POLITECNICO DI MILANO
DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE, E BIOINGEGNERIA
DOCTORAL PROGRAMME IN COMPUTER SCIENCE

MODEL EXPLAINABILITY THROUGH HUMAN KNOWLEDGE AND CROWDSOURCING

Doctoral Dissertation of:
Andrea Tocchetti

Supervisor:
Prof. Marco Brambilla

Tutor:
Prof. Davide Martinenghi

The Chair of the Doctoral Program:
Prof. Luigi Piroddi

2024 – XXXV Cycle

"Both knowledge and awareness are equivocal.
One's reality might be another's illusion."

To the friends I made at Politecnico di Milano, thank you for your kindness and the dinners together.

To the friends and colleagues at TU Delft, thank you for making me feel at home.

To the friends I shared passions, games, laughs, and adventures with, I will always treasure the time spent together and look forward to more.

To my dear colleague and friend, Lorenzo, for the time spent together through the hardships and the adventures we shared.

To my supervisor and friend, Marco Brambilla, for the support, guidance, and trust, for the laughs, the travels, and the meals together. You will always be a person I will look up to.

To Chiara, meeting you was an unexpected quest I undertook during this journey, and saving a princess from a tower has become a lifelong adventure.

Finally, to my beloved family, for always driving me to pursue my dreams, for their support, and love.

To all of you.

Thank You.

Abstract

The ongoing development of incrementally complex and high-performing ML and AI models has made them more opaque than ever, making it hard to understand their behaviour and decision-making process. Such complexity calls for a fundamental property that models must have, *i.e.*, they must be explainable, allowing human interpreters to understand their decision-making process. Consequently, the research community has focused on developing approaches to provide explanations for such black-box models faithfully. While several efficient methods have been developed, human interpretability still represents the most critical aspect. Humans might play various roles in this context, and their knowledge is fundamental to achieving such an objective. This PhD dissertation focuses on human-centred approaches in the context of Explainable AI, analyzing and developing techniques to involve humans, collect and structure their knowledge, and employ it towards explaining model behaviour. Besides XAI, crowdsourcing and gamification are essential to driving human involvement and behaviour. In this dissertation, fundamental literature on such topics of interest is provided. Then, the role of humans in XAI and their contribution towards model robustness are described. Explainability approaches are then developed in the context of Natural Language Processing (NLP) and Computer Vision (CV). In NLP, a formalization to organize human knowledge in various tasks is defined, and data is collected through crowdsourcing. In CV, an approach to describe the decision-making process of black-box models using human knowledge was developed and tested against state-of-the-art methods, reporting on its effectiveness. This dissemination focuses on humans as the core element to achieve high interpretability, driving model trustworthiness. Several perspectives and human-driven approaches demonstrate the fundamental need to engage humans in XAI, highlighting the relevance of such a research topic.

Sommario

Lo sviluppo di modelli di Machine Learning (ML) e Intelligenza Artificiale (AI) sempre più complessi e performanti ha reso difficile la comprensione del loro comportamento. Tale incremento di complessità evidenzia una proprietà fondamentale che i modelli di ML devono possedere, ovvero essere spiegabili, consentendo agli interpreti umani di capire come prendono le decisioni. Di conseguenza, la comunità scientifica si è concentrata sullo sviluppo di approcci per fornire spiegazioni accurate per tali modelli "black-box". Sebbene siano stati sviluppati numerosi metodi, l'interpretabilità umana rappresenta ancora l'aspetto più critico. In questo contesto, gli esseri umani possono ricoprire diversi ruoli oin tale contest, e la loro conoscenza è fondamentale per raggiungere questo obiettivo. Questa tesi di dottorato si concentra sugli approcci incentrati sull'uomo nel contesto dell'Intelligenza Artificiale Spiegabile (XAI), analizzando e sviluppando tecniche per coinvolgere gli esseri umani, raccogliere e strutturare la loro conoscenza e utilizzarla per spiegare il comportamento dei modelli. Oltre alla XAI, il crowdsourcing e la gamification sono stati utilizzati per incentivare il coinvolgimento umano e guidare il loro comportamento. Una panoramica degli argomenti di interesse menzionati nella letteratura è presentata. Successivamente, il ruolo degli esseri umani nella XAI e il loro contributo alla robustezza dei modelli è descritto. Gli approcci di spiegabilità proposti in questa tesi sono sviluppati nel contesto del Natural Language Processing (NLP) e della Computer Vision (CV). Nel NLP, si propone una formalizzazione per organizzare la conoscenza umana per svariate attività, raccogliendo i dati tramite crowdsourcing. Nella CV, è stato sviluppato e testato un approccio per descrivere il processo decisionale dei modelli "black-box" utilizzando la conoscenza umana, confrontandolo con i metodi più avanzati e riportandone efficacia. Questa ricerca pone gli esseri umani come elemento centrale per raggiungere un'elevata interpretabilità, aumentando così l'affidabilità dei modelli. Diverse prospettive e approcci guidati dall'uomo dimostrano la necessità fondamentale di coinvolgere gli esseri umani nella XAI, evidenziando l'importanza di tale argomento di ricerca.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Over the last decades, the widespread use of Machine Learning (ML) models demonstrated its effectiveness in supporting and improving human capabilities in various contexts like computer science, economics, medicine, and many more while driving technological advancement like never before. The demonstrated effectiveness of such models on general and domain-specific tasks has driven the development of specialized models capable of achieving even higher performance. For example, the recent development of Deep Learning (DL) and Deep Neural Networks (DNN) outperformed state-of-the-art models on tasks such as image classification, text translation, and many more. Furthermore, recent technological advancements and the rise of high-performance machine learning approaches contributed to developing a novel research field. Artificial Intelligence (AI) aims to design and develop software systems capable of replicating human-like cognitive abilities and behaviours to perform complex tasks efficiently, even those humans are not usually capable of accomplishing. Nowadays, AI systems permeate our daily lives through voice assistants, self-driving cars, and many more, easing our efforts and enhancing our capabilities. Despite the widespread excitement about such accomplishments, the scientific community quickly apprehended that such systems cannot rely on performance alone to become valuable human assets. Indeed, most complex, high-performing AI systems were lacking an essential feature. Due to their intricacy, their behaviour was not promptly understandable to the people using or developing them, leading to a loss of trust and hindering their applicability to real-world scenarios.

Problem Statement. The fundamental need to understand models' behaviour brought forth the necessity of developing methodologies to faithfully represent their rationale in a human-understandable fashion, allowing human interpreters to comprehend their

decision-making process. The research on Explainable AI (XAI) establishes this aim as its primary focus. Furthermore, the differences in how humans and machine learning systems learn, explain, and represent knowledge make it fundamental to bridge the gap between model and human behaviour to improve or enable humans' understanding of such systems. In particular, AI systems must be made explainable regardless of whether the humans employing them are developers or end-users, experts or novices, while still accounting for the user's expertise, consequently shaping rationales to provide the information and representation that better convey the explanation.

A faithful, complete, and understandable representation of a model's behaviour would increase human trust and acceptance. On the other hand, it would also be helpful to inspect such systems, allowing researchers to understand their faults and consequently driving models' performance even higher. Acknowledging the intrinsic two-sided nature of explainability, *i.e.*, humans and models can be considered two sides of the same coin; researchers have recognized the need to involve humans in such a complex process. While XAI experts and researchers are essential in building explainable systems and studying methods to make complex models understandable, end-users are also important as final beneficiaries of models and their explanations and sometimes knowledge providers concerning XAI-related tasks. Although human involvement contributes towards answering various research questions, designing and developing ad-hoc techniques is necessary to make such tasks understandable and feasible to humans. Researchers involve humans through well-known crowdsourcing techniques to collect knowledge, assess the understandability of models and their explanations, and accomplish various explainability tasks. These are usually designed to make them as clear as possible to non-experts, sometimes combining them with gamification and other playful design techniques to make them more enjoyable and ease the process.

