format and levels of detail (Liu and Zaharia, 2022).

5.4.3.9 Choosing a technique

The choice of which technique to use not only has to consider whether the main goal involves explainability or interpretability, but is also heavily predicated on what the key problem under consideration is: the precise nature of the task and its domain, the complexity of the questions being asked or the behaviors being analyzed, the desired broadness of applicability, the architectures of both the hardware and the software involved, and above all the type of audience for which explainability or interpretability is important, including their presumed knowledge, specific requirements, and intents (Barredo Arrieta et al., 2020; Bhattacharya, 2022b; Zini and Awad, 2022; Nannini et al., 2023).

5.5 Key takeaways for additional exploration

5.5.1 Important references (for backward snowballing)

Several references piqued our curiosity during the reading in full of the texts selected for the systematic literature review. While all of them seemed relevant for our interests, their adequate appreciation falls beyond the scope

of the present endeavor. That activity constitutes a possibly fruitful avenue for future work, and might help to improve on the findings discussed in this chapter. They are listed in full in Appendix B.

5.5.2 Interdisciplinarity

The importance of interdisciplinarity for the consideration of explainability and interpretability in artificial intelligence systems was another topic that ended up drawing our attention during the literature review. We cannot be sure how much the author of this work's previous background in the social sciences contributed to that specific outlook, but we believe the subject was mentioned often enough in the texts, as either inter or multidisciplinarity (Barredo Arrieta et al., 2020; Simon and Barr, 2023; Nannini et al., 2023; Räuker et al., 2023), for it to possibly warrant the attention of any researcher studying the themes that we have been discussing here. Specific disciplines other than computer science and engineering mentioned in the review included psychology, ethics, law, as well as the social sciences in general.

Psychology, especially its cognitive strand, was one of the main disciplines mentioned in the literature concerning explainability and interpretability. Experiments in the field are said to be able to help in reducing subjectivity when assessing the creation of convincing intelligibility arguments, taking into consideration the cognitive skills, capacities and limitations of the individual human and their mental models. Concepts of this field can help to draw insights about convincing a person and measuring whether something has been understood or put clearly (Barredo Arrieta et al., 2020; Simon and Barr, 2023).

Philosophy is another discipline that also appeared as relevant in the review, especially the subfield of **ethics**. Considering ethical restrictions is said to be paramount in order to create human-centered applications and to specifically model the artificial intelligence components of a system. The ethical implications of predictions and decisions taken based on them gain special significance as machine learning, and especially deep learning, models are employed to address increasingly sophisticated problem solving, which necessitate reflecting on ethical explanations and interpretations (Barredo Arrieta et al., 2020; Simon and Barr, 2023; Nannini et al., 2023).

On a closely related note, the study of **law** appears as an equally relevant related disciplinary field that should be taken into account in interdisciplinary research about explainability and interpretability. The development and the results of these methods can have considerable impact due to regulations and policies, such as in the case of the European Union General Data Protection

Regulation's right for explanation. The question of a computerized system's legal liability is also an open and relevant one that is closely related to their intelligibility, and the idea of making laws that require organizations to reveal a certain level of understandability seems to be spreading, such that it might become an unavoidable requirement in the future (Barredo Arrieta et al., 2020; Loh et al., 2022; Meador and Goldsmith, 2022; Nannini et al., 2023).

Other disciplines in the **humanities**, aside from the previous two, were also found to be relevant to researching and developing systems considering explainability and interpretability as possible requirements, like the study of languages to help in collaborative sense-making, for instance by sharing the context producing their findings, wherein context refers to sets of narrative stories surrounding ways of understanding artificial intelligence systems; and even the arts, since there is a measure of creativity involved not only in dealing with data but also in choosing the best combination of explainability and interpretability techniques to be utilized for dealing with these technologies (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021b).

The most prevalent mention to other traditions when interdisciplinarity was the word of the day in the review revolved around the **social sciences** in general, with occasional mentions to sociology in particular. Looking at the social aspect, or the social ties of people involved in the development and deployment of artificial intelligence explainable and interpretable systems are said to be crucial activities that are still in progress, with plenty of room and interest for collaboration. The distinct sociotechnical impacts of these computational artifacts is also a theme that appeared central to interdisciplinary considerations (Barredo Arrieta et al., 2020; Correia and Lindley, 2022; Li et al., 2022; Simon and Barr, 2023; Nannini et al., 2023).

The social sciences were also mentioned as disciplines that can help analyze the complexities of evaluating what counts as good explanations and interpretations, by producing more sophisticated accounts of how to assess a system's performance that is not simply restricted to metrics like model accuracy. The diversity of outcomes, for instance, is thought to be an example of a significant indicator of the quality of both the data used for machine learning and of the distinct types of results that can be expected from applications themselves and associated intelligibility techniques, for instance producing diverse and nuanced suggestions rather than highly-scoring yet similar results in a recommendation task. Another example involves sampling techniques developed as a part of basic methods in the social sciences, a technique that can help to gauge variability or bias in the estimates of a system, and which involves replicating existing observations and resampling

in order to derive multiple estimates and see whether they correspond to what is known about the original population (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021a).

Another important aspect that the social sciences are said to aid in is the choice of evaluation questions and the adequate characterization of the audience for which explanations and interpretations should be delivered, since they should be differently accessible to complex sets of stakeholders. That involves determining the right level of abstraction and the content of the results that are conveyed, which in turn depends on the kind of understanding each community within a specific audience is supposed to possess. To deal with the difficulties that can arise in adequately generalizing explanations and interpretations, principles of user-centric design and studies on Human-Computer Interaction are invaluable allies, together with the social context in which these technologies are adopted. A considerable amount of human experimentation and creativity in the loop are also routinely called for in improving explainability and interpretability (Barredo Arrieta et al., 2020; Bhattacharya, 2022b; Liu and Zaharia, 2022; Nannini et al., 2023; Räuker et al., 2023).

One final topic should be mentioned as part of the contribution the social sciences can make for the consideration of explainability and interpretability, and that has to do with ways to ensure fairness and avoid undue bias not only in data but also in the development and functioning of artificial intelligence systems and models. Protected attributes such as gender or race that in most cases should not be encoded as part of a problem's dimensions can be removed from an architecture, and that can be ensured with recourse to adequate explainability and interpretability techniques. And care should always be taken in order to avoid creating systems and models that irresponsibly reflect prior sociotechnical disparities and harmful prejudices, which is something that frequently can be predicated on choices informing the design of applications (Zini and Awad, 2022; Nannini et al., 2023; Räuker et al., 2023).

5.5.3 Assessment and metrics

The results of artificial intelligence systems and models are often reduced to specific metrics, accuracy being chief among them but far from the only one. However, as the texts under review point out, there are lots of different ways of assessing performance, and explainability and interpretability related measurements can be just as important as the ones dealing with exactly how precise an application is expected to produce predictions. Test accuracy alone

can be an especially misleading metric, since that does not imply that the learned solution is sufficiently adequate or general, to use a couple of examples. Fixating on best-case scenarios for metrics can be equally illusory, which might end up overestimating the value of a proposed solution. (Barredo Arrieta et al., 2020; Li et al., 2022; Räuker et al., 2023).