Research Objective. This PhD thesis focuses on exploring and researching various aspects of the Explainable AI research field, with a strong focus on humans and aimed at developing methods and approaches to collect and use their knowledge to address and support model explainability. In particular, this PhD thesis reflects the steps undertaken and the topics addressed throughout the research. It touches on several sub-topics in the fields of interest, framing them from various perspectives rather than focusing on a single one. We first explore crowdsourcing-related topics in various contexts to understand which principles and approaches are of fundamental interest in involving human actors properly. Then, we perform explainability-related research to understand how human knowledge can contribute to this field and which roles humans play in XAI. Finally, we develop methods to collect and employ domain-specific knowledge towards building and improving explanations. We strive to demonstrate the impact of human involvement and human knowledge towards building explainable models.

Structure of the Thesis. This PhD thesis is structured as follows. Chapter 2 discusses the backgrounds and related works on Explainable AI, Human-Computer Interaction (HCI), and Crowdsourcing. Furthermore, topic-specific backgrounds will be addressed in each chapter, providing detailed information about the subject of interest. Chapter 3 describes the research on Crowdsourcing aimed at designing and developing methods promoting participation, employing gamification and empathy as the main

drivers. Chapter 4 describes the work carried out to organize and conceptualize the literature on the involvement of humans and robustness in XAI. Chapter 5 frames XAI from the Natural Language Processing (NLP) perspective, addressing human knowledge's structuring, formalization, and collection for a chosen set of tasks of interest. Chapter 6 frames XAI from the standpoint of Image Classification tasks, providing a novel methodology for local explanations for such tasks and a first approach to merge them into category-wise global descriptions. Chapter 7 offers final remarks on the topics addressed in this thesis, identifying future challenges and outlining new research lines. Furthermore, grey boxes providing a brief summary of the most important concepts described in a section are provided.

Published Content. Previously published content will be employed in writing this thesis, enriching it with novel content. Each chapter reports the list of articles involved in its writing, detailing the candidate's contribution. For what concerns background and related works, the content will be re-organized, improved, and distributed in either the generic background (*i.e.*, Chapter 2) or domain- and context-specific one (*i.e.*, the one explained in each chapter). The following publications were used

1. A. Tocchetti, L. Corti, M. Brambilla, and D. Di Marco. "A Web-Based Co-Creation and User Engagement Method and Platform". In: Lecture Notes in Computer Science. Springer International Publishing, 2021, pp. 496-501. DOI: 10.1007/978-3-030-74296-6_38. URL: https://doi.org/10.1007/978-3-030-74296-6_38

2. A. Tocchetti and M. Brambilla. "A gamified crowdsourcing framework for data-driven co-creation of policy-making and social foresight". In: CEUR Workshop Proceedings 2736 (2020), pp. 34-44. URL: http://ceur-ws.org/Vol-2736/paper6.pdf

3. Andrea Tocchetti, Silvia Maria Talenti, and Marco Brambilla. "My Lockdown Escape": A Data Collection Approach based on Gamification and Crowdsourcing for Subjective Perspectives, Self-Empathy, and Memories about Past Experiences". In: Proceedings of the 3rd Empathy-Centric Design Workshop: Scrutinizing Empathy Beyond the Individual. EmpathiCH '24. Honolulu, HI, USA: Association for Computing Machinery, 2024, pp. 14–20. ISBN: 9798400717888. doi: 10.1145/3661790.3661794. url: https://doi.org/10.1145/3661790.3661794.

4. Andrea Tocchetti, Lorenzo Corti, Marco Brambilla, and Irene Celino. "EXP-Crowd: A Gamified Crowdsourcing Framework for Explainability". In: Frontiers in Artificial Intelligence 5 (2022). ISSN: 2624-8212. DOI: 10.3389/frai.2022.826499. URL: https://www.frontiersin.org/article/10.3389/frai.2022.826499.

5. Andrea Tocchetti, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie Yang. 2024. A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities. ACM Comput. Surv. https://doi.org/10.1145/3665926

6. Tocchetti, A.; Brambilla, M. The Role of Human Knowledge in Explainable AI. Data 2022, 7, 93. https://doi.org/10.3390/data7070093

7. Andrea Tocchetti, Jie Yang, and Marco Brambilla. "Rationale Trees: Towards a Formalization of Human Knowledge for Explainable Natural Language Processing". In: Proceedings of the 4th Italian Workshop on Explainable Artificial Intelligence colocated with 22nd International Conference of the Italian Association for Artificial Intelligence(AIxIA 2023), Roma, Italy, November 8, 2023. Vol. 3518. CEUR Workshop Proceedings. CEUR-WS.org, 2023, pp. 29-46. url: https://ceur-ws.org/Vol-3518/paper3.pdf.

8. Matteo Bianchi, Antonio De Santis, Andrea Tocchetti, and Marco Brambilla. Interpretable network visualizations: A human-in-the-loop approach for post-hoc explainability of cnn-based image classification. In Kate Larson, editor, Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pages 3715–3723. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track. DOI: 10.24963/ijcai.2024/411, URL: https://doi.org/10.24963/ijcai.2024/411

While it is not directly discussed throughout this dissertation, the following article is used for additional discussion in the conclusions.

1. Mathyas Giudici, Federica Liguori, Andrea Tocchetti, and Marco Brambilla. Unveiling human-ai interaction and subjective perceptions about artificial intelligent agents. In Kostas Stefanidis, Kari Systä, Maristella Matera, Sebastian Heil, Haridimos Kondylakis, and Elisa Quintarelli, editors, Web Engineering, pages 414–418, Cham, 2024. Springer Nature Switzerland. DOI: doi.org/10.1007/978-3-031-62362-2_36, URL: https://doi.org/10.1007/978-3\-031-62362-2_36.

Moreover, part of Chapter 2 will be soon published in a Book while part of Chapter 5 will be published at the AHFE2025 conference.

1. A. Tocchetti, M. Bianchi, R. Campi, A. De Santis, and M. Brambilla. "On the Principles and Effectiveness of Gamification in Bidirectional Artificial Intelligence and Explainable AI". In: Bi-directionality in Human-AI Collaborative Systems. Elsevier. ISBN: 9780443405532.

2. A. Tocchetti, V. Naldi, and M. Brambilla. "Web-based Human-centered Explainability of NLP Tasks with Rationale Mapping Theory". In: Proceedings of the 16th International Conference on Applied Human Factors and Ergonomics. July 26-30, 2025, Florida, US.

CHAPTER 2

Background and Related Works

This chapter discusses the literature on the most essential topics in the context of this PhD thesis. In particular, Explainable AI and its definition, Crowdsourcing, and Gamification are introduced in the context of interest, and some relevant aspects are discussed. Further details on domain- and context-specific literature will be provided in each chapter. Part of this chapter will be soon published as follows

1. A. Tocchetti, M. Bianchi, R. Campi, A. De Santis, and M. Brambilla. "On the Principles and Effectiveness of Gamification in Bidirectional Artificial Intelligence and Explainable AI". In: Bi-directionality in Human-AI Collaborative Systems. Elsevier. ISBN: 9780443405532.

2.1 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a field of AI aimed at exploring and understanding the behaviour of AI models. XAI approaches can be classified based on various features and perspectives.

The most well-known Artificial Intelligence (AI) branch is Machine Learning (ML), in which models are trained to perform predictions, classifications, and other tasks by learning from data. Over the last few years, the development of Deep Learning (DL) and Deep Neural Networks (DNN) overcame state-of-the-art models on several tasks, achieving better performance at the expense of their interpretability. Indeed, most DNNs are referred to as *black-box* (or opaque) models, *i.e.*, systems with known input(s) and output(s) and complex to understand internal logic. They are opposed to *white-box* models whose decision-making process is either known or promptly under-

standable. Understanding a model's logic is essential to ensure its correct usage, to build trust among its users, and to avoid deploying models that might misbehave (*e.g.*, due to biases) or potentially harm its users. AI researchers, developers, and experts are hence involved in an endless quest to provide evidence about the correct functioning of such models, making them transparent and trustworthy in the eyes of their users while constantly striving to achieve better performance. The primary goal of Explainable Artificial Intelligence (XAI) is to obtain human-interpretable models while achieving various objectives, like assisting individuals in making more informed choices, integrating algorithms with human values, and many more [9]. Similarly, Gohel et al. [152] stated the main focus of XAI is to answer wh-questions related to the model's output (*e.g.*, "why is that answer obtained?") to provide the following features.