All of the above considerations are equally important when it comes to taking into account the explainability and interpretability of a system or model as well. Even though assessing the quality of techniques that are responsible for improving systems' understandability is something challenging, also in part due to their considerable diversity in methodologies and varying levels of complexity, it is a crucial activity to undertake (Barredo Arrieta et al., 2020; Li et al., 2022; Loh et al., 2022).

Assessment of explanations and interpretations can be done through several different means, lots of them being quantifiable. The literature claims for not only objective but also specifically numeric metrics, in order to support the use of measurement procedures and tools for evaluating the explainability and interpretability of distinct systems or models. The texts under review agree that, at least so far, there exist few ways of doing so, especially in straightforward, quantitative manners, since most methods supply visualizations or natural language outputs, which are often notoriously difficult to quantify (Barredo Arrieta et al., 2020; Bhattacharya, 2022b; Li et al., 2022).

Having precise ways of assessing different explainability or interpretability approaches also allows for comparison methodologies contrasting them, and opens the possibility or creating rigorous benchmarks similar to the ones that enabled considerable improvement of artificial intelligence systems. However, the texts under review point out that there is no unified criteria for doing so, despite how relevant their creation appears to be, including the establishment of unique and well designed testing frameworks. (Barredo Arrieta et al., 2020; Li et al., 2022; Zini and Awad, 2022; Simon and Barr, 2023; Räuker et al., 2023)

At the same time, the importance of qualitative approaches is not understated in the literature, since trying to gauge simply how understandable to a human explanations or interpretations are unavoidably involves their subjectivity. The entire process is already largely predicated on human assessment to begin with, since for instance supervised learning already involves annotated data. Since user communities' subjective perceptions are crucial to this kind of assessment, and they can include personality traits, individual experiences, subject matter knowledge, beliefs, and cultural influence, Human-Computer Interaction expertise and knowledge from the social sciences is once again crit-

ical (Barredo Arrieta et al., 2020; Bhattacharya, 2022b; Li et al., 2022; Zini and Awad, 2022; Simon and Barr, 2023; Nannini et al., 2023; Liu et al., 2023). We would like to add that doing so helps not to conflate the presence of subjectivity with lack of preciseness.

The texts in the review mention several possible key criteria for assessing the intelligibility of specific systems or models. They revolve around how satisfactory, useful, educational, reliable, costly and customizable the outcomes of explainability and interpretability techniques are. Let us go over each one in turn.

The outcomes of explainability and interpretability techniques can be assessed according to **how satisfactory** they are to their user communities. Satisfaction here can involve how coherent they are with previous knowledge, addressing existing beliefs surrounding the problem at hand (even if the result can end up questioning prior assumptions), but also conciseness, the right amount of simplification or detailing. At the same time, explanations or interpretations that involve non-obvious outcomes are just as important for satisfactory responses, including spelling out causal links whenever possible, and allowing users to contribute to satisfaction scales (Barredo Arrieta et al., 2020; Bhattacharya, 2022a).

Another way of assessing how good explanations or interpretations are has to deal with **how useful** they are deemed to be. That means they should be actionable, taking into account the specific needs of distinct user communities. Contrastive outcomes tend to be particularly relevant in this case, for instance including different predictions for what-if scenarios that are actually within reach of a person being affected by a system's prediction. These are usually very application-dependent since they require not only points of comparison but adequate knowledge of the domain in question to know which features can be the subject of changes and within which ranges (Barredo Arrieta et al., 2020; Bhattacharya, 2022a).

How educational the results of explainability and interpretability techniques are is another form of assessing their quality, since ideally those outcomes should be able to induce the improvement of the mental models the audience works with. It is perfectly reasonable to expect that good explanations and interpretations should enrich users' conceptualizations and knowledge about the domain in question, even if that is not supposed to be their primary function (Barredo Arrieta et al., 2020).

One more important aspect for explainability and interpretability assessment is **how reliable** or trustworthy their outcomes are. Good explanations and interpretations are expected to be consistent and demonstrate high fidelity

to their underlying systems or models, being adequately free of undue bias. In addition, it is very advisable to pair techniques' outcomes with confidence estimates, for instance quantifying the amount of uncertainty involved in a prediction (Barredo Arrieta et al., 2020; Bhattacharya, 2022a; Räuker et al., 2023).

It is also important to consider **how costly** explainability and interpretability are to implement, not only because it might be difficult to maintain a model's accuracy if it also needs to be adequately understandable (a topic to which we will return in the next subsection), but because even straight up designing and developing good explanations and interpretations, as well as the means to assess them, usually depends on human participation, due to the lack of univocal ground truths and the relevance of subjective understandings. Those end up being more resource-intensive goals, that can even end up giving more investment return, but the associated costs also need to be taken into account (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021a; Bhattacharya, 2022a; Liu et al., 2023).

Finally, **how customizable** explanations and interpretations are is also something relevant for their quality, revolving around how much the expertise and assumed knowledge of the audience can be involved in the process. Intelligibility can be gained through an adequate amount of simplification or detailing, depending on each use case, bringing about the need for employing different typologies. Not only the purpose of understandability outcomes can vary across tasks and stakeholders, but also their complexity, completeness, interactivity, or format. And in case providing personalized outputs is on the table, ensuring compliance with data privacy can be even more important than usual (Barredo Arrieta et al., 2020; Bhattacharya, 2022a; Nannini et al., 2023; Simon and Barr, 2023).

There are myriad **ways** that explainability and interpretability can be effectively assessed and evaluated, including checklists, scales, measuring the size of outcomes, the complexity of the features involved, the cognitive processing time required to understand and successfully act upon an explanation or interpretation, as well as retraining machine learning models according to alterations based on the comprehension achieved and comparing how it behaves then, measuring the degradation of its new performance, employing adversarial samples, or even randomizing a system's hyperparameters or data samples before doing the same (Barredo Arrieta et al., 2020; Bhattacharya, 2022a; Li et al., 2022; Räuker et al., 2023).

In addition to those, several of the methods for assessing the quality of the explainability and interpretability of an artificial intelligence system or model

necessarily involve deeper human participation, with real tasks being evaluated. Those can vary in type from application-grounded evaluations, which involve including the comprehensibility technique to be assessed by domain experts in a real environment through an almost-finished product (a time-consuming and high cost experiment), to functionality-grounded evaluations, which does not involve direct human experimentation, instead dealing with proxy tasks to evaluate the quality of explanations or interpretations: these are not only the less expensive and usually faster to implement tests, and even though their results might also be less promising, they are particularly useful as alternative approaches in use cases where human subject experiments are restricted and unethical (Bhattacharya, 2022a).

Figure 5.8: Different levels of evaluation for the quality of the explainability or interpretability of an artificial intelligence system. Source: Bhattacharya (2022a)

In the middle ground there are also the so called human-grounded evaluations (see Figure 5.8), which employ non-expert users on more straightforward tasks in comparison (which are faster and less expensive than the ones belonging to the first type, but tend to yield less definite results), using methods like A/B testing, counterfactual or forward simulations. A/B testing provides different explanations or interpretations to users, who are asked to select the best one and generate an aggregated score based on the voting, as well as other possibly relevant metrics such as click-through rate, screen hovering time, and time to task completion. In counterfactual simulations the participants are presented with the input and output of the model as well as explanations or interpretations for a certain number of data samples, and are asked to provide certain changes to the input features in order to change the model's final outcome to a specific range of values or a specific category. Lastly, in forward simulations, participants are provided with the model inputs and their corresponding explainability or interpretability techniques, and then asked to simulate the model prediction without looking at the ground-truth values, contributing to the generation of an error metric used to find the difference be-

tween human-predicted outcomes with the ground-truth labels (Bhattacharya, 2022a).