- **Transparency and Informativeness**. XAI enhances model transparency, making them expressive enough to be human-understandable and to assess their performance, justification, and vulnerabilities.
- **Trust and Confidence**. Providing explanations to users improves their trust in the model and its outcomes, making humans favour the model's outcome.
- **Bias Understanding and Fairness**. XAI promotes fairness and mitigates prediction bias when interpretations are provided.

Other than the core objective of XAI, its definition, and the properties it develops, several terms are available in the literature to define what it means to explain a model or to make it explainable. These will be addressed in section 2.1.2 "Formal Definition of Explainability", following which a comprehensive definition of the main objective of XAI will be provided.

Explainability techniques can be classified in XAI based on various features and perspectives. In particular, they can be Model-specific or Model-agnostic **(I)**, Ante-hoc or Post-hoc **(II)**, and Local or Global **(III)**.

- **(I) Model-specific XAI techniques** are custom-made approaches that can only be applied to specific model categories as they are designed and implemented for specific network structures and their features. Some examples from the literature include approaches for Natural Language Processing (NLP) [15, 169, 240] and Computer Vision (CV) [72, 213, 382] models.
- **(I) Model-agnostic XAI techniques** can be applied to any ML model to inspect and understand its decision-making process. Their applicability makes them more flexible compared to model-specific approaches. LIME [355] and SHAP [262] are two of the most well-known model-agnostic approaches described in the literature.
- **(II) Ante-hoc XAI techniques** incorporate explainability approaches into a model architecture or learning process, generating a model with improved interpretability by design [353]. Examples include Decision Trees (DT) [90], Generalized Additive Models (GAMs) [66], and Bayesian Rule Lists (BRL) [244].
- **(II) Post-hoc XAI techniques** are applied to a trained model after it has produced its prediction to inspect its decision-making process and understand how

the model generated its output [353]. A few examples include Features Importance [151] and Interaction [262], and Scoped Rules [356].

- **(III) Local XAI techniques** produce explanations that provide insights into the model behaviour for a single data instance (*e.g.*, a figure in a classification model). A few examples include GradCAM [382] and LIME [355].
- **(III) Global XAI techniques** produce explanations that provide insights into the complete model behaviour, not limited to a specific data point. A few examples include SHAP [262] and GAM [193].

Concerning the model explanation process, some models are built with the intrinsic capability of providing explanations alongside their output(s). In contrast, others require applying explainability techniques to achieve the same goal. Whenever an XAI technique is applied to a model, explanations are generated and supplied to human interpreters who aim to understand the model's decision-making process. Explanations can be presented to a human interpreter in various ways (*e.g.*, highlights and textual explanations). These are not only based on the model and the design of the XAI approach but also on the expertise and background of the people for which the explanation is produced, as it is essential to make them understandable. These aspects will be explored thoroughly in the following sections and chapters.

2.1.1 Reasons for Explainability

There exist several reasons to explain ML and AI systems, the most important one being preventing potential negative effects on human lives. Improving, debugging, and understanding models are essential motivations as well.

Explainable AI is of fundamental interest nowadays since AI systems are widely applied in many aspects of our daily lives. Developers, researchers, and end-users benefit differently from such systems and as such, they have different reasons to consider model explanations important [461]. From a research and development perspective, understanding a model's logic would allow developers to assess whether the system behaves as desired, consequently allowing fixing potential faults and improving its performance. In real-world use cases, end-users would experience improved trust, consequently contributing to the widespread usage of such systems [367].

From a general perspective, it is crucial to understand model behaviour as the unjustified usage of their outcomes might negatively impact human lives. In her book "Weapons of Math Destruction" [320], Cathy O'Neil analyzes real-life scenarios in which the improper usage of AI models negatively affected people's lives. The most critical problem she identified is humans' unjustified trust in such systems. Additionally, when describing these models, she emphasizes that opacity is one of the features characterizing the so-called "Weapons of Math Destruction". Such a statement implicitly suggests that when applying models lacking transparency, it is essential to implement mechanisms to understand their behaviour to prevent severe consequences. One of

the most well-known events that raised public awareness about this problem is the improper application of a decision-making system in judging the recidivism of convicts in the USA. American courts used to apply the predictions made by the COMPAS system without even questioning them as they could not understand its behaviour properly, potentially delivering incorrect judgements. Indeed, further inspections revealed a strong bias towards flagging black people as future criminals at almost twice the rate as white people due to an intrinsic data bias. Such improper behaviour is not only strictly associated with the model's misbehaviour but also with its development and application. COMPAS's developers should have assessed the system's behaviour to ensure proper functioning. At the same time, court judges should have employed the system to support their decision rather than unthinkingly applying and trusting its prediction. Such an analysis emphasizes the fundamental role of humans in what concerns models and decision-making. Regarding such human-related problems and the applicability of AI, a few references can also be found in the General Data Protection Regulation (GDPR), a document redacted by the European Union in May 2018. While the document describes about 100 data privacy regulations, a few can be traced back to the explainability problem. In particular, article 71[1] is of fundamental interest as it regulates the automated processing of personal data, the decision-making process carried out from such processing, and its transparency. The latter became a mandatory feature for any AI system that involves *any form of automated processing of personal data evaluating the personal aspects relating to a natural person [...] where it produces legal effects concerning him or her or similarly significantly affects him or her*. Moreover, data subjects can *obtain an explanation of the decision reached after such assessment*. This last fundamental statement makes understanding the applied model's rationale necessary to provide complete and clear explanations. These examples provide concrete evidence that Explainability is a need that encompasses society and industry alike.

While there are several ethical reasons to explain model behaviour, these can be summarized into four categories from a research and development perspective [3, 461].

- **(I) Explain to Justify -** Explanations should back up model-supported decision-making to justify the final decision. Furthermore, explanations should allow the labelling of the model's decision as fair and ethical, consequently contributing to building trust in the model.
- **(II) Explain to Control -** Explanations should improve model transparency to allow the identification of potential flaws and perform model debugging.
- **(III) Explain to Improve -** Explanations should allow model developers to improve their model's performance.
- **(IV) Explain to Discover -** Explanations should support extracting and learning novel knowledge and patterns.

Even though these categories are mainly concerned with models, humans are the core of Explainable AI as developers, final users, and experts are involved regardless. Research falling in the first category **(I)** apply inherently explainable systems or explainability approaches to provide evidence for models of interest, mainly involving humans

[1] GDPR, Article 71 - https://www.privacy-regulation.eu/en/recital-71-GDPR.htm (last accessed on 17 September 2024)

to assess the understandability of the explanations. The latter is primarily achieved by showing explanations to domain experts or end users and collecting feedback. The second **(II)** category involves explainable systems and tools through which experts, AI researchers, developers, and sometimes end-users explore the capabilities of a model to identify potential errors and improperly learned knowledge. These systems usually involve multiple views and representations to detail potential perspectives that might have led the system to err. Similarly, faithfully-crafted explanations can be inspected to understand any flaw in the model's behaviour. Research from the second **(II)** and third **(III)** groups are somewhat entwined. In particular, identifying faults in a model's behaviour would allow one to fix them, consequently improving performance. In this group **(III)**, humans are usually provided with a model's behaviour, and they are asked to validate a model's capabilities and provide potential improvements by suggesting knowledge that the system might need to learn. Concerning the last group **(IV)**, humans might be involved in assessing the understandability of the knowledge acquired by a model through a set of explanations elicited. For example, novice chess players might try to learn game strategies from explanations extracted from AlphaZero [395] while expert users might discover new strategies and game patterns as the AI model might have acquired such knowledge.

Despite the need for Explainability, whether and when it is needed is still an argument of discussion in the AI community. Holm [183] states that using black-box models is motivated *when they produce the best result, when the cost of a wrong answer is low, or when they inspire new ideas*. Additionally, Explainability might not be mandatory in low-stakes scenarios where trusting a model without understanding its behaviour would not cause any harm, even if it would misbehave. In high-stakes scenarios, there are a few conditions and situations in which explaining the system's behaviour is not fundamental. For example, Miller states that *if an AI model yields accurate predictions that help clinicians better treat their patients, it may be useful even without a detailed explanation of how or why it works*[2]. While explanations might be sometimes unnecessary, a few experiments [125, 335] revealed that explaining a model's behaviour may also generate unmotivated trust. Consequently, it is fundamental to understand the role of Explainability based on the context in which the model to explain or inspect is applied and the scope in which the model deserves trust even without explainability [167].

In conclusion, explaining AI models is essential to ensure their correct functioning, to build trust and consequently drive their usage, and to avoid or detect and fix potential biases and misbehaviour. In rare cases, Explainability might be unnecessary, depending on the context. In general, proper and attentive explanations are deemed fundamental towards developing and deploying AI systems in modern society to avoid affecting people's lives negatively.

[2]Should AI Models Be Explainable? That depends - https://hai.stanford.edu/news/should-ai-models-be-explainable-depends (last accessed on 16 September 2024)

2.1.2 Formal Definition of Explainability

While several definitions of XAI and its declinations have been described in the literature, common elements and similar perspectives can be identified, finally leading to a comprehensive definition.

Acknowledged the relevance of explaining AI models, it is necessary to understand what it means to explain, what defines an explainable AI model, and some fundamental notions that are yet to find a unique definition in the XAI literature [461]. This section does not aim to provide a unique definition for each concept but to summarize the most important ones and their features.

The fundamental notion behind the concept of Explainable AI is that of "explanation". Considering a context-agnostic definition of the term, Oxford Languages[3] defines an explanation as *a statement or account that makes something clear* or *a reason or justification given for an action or belief*. Despite these definitions being detached from the Explainable AI research field, they might be considered a basic definition of a model's explanation if properly contextualized. On the other hand, a discipline-specific definition must consider the main elements of the domain of interest, in this case, the model, its input(s) and output(s), its domain of application, and the people involved with the model (*i.e.*, its developers, its end-users, etc.). In this regard, Guidotti et al. [160] define an 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*. Such a definition encompasses two essential elements. The first is the accuracy (also known as faithfulness) of the explanation, *i.e.*, the capability of an explanation to properly represent a model's behaviour. The second is the comprehensibility of the explanation, *i.e.*, the capability of an explanation to be understood by a human interpreter. The most crucial aspect that can be observed from this definition is the existence of two sides concerning explanations, *i.e.*, the human and the model, which can also be found in most definitions in XAI. In fact, explanations must be made understandable for human interpreters to clarify the model's behaviour.

Strictly related to the concept of explanation is the one of Explainability. Regarding its definition, Gohel et al. [152] define it *as a need and expectation that makes a decision of an intrinsic AI model more transparent* that also develops *a rationale approach to implementing actions driven by AI and also be helpful for end users to understand*. Such a definition recalls the need for model transparency (*i.e.*, the capacity of a method to explain how a model works, even when it behaves unexpectedly [461]) so that its actions can be understood by its end users, hence focusing once again on the need for human involvement in this complex field. Saeed et al. [367] affirm that *Explainability provides insights to a targeted audience to fulfil a need.* Similarly, Ali et al. [9] extends this definition by stating that *Explainability provides insight into a model's decision to the end-user to build trust that the AI is making correct and non-biased decisions based on facts*. The same authors [9] collected other definitions for Explainability from the literature: *the capacity to make automatic interpretations and describe the inner*

[3] Oxford Languages - https://languages.oup.com/

workings of an AI system in human terms, *the process of elucidating or revealing the decision-making mechanisms of models*, and a process *related to the ability to understand why AI models make their decisions*. While several definitions of Explainability were collected, and even more are provided in the literature, they all revolve around the core concept of making models and the process leading to their decisions understandable to humans, either intrinsically when white-box or self-explainable models are applied or through external explainability approaches.

In the explainability research field, multiple terms are sometimes used interchangeably when describing the feature of a model to be explainable. While many exist [461], we claim interpretability and understandability to be the most important for this research as they are strictly associated with the human side of Explainability. Concerning interpretability, Saheed et al. [367] define it as *the degree to which the provided insights can make sense for the target audience's domain knowledge*. Ali et al. [9] summarize interpretability as a process that *enables developers to delve into the model's decision-making process, boosting their confidence in understanding where the model gets its results and make their decisions*. Similarly, Guidotti [160] defines it as *the ability to explain or provide meaning in understandable terms to a human*. Besides, understandability is defined as *the characteristic of a model to make a human understand its function (i.e., how the model works) without any need for explaining its internal structure or the algorithmic means by which the model processes data internally* [35]. Detaching the core of the definition from the model itself, Vilone et al. [461] summarizes it as *the capacity of a method of Explainability to make a model understandable*. In this case, one might observe that these definitions mainly differ because of their perspective. The first one considers a model whose decision-making process is promptly understandable. The second focuses on an external method applied to a model to make a human interpreter understand its functioning. Such a difference is just a small proof that some terms are not uniquely defined in the literature, making it essential to provide their interpretation to clarify their meaning to the reader as they are affected by multiple factors (*e.g.*, the context).

Another term broadly used is explainable Artificial Intelligence (not to be confused with the name of the research field). Arrieta et al. [35] provides the following definition

> *Given an audience, an eXplainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*

Such a characterization makes a series of important statements. First, it clearly states that the algorithm must be able to *produce details or reasons to make its functioning clear or easy to understand*. This statement exemplifies so-called self-explaining systems, *i.e.*, models producing their outcome and its corresponding explanation together (*e.g.*, decision trees and rule-based models). Such models are either inherently explainable or trained using data and explanations (*i.e.*, human rationale) [300], finally generating models capable of explaining their behaviour. In the second place, the authors consider the audience as a relevant entity, thus acknowledging that the interpreter influences the understandability of an explanation. Indeed, understanding how to shape explanations properly [301] is as essential as understanding how they are perceived

by the audience [301, 379, 487]. For example, an AI expert would probably prefer a detailed model description, a non-expert user would likely favour a small set of examples [201] representing the system's behaviour, etc. The last aspect addressed is that an explanation must be *clear or easy to understand*. This concept is strongly subjective as it depends on various human-related factors, such as the user's expertise with AI and ML systems, the context in which they are born, their education, and many more [408]. Therefore, it is fundamental to properly understand how to tailor explanations depending on the audience's characteristics. Moreover, the proposed definition depicts a system inherently able to explain its behaviour while not explicitly considering models requiring the application of so-called post-hoc explainability techniques.

Despite the plethora of XAI-related terms and definitions available in the literature, common elements can be identified regardless of their perspective and context.

- **Model** - An AI model whose behaviour is unclear or hard to understand for which explanations are generated to explain their decision-making process or identify potential biases and flaws. The model might also generate explanations by itself (so-called self-explainable models).
- **Audience** - A group of people related to the AI model. These can be stakeholders, final users, developers, researchers, and many more [112]. Depending on their relationship with the model, they have diverse needs and usually different expertise with AI. Explanations are to be provided accordingly, as it is of fundamental interest to make them properly understand a model's behaviour from their perspective.
- **Explanation** - The mean by which proofs of a model's behaviour are provided to the audience. They must be shaped based on different factors, like the context in which the model operates, the target audience expertise, etc., to make them human-understandable. Furthermore, explanations must faithfully represent the model's behaviour and decision-making process.
- **Behaviour** - The main element to be understood by the audience. It might be a single decision, the complete model's behaviour or, more generally, its decision-making process.

Considering the broad and multifaceted scenario concerning the definition of Explainability and its features, we deem the final aim of the XAI research field to be *the development of inherently explainable systems and explainability techniques that faithfully explicit the behaviour of complex machine learning models, tailoring their explanation in an understandable way for humans to improve their trust or identify a system's flaws and biases*.