5.5.4 Future challenges and possible drawbacks

This subsection is dedicated to the challenges and the possible drawbacks that are involved in explainability and interpretability, as reported in the literature. Future research avenues are addressed in the next chapter.

The first challenge has a very straightforward connection with our initial objective, since it involves **reaching consensual definitions**, not only for the concepts of explainability and interpretability, but also for how they should be employed. Agreement on the vocabulary and the meanings of the terms used in the field is considered an imperative, since there are numerous cases where there is no consistency on what is being taken into account. One of the reasons for that is the field having a relatively young quality, without a standardized terminology, which the texts consider is achievable, reducing ambiguities. In the same vein, attempting to define what constitutes good outcomes for explainability and interpretability techniques, and how to best utilize them, are active research subjects. At the same time, that should be done without forgetting that even the definitions, quality indicators and proposed usages must be tailored to a diverse community of stakeholders and a manifold of purposes (Barredo Arrieta et al., 2020; Meador and Goldsmith, 2022; Nannini et al., 2023; Zini and Awad, 2022).

The second challenge is closely related to the first, since it deals with **creating benchmarks** for adequately measuring explainability and interpretability. The derivation of general metrics to assess the quality of these requirements remains an open issue, and these are important to allow meaningful evaluations and comparisons between different methods that produce useful insights about a system or model. The costs involved in creating these metrics and benchmarks is far from negligible, especially since in several cases their validation necessitates human interaction in real environments. However, the potential for these benchmarks to improve not only the outcomes of explainability and interpretability techniques, but also the original applications themselves is significant (Barredo Arrieta et al., 2020; Liu and Zaharia, 2022; Nannini et al., 2023; Räuker et al., 2023).

Next we find the **relation between accuracy and intelligibility**, sometimes also called the performance versus transparency trade-off, when the latter is not considered as part of the former. While some texts indicate that these are necessarily opposite to each other, thinking that an exclusive focus

on performance leads to systems being increasingly opaque, the literature also mentions that this is a statement filled with misconceptions. As others texts under review indicate, explainability or interpretability, on one hand, and accuracy or privacy preservation, on the other, are not necessarily contradictory measures, as they can be improved simultaneously and even causally (see Figure 5.9). The relationship between these terms is thus a complex, intricate, and contextual one, and others factors like cost increases, growth of undue bias, higher computing demands, and loss in usability, all have to be considered in order to adequately assess how performance, in a more encompassing definition, is affected when explainability and interpretability are part of the equation (Barredo Arrieta et al., 2020; Bhattacharya, 2022a; Li et al., 2022; Nannini et al., 2023; Räuker et al., 2023).

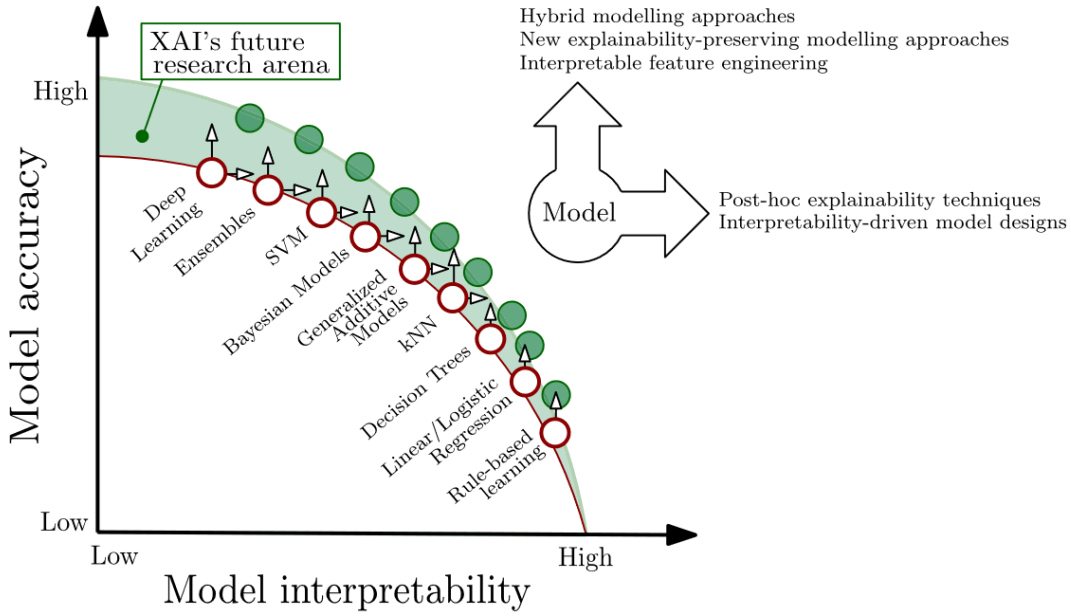

Figure 5.9: An abstract representation of the trade-off between explainability/interpretability and accuracy, and the potential improvement techniques could bring about. Source: Barredo Arrieta et al. (2020)

Another crucial challenge described by the literature is **including confidence measures** as part of explainability and interpretability outputs and assessments. That also means dealing openly with uncertainty and doing your best to inform your audience about the approximation dilemma, that is, the fact that simplifications are usually needed to convey explanations and interpretations about complex systems, while losing fidelity in the process. Users should then be informed about the share of epistemic uncertainty (namely, the uncertainty due to lack of knowledge) involved in an outcome, which becomes particularly relevant for instance in the utilization of large language models,

since their outputs might not reflect factual information. And while more complex models might enjoy more flexibility than their simpler counterparts, it is not necessarily true that they are inherently more accurate. In that vein, also estimating the confidence of the explanations or interpretations themselves is thus still important work to be done (Barredo Arrieta et al., 2020; Bruce and Fleming, 2021b; Nannini et al., 2023; Räuker et al., 2023).

A different challenge around explainability and interpretability revolves around **arriving at the reasoning** behind a system or model's outcomes. That is even more important when harmful predictions are made by applications, since simply recognizing them is not enough if one does not arrive at how they are made in the first place. Since even when there are no true relationships to be found within the data, models can produce seemingly reasonable correlations based on it, identifying the exact root causes that lead to a prediction becomes paramount. Additionally, systems that can demonstrate their reasoning also become more resilient to adversarial attacks (Bruce and Fleming, 2021b; Liu et al., 2023; Simon and Barr, 2023).

One more specific challenge indicated by the reviewed texts is **using attention weights** as proxies for explainability or interpretability. That is a popular choice that is still an area of active research, since there is still contention as to how much these weights can work as noisy predictors or relative importance of specific regions of the input data. While attention weights are more inherently understandable than the parameters of general deep networks, they become less intelligible in the deeper layers as compared to the outer ones, which brings unique difficulties in using them as justifications for a system or model's decisions (Chaudhari et al., 2021; Zini and Awad, 2022).