2.2 Crowdsourcing in XAI

Crowdsourcing is fundamental to engaging crowds with heterogeneous features in tasks requiring human knowledge and skills.

Data and knowledge are the most valuable assets for ML and XAI practitioners and researchers. Nowadays, data for simple or well-known ML tasks can be easily acquired

from the web, *e.g.*, from open-access data sources, repositories, and even dedicated websites and communities (*e.g.*, Kaggle[4] and Hugging Face[5]). Moreover, complex or custom ML processes may require knowledge that is not openly available or easily obtainable, and tasks requiring human knowledge cannot be promptly addressed without direct human intervention. Hence, it is essential to develop effective approaches to involve human actors in providing knowledge when needed. The ML and AI community found a solution to such needs in Crowdsourcing.

Crowdsourcing is a well-known technique that enables researchers and companies to involve a heterogeneous crowd in delivering their knowledge and skills (so-called wisdom of the crowd) to achieve specific goals [242]. These include collecting various types of data, including people's ideas and preferences [27], or the accomplishment of a task, *e.g.*, labelling a large number of images [283]. Crowdsourcing involves three core entities, *i.e.*, the crowd, the initiator, and the process [119]:

- The *crowd* is a group of geographically scattered anonymous members of a virtual community that feature various levels of knowledge, skills, and experience [208], which can span from expert roles (including scientists or domain experts), to generic citizens or amateurs [242].
- The *initiator* is typically an organization that or an individual (*e.g.*, a researcher) [242] who shares open calls within crowdsourcing communities to outsource internal tasks [187], trading crowd workers' expertise and time with monetary rewards.
- The *process* describes the initiation, execution, and result as a collection flow, typically implemented within a crowdsourcing platform through standardized interfaces or ad hoc applications.

Crowdsourcing inherently provides various advantages, like involving a crowd with diverse knowledge [321] and features, scalability on demand, and cheap data collection compared to standard methods, thanks to the existence of Crowdsourcing communities and platforms (*e.g.*, Prolific[6]). However, its capability to interest a large crowd is still a fundamental challenge, *e.g.*, high-complexity tasks may lead to low engagement as they may require specific or expert knowledge. Furthermore, its effectiveness is not only tied to the crowd but also the task and its parameters (*e.g.*, reward, complexity, clarity, etc.) [321], making tailoring an effective crowdsourcing process a challenging but essential objective. Hence, researchers have developed techniques and design patterns [47] striving to ease process complexity and to involve an even broader audience of crowd workers.

2.3 Gamification & Gameful Design

Gamification contributes to designing enjoyable applications by leveraging different kinds of motivation through design elements.

[4]Kaggle - https://www.kaggle.com
[5]Hugging Face - https://huggingface.co
[6]Prolific - https://www.prolific.com

As crowdsourcing was becoming a viable approach for data collection in complex tasks, maintaining high levels of user involvement and motivation was identified as a fundamental need to achieve a complete and accurate outcome. Across the myriad potential improvements, researchers tested gameful approaches and design patterns, unveiling their effectiveness towards satisfying those needs. Such approaches are mainly categorized as Gamification techniques, although similar terms are broadly employed in the literature (*e.g.*, Human Computation Games, Games with a Purpose, etc.) [380]. The term has not allowed for a unique definition, as researchers have provided several descriptions, perspectives, and principles characterizing it. Indeed, while a more general view about such approaches encompasses the adoption and institutionalization of games and their influence, a more specific one characterizes their capabilities in producing states of desirable experience and high engagement levels, making them a valuable asset for enhancing non-game services [99]. Indeed, one of the most well-known definitions describes Gamification as the use of game design elements in non-game contexts [99]. Gamification techniques rely on and leverage two fundamentally different types of human motivation, *i.e.*, intrinsic and extrinsic motivation [365].

Intrinsic motivation factors drive people's behaviour based on their inherent need to seek novelty, challenges, and learn. In particular, Relatedness, competence, and autonomy are three innate needs associated with intrinsic motivation [158, 365]:

- *Competence* is the need to learn new skills and master tasks (*e.g.*, it can be satisfied by providing the user with interesting challenges or feedback).
- *Autonomy* is the need to feel in control of one's own behaviours and goals (*e.g.*, it can be satisfied by promoting voluntary participation).
- *Relatedness* is the need to feel belongingness and connectedness with others (*e.g.*, it can be satisfied by connecting the user with a meaningful community).

Extrinsic motivation factors leverage people's desire to achieve a separable outcome (*e.g.*, a monetary or a physical reward).

Opposite to the Self-Determination Theory, Zichermann [532] argues intrinsic motivation is unreliable, making satisfying core intrinsic values impossible or unnecessary while discussing the capabilities of extrinsic motivation in improving performances despite hindering intrinsic motivation. Furthermore, extrinsic factors can be crafted to be perceived or become internalized as intrinsic in an effective strategy [532].

The effects of gamified approaches have been studied for a very long time in various contexts, *e.g.*, medical and healthcare [10], policy-making [42, 441–443], educational [233, 306] and many more, demonstrating the all-around applicability of such a paradigm. In particular, achieving long-lasting user involvement and high participation effort is mainly achieved by leveraging intrinsic motivation [524]. In contrast, extrinsic motivation is beneficial when an initial burst of users is to be gained or when tedious tasks are presented [62]. Extrinsic motivation is also the main driving factor in web crowdsourcing platforms (*e.g.*, Prolific) in which people are awarded money in exchange for completing tasks. On the other hand, intrinsic motivation mainly drives people to partake in social platforms in which they perform actions that benefit local

communities, research groups, or themselves [42, 441–443]. In summary, both types of motivations are considered valuable depending on the context and the environment in which they are applied and which objective must be achieved.

In practice, Gamification applies various design elements (*e.g.*, leaderboards, points, achievements, etc.), each driving or hindering either motivations and needs [368]. Classic gamification elements (*e.g.*, avatars, animations, challenges, etc.) were proven to be effective in improving user attention and enjoyment [448], consequently improving participation rate and reliability of the answers in data collection tasks [418], as well as improving user engagement and the system's capability in driving behaviours [7]. The effectiveness of each design approach and how it affects motivation depends on several factors, like the context, the audience involved, and many others. Furthermore, as the usage of design elements driving extrinsic motivation grows, intrinsic motivation is hindered [365]. The following sections will discuss these approaches and their effects in the context of AI and Explainable AI, finally summarising the most important design elements and principles. Other contexts were explored and researched, too, and will be discussed in the following chapters.

2.4 Gamification in AI & Explainable AI

This section considers areas and applications for the human-in-the-loop AI and Explainable AI research field to uncover how Gamification is applied and shapes them, focusing on the second. Two main areas will be addressed: Computer Vision, *i.e.*, a branch of AI involving visual content (*e.g.*, images and videos), and Natural Language Processing, *i.e.*, a branch of AI involving textual content and text-based tasks (*e.g.*, sentiment analysis). Examples of gamified approaches applied to other AI-related areas will also be provided. These will also be organized based on their objectives (*e.g.*, data collection, model assessment, etc.). Gamified systems will be described, highlighting gamified processes and the most critical design elements they exploit. Finally, a discussion of the main principles and design approaches inspected will be provided.

2.4.1 Gamification for Artificial Intelligence

> Gamification has been applied to AI for various purposes, developing approaches to mainly collect human-generated data or validate models' knowledge.

While AI can be applied to any application field and scenario, many complex AI models are built and deployed to process images and human text. Therefore, we focus on these two use cases as reference scenarios.

Computer Vision. In the context of Computer Vision, gamified approaches are mainly applied to label pictures, videos, or parts of them. Kotlinski et al. [222] designed a competitive game named *Detective Pig* to collect labels for complex visual data to enhance the performance of computer vision models. Players are asked to submit complex or unusual pictures, label them by choosing a category, and guess the class of other players' images. Players have limited time to guess the label, and any wrong labelling will deduct time from the clock. The more pictures they upload or guess, the

more points they earn. Players could also provide feedback on figures through likes or dislikes. Inspired by a well-known game named *Codenames*, Bitton et al. [43] collected a gamified association benchmark for multimodal machine commonsense reasoning and association abilities. A player plays as the *spymaster*, providing a single-word cue that joins a certain number of pictures and making the association hard for a model but easy for the users to guess. In contrast, an AI model and three other players try to guess which pictures the spymaster refers to by knowing the word and the number of associated pictures.

Natural Language Processing (NLP). In NLP, Gamification is mainly applied to collect knowledge and generate labelled data sets for complex tasks. Ogawa et al. [316] designed a gamified platform to collect and annotate task-oriented, situated, dialogue data, striving to motivate workers and reduce costs. The authors extended Minecraft's functionalities, allowing data collectors to customize their game instances and match couples of players to complete tasks. Players can annotate their text using a pre-defined list of buttons as they type into the chat. Ohman et al. [317] proposed a gamified framework for sentiment analysis and emotion detection to generate a dataset with multi-dimensional annotations. Their approach is aimed at satisfying intrinsic motivation [365]. Leaderboards and statistics promote Relatedness, while rank and prestige promote competence. Rank is based on a score assigned to players as they correctly annotate sentences. The score is, in turn, based on the similarity with gold annotations. Eisenschlos et al. [115] gamified a data collection approach to generate a dataset named *FoolMeTwice* containing challenging entailment pairs. Players play two different phases. In the first one, they are asked to author challenging claims supported or refuted by some textual evidence with the help of an Information Retrieval (IR) system, avoiding lexical overlapping. In the second, voters are shown two claims from other players and asked to guess which one is incorrect. The less time and evidence requested to guess, the higher the score assigned to the player. The claim's author is awarded any point lost by the voter in the second phase and whenever voters are not fooled. Venhuizen et al. [458] designed a game with a purpose named *Wordrobe* in which players are asked to answer multiple-choice questions on word senses. Players are awarded points and achievements as they complete individual or group questions, called drawers. The more complicated the questions, the more points are granted. The score is awarded based on their agreement with other players and a betting percentage of choice that will increase the points earned when answering a question correctly and reduce in case of an incorrect answer. Bos et al. [46] employed the game with a purpose designed by Venhuizen et al. [458] to examine semantic relationships expressed in noun-noun compounds. Amspoker et al. [16] developed a game named *FrameGame* for semantic role labelling involving short-story writing and expanding the FrameNet dataset. Players are asked to write a story based on the lexical units of the frame shown on the screen and to annotate frame-evoking lexical units in the story. Players can see other players' work for the same frame. Talmor et al. [435] developed a gamified application to test the limits of AI models, having participants create questions that mislead the model. Players author yes/no questions. They are shown the model's answer and asked to mark whether it was correct. Players are rewarded points whenever the model is incorrect, awarding an increased amount whenever they abide by specific suggestions

provided by the application (*e.g.*, using specific phrases). Humans then validate questions, deducting points from the player whenever the content submitted is incorrect or poorly written. This design allows players to ask complex questions without cheating while giving the game designer some control over the data submitted. Furthermore, the model is enhanced through a model-in-the-loop approach, training it on the newly collected samples. Finally, a dataset of 14.3k complex yes/no questions named *Commonsense QA 2.0* is created.

2.4.2 Gamification for Explainable AI

Proven the effectiveness of Gamification in AI, researchers developed approaches in XAI involving humans in data collection and improvement tasks, as well as model assessment and enhancement ones.

When considering Explainable AI, a few researchers recognized the potential of gamified approaches to achieve objectives like collecting data, assessing model explainability, or providing feedback to improve data collection.

Data Collection. In the context of knowledge elicitation, Balayn et al. [27, 30] proposed *FindItOut*, a Game With A Purpose (GWAP) based on the popular game *Guess Who* to elicit generative and discriminative knowledge as relationships between objects in pictures. The latter is achieved by involving people in a two-player gamified activity in which each player is shown a common board with images representing concepts. Each player is assigned one image and aims to guess the opponent's concept by asking questions shaped as relation-input couples. As questions are asked, players flip cards from their board downward until they guess the right one. The entwining between questions, flipping, and the final guess allows knowledge to be elicited.

XAI Assessment and Improvement. Aiming to assess the interpretability of model explanations, Lu et al. [261] developed a gamified approach based on the well-known Peek-a-Boom GWAP. A human actor plays the role of Peek, aiming to guess the content represented in a picture. An XAI method plays the role of Boom, incrementally revealing parts of a chosen image so Peek can guess the image content. Players are asked to perform an image classification task, while the XAI method shows the most critical non-revealed pixels whenever Peek requires it. They deem the most interpretable XAI methods to be those whose content can be guessed with a low pixel exposure rate. Fulton et al. [135] and Morrison et al. [294] perform model assessment similarly. Fulton et al. [135] explore how humans interpret AI explanations through a multiplayer GWAP for image recognition. The explainer is given the source image and its feature visualizations (*i.e.*, a technique that visualizes the features learned by the model by activation maximization [318]), and they provide the guessers with the explanations they believe would lead them to guess the correct answer the quickest. Explanations are provided over time; the faster a guesser guesses, the more points they receive. A textual explanation is also given if the guessers cannot guess based on their provided content. Morrison et al. [294] extended this gamified approach by developing a Gamified evaluation tool named *Eye into AI* to compare and improve XAI methods. The game involves several

rounds with different functions. In the explainer's round, the player is given a source picture of choice and a list of explanations (*i.e.*, the best and worst five from the explainability method's perspective). They are asked to choose the top four out of ten that would allow a player to guess the original image and rank them. In the guesser's round, the player is given the top explanations (in the first match) and random explanations (in the second match) using the same XAI technique. Similarly to Fulton [135], explanations are slowly revealed over time; the lower the number of displayed explanations, the higher the score. If the player cannot guess based on the content they were given, a textual explanation is also provided, after which they are shown a list of four options to choose from if they still can not guess. The authors compute the agreement between humans and XAI approaches to understand which XAI techniques are less intuitive. The exposure and the number of explanations required to lead to a correct interpretation are employed to rank the interpretability of XAI techniques, similar to Fulton et al. [135]. Ma et al. [263] proposed a gamified web platform to assess explainable AI approaches (*i.e.*, local feature importance and counterfactuals) engaging non-expert users in the context of time-series forecasting. Through qualitative options, users are asked to estimate the relationship between the model's predicted price and a hypothetical estimated price. Each player plays five rounds. In each round, the user can ask for model explanations. The closer their hypothesis is to the model prediction, the higher the score they are awarded. Recently, Humer et al. [192] applied a gamified application, named *Forestly*, to involve 500 participants in assessing which explainability techniques are more trustworthy in high-stakes scenarios. In particular, they were asked to identify mushrooms as poisonous or edible with the help of an AI-based application, potentially supported by a specific XAI technique. They were shown the picture of a mushroom, the outcome of an XAI approach depending on the group they were assigned to, the predictions from a model with its probability, and the name of the species and whether it is edible. Then, participants were asked whether they would pick the mushroom for later consumption, making it more realistic. The gamified experience was also expanded by implementing an interactive game named *AI Forest - The Mushroom Hunting Game*.