Finally, one equally important avenue for continued investigation has to do with the **ethical implications** of explainability and interpretability, since their providers are directly part of the resolution of complex problems with a large gamut of impacts. As explanations and interpretations become ever more critical for oversight purposes, opportunities arise both for the creation of metrics and standards, on one hand, and of ethics-washing and responsibility avoidance, on the other. In order to prevent the latter, policies like accountability measures should be clearly implemented, which can only happen within clear regulatory baselines and mechanisms designed for legal recourse and public scrutiny, which research can help come about. While the current focus seems to be on providing guidance to the industry and promoting innovation, a more balanced approach would leverage commercial interests with human rights, to the benefit of a broader community of stakeholders (Simon and Barr, 2023; Nannini et al., 2023).

6
Conclusion

This work was dedicated to answering the question of how 'explainability' is defined in the domain of foundation models, through a systematic literature review. We arrived at this question after an initial exploratory contact with the topic, which was relayed in Chapter 2, and explained how we would tackle the question in Chapter 3. An overview of our findings was given in Chapter 4, and they were further detailed in Chapter 5. The question had to be refined over the course of the work due to our results: interpretability joined center stage as another equally important term for consideration, while the scope had to be broadened from foundation models to machine learning more generally. That happened because not only was the literature considering foundation models a lot more scarce, but also because comparisons with other types of artificial intelligence systems or models ended up being crucial for understanding the definitions of the terms in question.

Our endeavor has specific limitations that future work can mitigate, possibly using our results as a basis. We did not dive into the workings of specific explainability or interpretability techniques, although we mentioned all the ones that were found during the review. Another fruitful avenue of further research could involve choosing one particular application among the ones mentioned as part of our results to see how the definitions that we concerned ourselves with can vary according to distinct use contexts. We also did not engage in forward snowballing, which could be done by looking at the texts that cited the work found in the current review: that could show how the definitions have changed or persisted over time from their inception. Finally, it would be possible to simply repeat our same queries after some time has elapsed: the mere passage of time can help in answering the original research question, as more texts about explainability and interpretability centered around foundation models get published.

In conclusion, and following what part of the texts under review do, we found out it was useful to keep a distinction between the definitions of the terms explainability and interpretability. **Explainability** can thus be defined as the capacity of an artificial intelligence system to make their mechanisms and inner workings understandable, focusing on matters like their

algorithmic operation, and are usually the province of the developers and designers in charge of creating and maintaining them. **Interpretability**, in a complementary manner, can be defined as the capability of an artificial intelligence system to make their inputs, outputs and the relationships between them, understandable, and what are the possible repercussions involved in them in practice, often being of more direct interest to stakeholders like policymakers and people directly affected by decisions based on these outputs.

In our review, we explored a vast range of possible solutions that can help systems achieve greater explainability and interpretability, and we contributed with an expanded vocabulary that can help classifying distinct techniques and approaches to do so. These methods can inform the creation of specific computational artifacts, or the improvement of currently existing ones, but these, as well as the insights they generate, are complex sociotechnical constructs themselves. As such, they are just as subject to similar technical limitations, human preferences and cultural inclinations as any other applications of our collective ingenuity. We hope that the themes that received consideration in this work can help build, evaluate, and improve artificial intelligence systems that are not only technically more sound but also socially responsible, making it so that the trust that we deposit in them can be rightfully earned.

Bibliography

Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Alarcon, G. and Willis, S. (2023). Explaining Explainable Artificial Intelligence: An integrative model of objective and subjective influences on XAI. In *Proceedings of the 56th Hawaii International Conference on System Sciences*.

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Berthouze, N. (2020). Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). Towards Robust Interpretability with Self-Explaining Neural Networks.

Arpaci-Dusseau, R. H. and Arpaci-Dusseau, A. C. (2018). *Operating Systems: Three Easy Pieces*. CreateSpace Independent Publishing Platform, Erscheinungsort nicht ermittelbar.

Balayn, A., Rikalo, N., Lofi, C., Yang, J., and Bozzon, A. (2022). How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models? In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–16, New York, NY, USA. Association for Computing Machinery.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Bhattacharya, A. (2022a). Foundational Concepts of Explainability Techniques. In *Applied Machine Learning Explainability Techniques: Make ML Models Explainable and Trustworthy for Practical Applications Using LIME, SHAP, and More*. Packt Publishing.

Bhattacharya, A. (2022b). Model Explainability Methods. In *Applied Machine Learning Explainability Techniques: Make ML Models Explainable and Trustworthy for Practical Applications Using LIME, SHAP, and More*. Packt Publishing.

Biolchini, J., Mian, P. G., Natali, A. C. C., and Travassos, G. H. (2005). Systematic review in software engineering.

Biran, O. and Cotton, C. V. (2017). Explanation and Justification in Machine Learning: A Survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*.

Bruce, P. C. and Fleming, G. (2021a). Background: Modeling and the Black-Box Algorithm. In *Responsible Data Science*, pages 27–48. Wiley.

Bruce, P. C. and Fleming, G. (2021b). Model Interpretability: The What and the Why. In *Auditing for Neural Networks*. Wiley.

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., and Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9:101895.

Chaudhari, S., Mithal, V., Polatkan, G., and Ramanath, R. (2021). An Attentive Survey of Attention Models. *ACM Transactions on Intelligent Systems and Technology*, 12(5):1–32.

Chazette, L. (2023). *Requirements Engineering for Explainable Systems*. Doctarate, Gottfried Wilhelm Leibniz Universität, Hannover.

Chazette, L., Brunotte, W., and Speith, T. (2022). Explainable software systems: From requirements analysis to system evaluation. *Requirements Engineering*, 27(4):457–487.

Chazette, L. and Schneider, K. (2020). Explainability as a non-functional requirement: Challenges and recommendations. *Requirements Engineering*, 25(4):493–514.

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4):2051–2068.

Confalonieri, R., Coba, L., Wagner, B., and Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining and Knowledge Discovery*, 11(1):e1391.

Correia, A. and Lindley, S. (2022). Collaboration in relation to Human-AI Systems: Status, Trends, and Impact. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3417–3422.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A Survey of the State of Explainable AI for Natural Language Processing.

DARPA (2016). Explainable Artificial Intelligence (XAI). Broad Agency Announcement DARPA-BAA-16-53, Defense Advanced Research Projects Agency.

Das, A. and Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey.

Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., and Li, Y. (2021). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, pages 1591–1602, New York, NY, USA. Association for Computing Machinery.

Doshi-Velez, F. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., and Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, 55(9):194:1–194:33.

Ehsan, U., Passi, S., Liao, Q., Chan, L., Lee, I.-H., Muller, M. J., and Riedl, M. O. (2021). The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *ArXiv*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):93:1–93:42.

Gunning, D. and Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58.

Gunning, D., Vorm, E., Wang, J. Y., and Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61.

Habiba, U.-E., Bogner, J., and Wagner, S. (2022). Can Requirements Engineering Support Explainable Artificial Intelligence? Towards a User-Centric Approach for Explainability Requirements. In *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, pages 162–165.

Haque, A. B., Islam, A. K. M. N., and Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective- A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, 186:122120.