Natural Language Processing. Significantly, few researchers employed Gamification in the context of Explainable AI in Natural Language Processing, making it an exciting area to explore. In particular, Sevastjanova et al. [384] combined Gamification and Explainable AI, developing a visual analytics technique, named *QuestionComb*, for interactive data labelling to generate high-level relationships and rules for texts in the context of Question Answering. The gamified activity aims to improve a model's quality, awarding users badges whenever they complete a level or successively improve the model's certainty. Users are provided with an interface to explore the dataset, choose and annotate an instance, organize the labelled instances, and finally, overview the rules based on the previous grouping. This structure allows users to explore the model performances and its internals (*i.e.*, the rules it learned) as the user labels data instances. Gamification is applied through well-known game dynamics, namely content unlocking (*i.e.*, only a part of the data is accessible at a time, highlighting enabled and disabled instances); freedom of choice (*i.e.*, users can label data instances of choice, unless they are not allowed to); collection (*i.e.*, users can group data instances through the interface, enhancing their reasoning); levels; and badges. Recently, Takan et al. [434]

applied Gamification to bring together proposals for AI fairness tests into a single platform to create novel sustainable fairness practices ultimately. Gamification allows users to see their scores and awards them by sharing their achievements with other users on the platform, leveraging extrinsic motivation. In the system, users can set rules with a description and develop tests to assess the conformity of texts to rules. The platform computes a score based on the rules and the tests using a ZDD (Zero-Suppressed Decision Diagram) score and centrality score, as well as a confusion matrix and mutation test scores, respectively. The sum of these scores constitutes the score assigned to the user and its influence on the game, so players with a higher score have a higher influence on the game.

Feedback. Gamification and Explainability were also combined to provide feedback to improve data quality. In particular, Shingjergji et al. [392] designed *Facegame*, a gamified application for collecting facial expressions, addressing the need for data and model interpretability simultaneously. Players are exhibited a picture with a facial expression to mimic. Their facial expression is shown through a camera. The closer their expression is to the given one, the higher the score they are awarded. The latter is also affected by the quality of the picture (*e.g.*, a dark picture might result in a lower score). Developers employ this score to provide feedback to players, motivating them to create better images and to get a better score. Feedback is provided through explanations as free text. These are generated by comparing the player's facial expression with the primary muscle movements associated with Explainability using a rule-based dictionary approach. These represent user feedback and a human-friendly, interpretable explanation of a facial emotion recognition model. A combination of Explainability and gamified approaches was also employed in other contexts, like security and education. Suhail et al. [424] implemented *ENIGMA*, a gamified approach for addressing the explainable security challenges for Digital Twins, providing security analysts with a controlled and supportive virtual training environment. Security analysts play as the attacking team, while an Explainable security assessment model explains the defending model's decision. Objective-game-based scenarios are generated by configuring the teams, roles, and resources. The game unfolds by simulating an attack on a digital twin instantiated on a virtual machine to not affect an ongoing process. A post-hoc explainability approach is applied to explain the defence mechanisms used by the AI model, making the final security decision understandable to security analysts.

Education. Striving to educate students and advance education on robust AI, Gelata et al. [143] implemented an open-source game-based platform named *Maestro*. Students partake in a competitive programming environment where they are challenged to achieve better model performances than some baselines and their competitors in goal-based scenarios. The authors leveraged intrinsic motivation through a leaderboard displaying students' scores, ranks, and errors, fulfilling correctness, competence, and autonomy [365].

2.5 Gamification Principles for AI and XAI

Common features and principles were identified from the gamified approaches in the literature (*e.g.*, the most commonly employed gameful design elements and design patterns).

This section will discuss the most commonly applied approaches, principles, and design patterns from the collected literature in AI and XAI.

Usage of Well-known Games. Some of the presented works employ and extend well-known games and game mechanics to implement their gamified approaches. *Guess Who* [27,30], *Codenames* [43], and *Minecraft* [316] are just some examples constrained in the context of AI. Such an approach reduces usability barriers as participants might know how to play the game beforehand. This effect might be reduced when games are extended to fit the authors' use cases as new mechanics might be introduced. In this case, it is essential to clarify how the extension updates or improves the original game. Furthermore, these games are usually board or video games, making them fun and appealing to the public by design, potentially increasing the quality of the outcome and engagement. On the other hand, custom gamified activities might be too complex for new players to understand, include unbalanced rewards or game mechanics, or require a lot of design, implementation, and testing efforts.

Multiplayer Game Design. Multiplayer games are designed to satisfy relatedness [365], making players cooperate for or compete towards a common objective. Cooperative games usually ease complex tasks or ensure higher data quality by combining knowledge from various actors. Instead, competitive games [27, 30, 135, 294] drive players to perform well as they are moved by their desire to win and achieve better results. From the players' perspective, it is essential to structure the competitive aspect to avoid generating hatred towards other players, which might cause strong disengagement. Such behaviour is more common when competitive players are engaged in an environment that strongly rewards veteran players, making it hard for novices to compete. On the other hand, it should not be that easy for novices to reach veterans' achievements. Consequently, in-game reward and penalty (*e.g.*, points) balance is fundamental to achieving short- and long-term engagement. When multiplayer games are set in place, it is necessary also to employ mechanisms that make the game playable even when one player is not active, *e.g.*, introducing a time limit for each turn would allow a game instance to proceed even with a single player. Similarly, a penalty or reporting system for leaving or not playing the game might be set.

Gameful Design Elements. In gamified approaches, multiple elements can be employed to drive motivation. Depending on the component, intrinsic or extrinsic motivation is driven [368]. From the collected literature, it can be observed that points [115, 135, 143, 222, 263, 294, 317, 392, 434, 435, 458] are the most commonly employed gamified elements, followed by leaderboards [143, 317] achievements [434, 458], and badges [384]. Gamified design based on time [115, 135, 222, 294] are also broadly applied. Points are among the most common in-game rewards that drive extrinsic moti-

vation. Players are awarded points for completing activities or performing high-quality tasks. When it comes to points, it is essential to keep awarding players throughout the gameplay, as these rewards are effective because the player feels a sense of satisfaction in increasing their score. Leaderboards are strictly related to points, as most rank players based on their score. Sometimes, the number of achievements or badges a player achieves is also used to generate leaderboards or resolve potential ties between players. When applying leaderboards, these must award the most performing players while driving the low-performing ones to do better to improve their rank. On the other hand, ranking should not make players feel a sense of hatred towards those with a higher score or discourage them from playing the game. The latter can be achieved through well-balanced rewards and applying time-based leaderboards (*i.e.*, other than commonly applied global leaderboards, one might generate leaderboards that consider the best matches performed by players over a specific period to award players for playing constantly and doing well, making it possible for everyone to get at the top after each leaderboard reset, *e.g.*, including daily or weekly leaderboards). Achievements (and badges) reward players for accomplishing specific objectives and award them status within the community. Objectives of different complexities must be considered, varying from simple ones (*e.g.*, completing the tutorial) to complex ones, making it rewarding in the beginning and challenging for the player to complete them. Time is also a fundamental design element when it comes to gamification [115, 135, 222, 294]. It drives the game design, making players earn more points the quicker they complete the activity when a correct answer is provided. As previously discussed, time can also be applied to other aspects of gamified services.

Other than these common design elements and approaches, An interesting gameful design was applied by Takan et al. [434]. They made it so that players who achieve better results not only achieve a higher score but also have a stronger influence on the game. This design approach acts as a form of status as a reward, granting good players an influencing reward and status towards the network of players.

Fun & User Engagements. The most important factor that makes Gamification effective is fun [85]. A gamified application must be enjoyable for users, finally creating long-lasting engagement. Furthermore, when players enjoy partaking in an activity, they provide extended and higher-quality content. While well-known games are most likely enjoyable for the public as they were pre-tested and well-designed, custom games must be assessed to validate their capability to keep the user engaged. The main challenge associated with this topic is that fun is highly subjective and strictly depends on the players' objectives [168]. Similarly, players might enjoy or dislike the gamified design of an application. Other than the presented gamified approaches in the context of XAI, various design patterns have been discussed in the literature, like the so-called *games for fairness and interpretability* [85]. These include *Human versus AI* and *Break the bot* games. The first category includes games where the human competes against the model to guess the right answer. In the second approach, players are shown an input and asked to change it to change the model the most. The winner is the one capable of creating a greater change.