Ibrahim, R. and Shafiq, M. O. (2023). Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. *ACM Computing Surveys*, 55(10):206:1–206:37.

Islam, M. R., Ahmed, M. U., Barua, S., and Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences*, 12(3):1353.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation.

Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. (2022). A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys*, 55(5):95:1–95:29.

Kaur, D., Uslu, S., Rittichier, K. J., and Durresi, A. (2022). Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, 55(2):39:1–39:38.

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. Technical Report 0400011T.1, Keele University and National ICT Australia Ltd., Keele and Sydney.

Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report / Keele University.

Kofod-Petersen, A. (2015). How to do a Structured Literature Review in computer science.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. (2022). The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective.

Laato, S., Tiainen, M., Najmul Islam, A., and Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research*, 32(7):1–31.

Lakkaraju, H. and Bastani, O. (2019). "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., and Zhou, B. (2023). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55(9):177:1–177:46.

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., Bian, J., and Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., and Tang, J. (2023). Trustworthy AI: A Computational Perspective. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–59.

Liu, Y. and Zaharia, D. M. (2022). Fundamentals of Deep Learning Explainability. In *Practical Deep Learning at Scale with MLflow: Bridge the Gap between Offline Experimentation and Online Production*. Packt Publishing.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., and Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 226:107161.

Meador, D. and Goldsmith, K. (2022). Explainable AI - Using LIME and SHAP. In *Building Data Science Solutions with Anaconda: A Comprehensive Starter Guide to Building Robust and Complete Models*. Packt Publishing.

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

Mohammadkhani, A. H. (2023). Explainable AI for Software Engineering: A Systematic Review and an Empirical Study. Master's thesis, University of Calgary, Calgary.

Mohseni, S., Block, J. E., and Ragan, E. (2021a). Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. In *26th International Conference on Intelligent User Interfaces*, IUI '21, pages 22–31, New York, NY, USA. Association for Computing Machinery.

Mohseni, S., Zarei, N., and Ragan, E. D. (2021b). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4):24:1–24:45.

Molnar, C. (2019). *Interpretable Machine Learning*. Lulu.com, Victoria, British Columbia.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI.

Nannini, L., Balayn, A., and Smith, A. L. (2023). Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1198–1212, New York, NY, USA. Association for Computing Machinery.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*.

Office, I. C. and Institute, T. A. T. (2023). Explaining decisions made with AI. Guidance 1, Information Commissioner's Office & The Alan Turing Institute.

Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., Tal, A. S., Hartman, A., and Kuflik, T. (2022). Mitigating Bias in Algorithmic Systems—A Fish-eye View. *ACM Computing Surveys*, 55(5):87:1–87:37.

Petticrew, M. and Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide*. Wiley, 1 edition.

Räuker, T., Ho, A., Casper, S., and Hadfield-Menell, D. (2023). Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club*, 123(3):A12.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.

Sado, F., Loo, C. K., Liew, W. S., Kerzel, M., and Wermter, S. (2023). Explainable Goal-driven Agents and Robots - A Comprehensive Review. *ACM Computing Surveys*, 55(10):211:1–211:41.

Saeed, W. and Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273.

Selbst, A. D. and Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Serrano, S. and Smith, N. A. (2019). Is Attention Interpretable?

Simon, C. and Barr, J. (2023). Understanding Explainable AI. In *Deep Learning and XAI Techniques for Anomaly Detection: Integrate the Theory and Practice of Deep Anomaly Explainability*. Packt Publishing.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods.

Sokol, K. and Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 56–67, New York, NY, USA. Association for Computing Machinery.

Tomaino, G., Abdulhalim, H., Kireyev, P., and Wertenbroch, K. (2022). Denied by an (Unexplainable) Algorithm: Teleological Explanations for Algorithmic Decisions Enhance Customer Satisfaction.

USA (2022). Blueprint for an AI Bill of Rights | OSTP.

Vellido, A., Martín-Guerrero, J., and Lisboa, P. (2012). Making machine learning models interpretable. In *The European Symposium on Artificial Neural Networks*.

Vilone, G. and Longo, L. (2021). A Quantitative Evaluation of Global, Rule-Based Explanations of Post-Hoc, Model Agnostic Methods. *Frontiers in Artificial Intelligence*, 4.

Votto, A. M. and Liu, C. Z. (2023). Transparent Artificial Intelligence and Human Resource Management: A Systematic Literature Review. In *Proceedings of the 56th Hawaii International Conference on System Sciences*.

Vouros, G. A. (2022). Explainable Deep Reinforcement Learning: State of the Art and Challenges. *ACM Computing Surveys*, 55(5):92:1–92:39.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057. PMLR.

Yeh, C.-K., Kim, B., Arik, S. O., Li, C.-L., Pfister, T., and Ravikumar, P. (2022). On Completeness-aware Concept-Based Explanations in Deep Neural Networks.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929.

Zhu, J., Liapis, A., Risi, S., Bidarra, R., and Youngblood, G. M. (2018). Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8, Maastricht, Netherlands. IEEE Press.

Zini, J. E. and Awad, M. (2022). On the Explainability of Natural Language Processing Deep Models. *ACM Computing Surveys*, 55(5):103:1–103:31.

A
Constructing the SLR Protocol

This appendix supplies the Systematic Literature Reviews Protocol in full (in Table A.1) and describes in detail how it was followed to completion. Different choices along the way would have led to consequently distinct results, and readers are welcome to experiment with them.

Table A.1: Protocol for SLR

Topic	**Response**
Introductory	
Title	SLR about AI Explainability
Introduction and background	See Chapter 2 for the background.
Objectives	See Chapter 1 for the objectives.
Methodology	

Continued on next page

Table A.1: Protocol for SLR (Continued)

PICOC Criteria: Keywords and Synonyms	Population: Foundation Models (Foundation Models, Large Language Models, LLM, Transformers, Neural Networks, Deep Learning, Artificial Intelligence) Intervention: Explainability (Explainability, Explicability, Interpretability, Explainable Artificial Intelligence, XAI) Comparison: Black boxes (Black Box, Inscrutable, Opaque, Bias, Biased) Outcome: Explanations (Explanation, Explication, Trust, Trustworthiness, Fairness, Transparency, Reliability, Responsibility) Context: Definition (Definition, Taxonomy) | Emergent Capabilities (Emergent Capabilities, Emergent Abilities, Emergent Properties)~~—| Obligations (Obligation, User Experience, UX, Regulation, Legislation) | Performance | Non Functional Requirement~~
Research questions	RQ1. *How is 'explainability' defined in the domain of foundation models?* RQ2. *What are the current efforts to explain emergent capabilites in AI?* RQ3. ~~*To whom are AI explanations owed, and what does that figuration entail?*~~ RQ4. ~~*How does trying to achieve explainability affect a model's performance?*~~ RQ5. ~~*What does considering explainability a (non-functional) requirement involve?*~~
Digital library sources	• ACM Digital Library • IEEE Digital Library • ~~Scopus~~ • ~~Web of Science~~

Continued on next page

Table A.1: Protocol for SLR (Continued)

Inclusion and exclusion criteria	Period: No restriction. Language: Excluding texts not in English. Type of literature: Excluding gray literature (reports, policy, working papers, newsletters, government). | No exclusion. Type of source: Including articles from journals. | Including articles from journals and conferences. Impact source: No exclusion. | Including articles from Q1 sources. Access: Excluding sources to which we do not have access. Relevant to research questions: Excluding texts not relevant to at least 1 of our first 2 research questions.

A.1 Building the search string

Each of the PICOC criteria and their synonyms gave rise to one or more sets of keywords which were used as experimental building blocks for the searches. The results derived from their combinations led the research process, as further detailed in the next section. In the case of the Population criterion, the different sets indicated how broadly or narrowly we would be able to define the technologies under consideration, as far as explainability was concerned. In the case of the Context criterion, the sets had to do with each particular Research Question we were interested in addressing. The possible mix and match building blocks for each criterion (in Extended Backus-Naur Form notation, where the square brackets denote optional parts and the vertical bars are syntax for alternatives) were then as follows:

Population

```
"Foundation Model" OR "Foundation Models"[ OR "Large Language
Model" OR "Large Language Models" OR "LLM" OR "LLMs" OR "LLM's"
OR "Transformers"[ OR "Neural Network" OR "Neural Networks" OR
"Deep Learning"[ OR "Artificial Intelligence"]]]
```

Intervention

```
"Explainability" OR "Explicability" OR "Interpretability" OR
"Explainable Artificial Intelligence" OR "XAI"
```

Comparison

```
"Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque" OR
"Bias" OR "Biased"
```

Outcome

```
"Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair" OR
"Fairness" OR "Transparent" OR "Transparency" OR "Reliable" OR
"Reliability" OR "Responsible"
```

Context

```
("Define" OR "Definition" OR "Definitions" OR "Taxonomy" OR
"Taxonomies")
|
("Emergent Capability" OR "Emergent Capabilities" OR "Emergent
Ability" OR "Emergent Abilities" OR "Emergent Property" OR
"Emergent Properties")
|
("Regulation" OR "Regulations" OR "Regulatory" OR "Legislation"
OR "Legislative" OR "Law" OR "Laws" OR "User Experience" OR
"UX" OR "Audience")
|
("Performance")
|
("Non Functional Requirement")
```

A.2
Comments on the exploratory mid-protocol search processes

We started a first batch of searches while still constructing the research protocol to see how exactly each repository would behave and what parameters we could tweak. Some terms were discovered and added during this stage, including nouns but mainly the adjective forms of the terms already chosen, after we noticed that the results which used wildcards did not seem consistent with our expectations (there was much more variation than imagined at first).

Initial searches in the ACM Library and IEEE Xplore seemed successful, and their results were organized as Zotero subcollections and briefly compared to further refine the search process, along with the respective search strings and search options used on each platform. Scopus and Web of Science searches were not successful, and neither was the use of Harzig's Publish or Perish software. Details on each one follow.

Scopus does not have an option that includes searching the full texts, so we opted for the most comprehensively relevant search in this case, which was "Title+Abstract+Keywords". Unfortunately, the platform limits the search of keywords for known English words, so it did not recognize the terms "Explainability", "Explicability" or "Interpretability", for instance, as valid keywords. After removing them, the search string was accepted, but there were no results. We removed all terms related to the Intervention criterion, but there were still no results.

We proceeded inputting only one set of keywords at a time, starting with the set for the Population criterion, and then added the set for the Context criterion, and came up with 680 documents, just to make sure the search was taking into account multiple sets of keywords using the syntax provided. After that we added a modified set for the Intervention criterion, consisting of only the keyword "Explainable", and the results came down to only 3 documents, none of which dealt with our Research Questions. Harzig's Publish or Perish software was then considered as an alternative, but we found out that it does not support searching in the Abstracts like Scopus' site does. That ended up restricting the results even more than the the website search. Because of all of these factors, Scopus was eliminated as a possible database for our review.

Since we did not have access to Web of Science (it is a paid platform and none of the institutions to which I belong currently subscribe to it), it was also discarded as a possible source for the review.

A.3 Most promising candidate queries

These were all given a descriptive alias to help discuss them during the search process and to indicate what are the most significant differences in each one. The results were aggregated in Zotero subcollections and titles, keywords and abstracts skimmed over whenever applicable to give a feel of how to refine further search options and strings. The final searches were run on July 7, 2023.

A.3.1
Most promising candidate queries for RQ1 and RQ2 at ACM

1. Alias: **Reduced Population terms to just FM's, RQ1 Context**
   Database: ACM Digital Library
   Collection: The ACM Guide to Computing Literature
   'Search Within' option used: "Anywhere" for all fields
   Publication Date: All dates
   Filters: None
   Results: **34**

   Search string:

```
("Foundation Model" OR "Foundation Models")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Define" OR "Definition" OR "Definitions" OR "Taxonomy"
OR "Taxonomies")
```

2. Alias: **Reduced Intervention scope for just Keywords, RQ1 Context**
   Database: ACM Digital Library

Collection: The ACM Guide to Computing Literature
'Search Within' option used: "Author Keyword" for Intervention criteria, "Anywhere" for all other fields
Publication Date: All dates
Filters: None
Results: **53**

Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Define" OR "Definition" OR "Definitions" OR "Taxonomy"
OR "Taxonomies")
```

3. Alias: **Reduced Intervention scope for just Abstracts, RQ1 Context**
   Database: ACM Digital Library
   Collection: The ACM Guide to Computing Literature

'Search Within' option used: "Abstract" for Intervention criteria, "Anywhere" for all other fields
Publication Date: All dates
Filters: None
Results: **97**

Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Define" OR "Definition" OR "Definitions" OR "Taxonomy"
OR "Taxonomies")
```

4. Alias: **RQ2 Context, no further restrictions**
   Database: ACM Digital Library
   Collection: The ACM Guide to Computing Literature
   'Search Within' option used: "Anywhere" for all fields
   Publication Date: All dates

Filters: None
Results: **12**

Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Emergent Capability" OR "Emergent Capabilities" OR
"Emergent Ability" OR "Emergent Abilities" OR "Emergent
Property" OR "Emergent Properties")
```

A.3.2
Most promising candidate queries for RQ1 and RQ2 at IEEE

1. Alias: **Reduced Population terms to just FM's, RQ1 Context**
   Database: IEEE Xplore
   Search 'in' option used: "Full Text & Metadata" for all fields
   Publication Date: All dates

Filters: None
Results: **24**

Search string:

```
("Foundation Model" OR "Foundation Models")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Define" OR "Definition" OR "Definitions" OR "Taxonomy"
OR "Taxonomies")
```

2. Alias: **Reduced Intervention scope for just Keywords, RQ1 Context**
   Database: IEEE Xplore
   Search 'in' option used: "Author Keywords" for Intervention criteria, "Full Text & Metadata" for all other fields
   Publication Date: All dates
   Filters: None
   Results: **47**

   Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Define" OR "Definition" OR "Definitions" OR "Taxonomy"
OR "Taxonomies")
```

3. Alias: **Reduced Intervention scope for just Keywords, RQ1 Context**
   Database: IEEE Xplore
   Search 'in' option used: "Abstract" for Intervention criteria, "Full Text & Metadata" for all other fields
   Publication Date: All dates
   Filters: None
   Results: **118**

   Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Define" OR "Definition" OR "Definitions" OR "Taxonomy"
OR "Taxonomies")
```

4. Alias: **RQ2 Context, no further restrictions**
   Database: IEEE Xplore
   Search 'in' option used: "Full Text & Metadata" for all fields
   Publication Date: All dates
   Filters: None
   Results: **3**

   Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Emergent Capability" OR "Emergent Capabilities" OR
"Emergent Ability" OR "Emergent Abilities" OR "Emergent
Property" OR "Emergent Properties")
```

A.4 Final search strings and options

Based on the number of results found in each case and on an initial perusal of the titles of the search results, these were the options deemed most promising to be subsequently subjected to the inclusion and exclusion criteria and to the qualitative assessment, if needed, before the information extraction. These options are recorded here in full since they were used to effectively and efficiently run the searches on each platform.

A.4.1
Final search strings and options for RQ1

- Database: ACM Digital Library
  Collection: The ACM Guide to Computing Literature
  'Search Within' option used: "Abstract" for Intervention criteria, "Anywhere" for all other fields
  Publication Date: All dates
  Filters: None
  Results: **97**

  Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Define" OR "Definition" OR "Definitions" OR "Taxonomy"
OR "Taxonomies")
```

- Database: IEEE Xplore
  Search 'in' option used: "Abstract" for Intervention criteria, "Full Text & Metadata" for all other fields
  Publication Date: All dates
  Filters: None
  Results: **118**

  Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Define" OR "Definition" OR "Definitions" OR "Taxonomy"
OR "Taxonomies")
```

A.4.2
Final search strings and options for RQ2

- Database: ACM Digital Library
  Collection: The ACM Guide to Computing Literature
  'Search Within' option used: "Anywhere" for all fields
  Publication Date: All dates
  Filters: None
  Results: **12**

  Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Emergent Capability" OR "Emergent Capabilities" OR
"Emergent Ability" OR "Emergent Abilities" OR "Emergent
Property" OR "Emergent Properties")
```

- Database: IEEE Xplore
  Search 'in' option used: "Full Text & Metadata" for all fields
  Publication Date: All dates
  Filters: None
  Results: **3**

  Search string:

```
("Foundation Model" OR "Foundation Models" OR "Large
Language Model" OR "Large Language Models" OR "LLM" OR
"LLMs" OR "LLM's" OR "Transformers")

AND

("Explainability" OR "Explicability" OR "Interpretability"
OR "Explainable Artificial Intelligence" OR "XAI")

AND

("Black Box" OR "Black Boxes" OR "Inscrutable" OR "Opaque"
OR "Bias" OR "Biased")

AND

("Explanation" OR "Explanations" OR "Explication" OR
"Explications" OR "Trust" OR "Trustworthiness" OR "Fair"
OR "Fairness" OR "Transparent" OR "Transparency" OR
"Reliable" OR "Reliability" OR "Responsible")

AND

("Emergent Capability" OR "Emergent Capabilities" OR
"Emergent Ability" OR "Emergent Abilities" OR "Emergent
Property" OR "Emergent Properties")
```

A.5
Notes on enacting inclusion and exclusion criteria

We started by exporting the final Zotero subcollection (with 213 unique entries) to a .csv file (encoding: UTF-8, without Byte Order Mark) and manually imported it to an inclusion and exclusion criteria spreadsheet (publicly available as a Google Sheet).

We then proceeded to downloading all of the full text resources when available using institutional access options (starting with PUC's and then IBM's), or using other commonly employed methods known to academics when there was no other alternative (including, but not limited to, buying books).

Whenever it was possible to dive into more specific applications, that was the label chosen. For example: 'Autonomous transportation' was chosen over 'Image classification', even if the former utilizes the latter.

We chose to differentiate 'Information extraction', as meaning from among a large text, find passages that are of a certain type (for instance, hate speech) and 'Text classification', which we took to be the activity of, while analyzing a data entry that includes texts, predict a certain classifier for that data entry.

The expression 'Autonomous transportation' was preferred to 'Autonomous driving', since the former included other modes of transportation relevant for some of the texts under review, such as by train or by boat, while the latter is mostly concerned with road vehicles like cars and trucks.

When an item could be classified in two or more different types of Applications, the most prevalent one was chosen, when evident. For instance, if a task that aims at 'Property prediction' used 'Information extraction' as a sub-step, the former was chosen. If there were no applications that stood out, the item was classified as dealing with 'Various' applications.

We took 'Information extraction' to also include the task of information retrieval: given an information as input, retrieve documents from a repository that are related to that information. In turn, 'Text generation' also includes the task of text translation.

A.6
Information extraction

This step constituted of reading through the selected texts integrally, highlighting and annotating the most important parts, and creating an information extraction form (publicly available as a Google Sheet).

The categories initially used in the information extraction form were as follows, with change occurring along the writing process that gave rise to the

ones effectively used:

- Title
- Year
- Author(s)
- Item Type
- Keywords
- Initial context and motivation
- Explainability characterizations
- Distinction between explainability and interpretability
- Explainability definition
- Interpretability definition
- Additional terms (understandability, intelligibility, comprehensibility, transparency)
- Benefits of explicability/interpretability
- Risks of non-explicability
- Techniques for explaining
- Application domains and specific applications mentioned
- Explainability assessment and metrics
- Important references (for backwards snowballing)
- Cited by (for forward snowballing)
- Main contribution summary
- Interdisciplinarity
- Future challenges and explainability drawbacks

B
Backward snowballing

This is the full list of interesting references on topic that were found during the systematic literature review and that were not otherwise addressed in the current work. Their reading can be a source of enrichment for the arguments advanced here.

- A Historical Perspective of Explainable Artificial Intelligence (Confalonieri et al., 2021)
- A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems (Mohseni et al., 2021b)
- A Quantitative Evaluation of Global, Rule-Based Explanations of Post-Hoc, Model Agnostic Methods (Vilone and Longo, 2021)
- A Survey of the State of Explainable AI for Natural Language Processing (Danilevsky et al., 2020)
- Attention is not Explanation (Jain and Wallace, 2019)
- Blueprint for an AI Bill of Rights | OSTP (USA, 2022)
- DARPA's explainable AI (XAI) program: A retrospective (Gunning et al., 2021)
- DARPA's Explainable Artificial Intelligence (XAI) Program (Gunning and Aha, 2019)
- Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study (Alqaraawi et al., 2020)
- Explainability fact sheets: a framework for systematic assessment of explainable approaches (Sokol and Flach, 2020)
- Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation (Zhu et al., 2018)
- Explainable Artificial Intelligence (XAI) (DARPA, 2016)
- Explaining decisions made with AI (Office and Institute, 2023)
- Explaining Explanations: An Overview of Interpretability of Machine Learning (Gilpin et al., 2018)

- Explanation and Justification in Machine Learning: a Survey (Biran and Cotton, 2017)
- Explanation in artificial intelligence: Insights from the social sciences (Miller, 2019)
- Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI (Mueller et al., 2019)
- Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods (Slack et al., 2020)
- How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models? (Balayn et al., 2022)
- "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations (Lakkaraju and Bastani, 2019)
- Interpretable Machine Learning (Molnar, 2019)
- Is Attention Interpretable? (Serrano and Smith, 2019)
- Making machine learning models interpretable (Vellido et al., 2012)
- Meaningful information and the right to explanation (Selbst and Powles, 2017)
- Methods for interpreting and understanding deep neural networks (Montavon et al., 2018)
- On Completeness-aware Concept-Based Explanations in Deep Neural Networks (Yeh et al., 2022)
- Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey (Das and Rad, 2020)
- Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) (Adadi and Berrada, 2018)
- Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark (Mohseni et al., 2021a)
- Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead (Rudin, 2019)
- The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective (Krishna et al., 2022)
- The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery (Lipton, 2018)
- The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations (Ehsan et al., 2021)

- Towards A Rigorous Science of Interpretable Machine Learning (Doshi-Velez and Kim, 2017)
- Towards Robust Interpretability with Self-Explaining Neural Networks (Alvarez-Melis and Jaakkola, 2018)
- Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle (Dhanorkar et al., 2021)
- "Why Should I Trust You?": Explaining the Predictions of Any Classifier (Ribeiro et al., 2016)

C
Learning Computer Science in Portuguese

Learning Computer Science in Brazil, as Portuguese speakers, entails delightful challenges all on their own, since there are so many terms that students and professors alike find hard to translate from English, from where they usually originate. When taking into account the fact that there is even outright dissent among university faculties on whether students should be even allowed to write programs using Portuguese, it becomes clear this state of things has probably more implications than people want to let on.

Over the years it has become a fun obsession for the author of this work to discover (and even propose) Portuguese translations for those terms whose meaning seem the hardest to convey from other languages, especially English, and even more so those that appear to be particularly stubborn in their supposed 'untranslatability'. Needless to say, this is as much a geopolitical stance as it is one of facilitating the accrual and transmission of knowledge; of, at the end of the day and most of all, fostering understanding. And if it were the case of taking a stand on the matter of preferring the usage of terms in English or in Portuguese, evidently our position would be that one should learn *both* and use whichever communicates best at each occasion.

To that end, readers will find below the biggest culprits and proposed translations for the most used terms during Computer Science teaching (and probably also in the industry, from what we can tell so far). This list does not aim to be authoritative, and purports to be a living construct, to which contributions are more than welcome.

Table C.1: Translations to Portuguese

Original term	**Portuguese Translation**
Angle brackets (< >)	Parênteses angulares | Colchetes angulares
Accuracy	Exatidão (para evitar Precisão, de Precision)
Affordances	Habilitadores | Capacitadores
API (Application Programming Interface)	Interface para programação (da aplicação)

Continued on next page

Table C.1: Translations to Portuguese (Continued)

Array	Arranjo | Vetor | Coleção homogênea
Assert	Assegurar | Garantir | Asseverar
Assertion	Caução | Garantia (é comum ouvir Asserção, aqui preferida para se traduzir Statement)
Assume	Admitir | Assumir | Considerar
Backend	Traseira | Fundos | Posterior | Retaguarda | Ré
Batch	Lote
Benchmark	Nivelamento | Comparação de desempenho
Broker	Atravessador | Intermediário | Corretor
Buffer	Armazenamento provisório | Reserva
Cluster	Agrupamento | Agregação
Commit (verb and noun)	Empenhar (to commit) | Empenho (a commit)
Deadlock	Impasse
Deployment	Lançamento (como de um foguete)
Deprecated	Descontinuado | Obsoleto | Ultrapassado
Dispatcher	Emissor | Despachante
Driver	Condutor
Endpoint	(Dispositivo) Terminal
Exploit (noun)	Vulnerabilidade
Feature	Característica | Propriedade
Fetch	Obter
Fitness	Aptidão
Framework	Arcabouço
Frontend	Dianteira | Frente | Anterior | Vanguarda | Vante
Fuzzy logic	Lógica gradativa | Lógica nebulosa
Guardrail	Salvaguarda
GUI (Graphical User Interface)	Interface gráfica (do usuário)
Hack	Macete
Hardware | Firmware | Software	Artigos rígidos | firmes | flexíveis

Continued on next page

Table C.1: Translations to Portuguese (Continued)

Hash (noun)	Suma | Picadinho
IDE (Integrated Development Environment)	Ambiente (integrado de desenvolvimento)
Kernel	Cerne
Linter	Papa-fiapos | Filtro de cotão
Look-ahead (verb, noun)	Antever, o antevisto
Outlier	Anômalo | Discrepante
Overfit	Superajustar
Overhead	Custo ou carga adicional | Ônus de operação
Owner (as in Data owner or Product owner)	Titular | Proprietário
Padding	Enchimento
Parser	Analisador sintático
Pipeline	Linha condutora | Duto
Production, In	Na produção (evita o ambíguo 'em produção') | Em funcionamento
Prompt (as in Command)	Indicador | Aceno | Sinal
Prompt (in Large Language Models)	Prenúncio | Comando
Raise (as in 'Raise an exception')	Sinalizar
Scaling	Dimensionamento
Scheduler	Planejador | Escalonador | Agendador
Singleton	Filho único | Conjunto unitário
Starvation	Inanição | Espera indefinida
Statement	Afirmação | Asserção
Stemming	Truncagem
String	Cadeia | Corrente
Support	Apoio | Aceitação | Atendimento (prestar)
Tail	Fecho | Cauda
Thread	Tarefa | Trecho (de código)

Continued on next page

Table C.1: Translations to Portuguese (Continued)

Toy (as in 'Toy example' or 'Toy model')	Lúdico
Trade-off	Oposição | Solução de compromisso | Dilema
Trap	Captura(r) | Interceptar | Interceptação
Transparent	Imperceptível | Invisível*
Turnaround	Conclusão (tempo para)
Type casting	Coerção | Conversão | Moldagem de tipos
Walkthrough	Passeio guiado
Wrangle (data)	Manejar (dados)

* The same ambiguity happens in English, derived from the meaning of Transparency in an organization's matters, which refers to being open to public scrutiny. Coincidentally, this is remarked upon by one phenomenal book on Operating Systems (Arpaci-Dusseau and Arpaci-Dusseau, 2018, p. 133 note 2).

D
Paralipomena

I do not say that artists cannot be seers, inspired: that the awen *cannot come upon them, and the god speak through them. Who would be an artist if they did not believe that that happens? If they did not know it happens, because they have felt the god within them use their tongue, their hands? Maybe only once, once in their lives. But once is enough.*

Nor would I say that the artist alone is so burdened and so privileged. The scientist is another who prepares, who makes ready, working day and night, sleeping and awake, for inspiration. As Pythagoras knew, the god may speak in the forms of geometry as well as in the shapes of dreams; in the harmony of pure thought as well as in the harmony of sounds; in numbers as well as in words.

But it is words that make the trouble and confusion. We are asked now to consider words as useful in only one way: as signs. Our philosophers, some of them, would have us agree that a word (sentence, statement) has value only in so far as it has one single meaning, points to one fact that is comprehensible to the rational intellect, logically sound, and—ideally—quantifiable.

[...]

The artist deals with what cannot be said in words.

The artist whose medium is fiction does this in words. *The novelist says in words what cannot be said in words.*

Ursula K. Le Guin,
The Left Hand of Darkness (1969)