Cheating & Exploits. When playing gamified activities, some players might try to cheat or "game the game", *i.e.*, exploiting unwanted gamified behaviours in the gamified application that lead to achieving high scores or performances without properly playing the game. Unless someone discovers and reports these behaviours, these situations are usually unknown to the authors. On the other hand, players might cheat using other methods, *e.g.*, they might get in touch with each other through internal or external channels to make arrangements to get better results. The main concern with cheating in gamified approaches is that the outcome (*e.g.*, a dataset) might be affected by such behaviours, leading to strongly biased or untrustworthy outcomes. These behaviours must be prevented by designing well-balanced and exploits-free games that properly award players based on their performance, avoiding including mechanisms that might lead players to cheat, or including mechanisms to report such behaviours [27]. This is one of the main reasons external rewards, like real-life rewards, are highly discouraged, as they might drive players towards finding exploits to get more and more of these rewards. In-game extrinsic rewards might achieve a similar effect with a lower scale and effect, as players strive for a feeling of greatness by getting high scores or rankings compared to their peers.

Gamification-AI Integration. From the collected literature, it is worth noticing that while most researchers focused on collecting data for [16, 27, 30, 43, 46, 115, 222, 316, 317,435,458] or assessing explainable systems and approaches [135,192,261,263,294], some of them combined explainable AI and Gamification to achieve other objectives (*e.g.*, providing feedback [392, 424] or educating people on these topics [143]). XAI-based feedback [392, 424] can help users perform activities involving AI systems, allowing them to understand a model's behaviour and potentially improve theirs, finally leading to an enhanced output. In these cases, gamified systems contribute by somewhat reducing the complexity of explainable AI content to the users as these must be user-understandable, similar to the gamified approach [294]. When it comes to education, integrating explainable AI allows learners to improve their understanding of a system's behaviour or whether the approaches they applied were effective and how to improve them [143] potentially.

Gamification Principles. In the considered literature, a series of theoretical principles driving gamified services and user behaviour have been identified. Although researchers employed intrinsic motivation principles [143], extrinsic ones are mainly applied (*e.g.*, points, leaderboards, etc.). When designing a gamified system, its objective must be transparent to make players become connected to its cause [16], motivating them to play and cooperatively achieve a common goal. Successfully achieving an objective requires a community of people actively contributing towards it [16], promoting Relatedness between their members [365]. Similarly, orthogonal game elements must be aligned with a stated goal [16].

Competence [365] is satisfied by creating challenging tasks through which players can constantly learn and improve, as well as allowing players to solve human computations problems at hand [294]. Implementing skill-intense tasks can be helpful to highlight players' creativity [16], and can be achieved by designing GWAPs with multiple complexity levels, awarding the players with badges and achievements whenever

they accomplish them and mastering skills to strengthen user awareness and possession [384]. Activities can be challenging when people performing them are uncertain about the outcome [143].

Furthermore, feedback on in-game behaviour and outcomes is essential in the design to effectively improve players' skills and competence [143, 294]. As players are engaged in a gamified system that includes various activities or tasks with multiple complexities, they must be free to choose among those while still being guided based on their skills [384]. The gamified process is not always about the players. Indeed, a gamified process must contribute towards the objective set by its creator, finally organizing players' data into task-relevant outcomes [294].

Reduced Knowledge Barriers. Masking the inherent complexity of explainability approaches and explainable models is another fundamental reason why Gamification can be considered an effective asset in this context. Indeed, some researchers explicitly have demonstrated that an XAI-based gamified approach is not too complex for non-expert users to understand [294]. Similarly, one might argue that most discussed articles intrinsically demonstrate it as they successfully achieve their objectives when engaging novices. One of the main reasons why Gamification works in this context is that users do not perceive the complexity behind the explanations or the explainable model as they are primarily involved in activities designed to be simple and enjoyable (*e.g.*, guessing games). This introduces a tradeoff between the design of the gamified approach and the knowledge needed by users who contribute towards the final objective. Indeed, the more the application is gamified and the resulting task is simplified, the less knowledge it will require. On the other hand, complex and less gamifiable tasks might require knowledgeable users, making it essential to engage experts and potentially making Gamification unnecessary.

CHAPTER 3

Crowdsourcing

This chapter discusses the research on crowdsourcing and engagement in various contexts (mainly policy-making) with the final aim of understanding fundamental principles to engage users properly. Understanding these principles and their role in designing crowdsourcing applications and approaches is essential in human-centred XAI. This chapter is mainly built upon the articles.

1. A. Tocchetti, L. Corti, M. Brambilla, and D. Di Marco. "A Web-Based Co-Creation and User Engagement Method and Platform". In: Lecture Notes in Computer Science. Springer International Publishing, 2021, pp. 496-501. DOI: 10.1007/978-3-030-74296-6_38. URL: https://doi.org/10.1007/978-3-030-74296-6_38

2. A. Tocchetti and M. Brambilla. "A gamified crowdsourcing framework for data-driven co-creation of policy-making and social foresight". In: CEUR Workshop Proceedings 2736 (2020), pp. 34-44. URL: http://ceur-ws.org/Vol-2736/paper6.pdf

3. Andrea Tocchetti, Silvia Maria Talenti, and Marco Brambilla. *My Lockdown Escape*: A Data Collection Approach based on Gamification and Crowdsourcing for Subjective Perspectives, Self-Empathy, and Memories about Past Experiences". In: Proceedings of the 3rd Empathy-Centric Design Workshop: Scrutinizing Empathy Beyond the Individual. EmpathiCH '24. Honolulu, HI, USA: Association for Computing Machinery, 2024, pp. 14–20. ISBN: 9798400717888. doi: 10.1145/3661790.3661794. url: https://doi.org/10.1145/3661790.3661794.

4. Andrea Tocchetti, Lorenzo Corti, Marco Brambilla, and Irene Celino. "EXP-Crowd: A Gamified Crowdsourcing Framework for Explainability". In: Frontiers

in Artificial Intelligence 5 (2022). ISSN: 2624-8212. DOI: 10.3389/frai.2022.826499. URL: https://www.frontiersin.org/article/10.3389/frai.2022.826499.

The PhD candidate contributed to researching the literature about Gamification, crowdsourcing, the design of human-centred applications, and empathy (1 to 4). Following this research, the candidate designed and implemented various applications (1 to 3) and methods (4) to engage users to collect their thoughts and opinions (1 to 3) and XAI-related data (4). In (1, 2, and 4), the candidate directly redacted the article and designed and implemented the method. In (3), the candidate contributed to the research with their expertise on the topic and by writing some background sections.

This chapter is organized to cover one context at a time, reporting on the essential principles learned and their applicability to designing an XAI platform to engage researchers and end-users. The considered contexts are policy-making, COVID-19, empathy, and XAI.

3.1 Gamified Crowdsourcing in Policy-making

Acknowledging the complexity of collecting citizen feedback and the recent success of web platforms and social media in policy-making, a gamified web platform has been developed to explore crowdsourcing applications and understand essential principles for their development.

Over the last decades, the communication between governments and citizens has worn thin, leading to an overall loss of faith and interest in politics. People feel unrepresented by their governments, believe their decisions do not affect society, and feel overlooked by public authorities, thus giving up on exercising their participatory rights. Indeed, political alienation is a typical trait of contemporary societies [271] both at local and national levels, mainly associated with young people. Nonetheless, involving citizens in public decision-making processes is gradually proving to be a new way to overcome long-lasting symptoms of a democratic deficit in modern societies, *i.e.*, the reluctance to state one's opinion publicly, declining voter turnout, diminishing participation in public debate within institutions, and many more. Consequently, public organizations are encouraging citizens to get involved in finding solutions to problems in the public sector for the sake of the common good. In this regard, the way people partake in political debates has changed over the last few years. Although voting is still perceived as the primary approach to participate in public debates, modern forms of participation (*e.g.*, discussions in online communities and social networks) are spreading among the youngest generations. These trends show a new approach to increase engagement, participation, and contribution in public debates and to shortening the gap between decision-makers and citizens is necessary.

Well-known and effective approaches to drive participation and engagement were analysed, finally contributing to the design and implementation of a framework combining crowdsourcing, Gamification, and data analysis to improve the crowd's collec-

tive engagement and contribution to policy-making decisions. The proposed approach moves the focus and objective of crowdsourcing and co-design practices from traditional, short-term goals to long-term visionary changes at the societal, economic, and political levels. Such a paradigm shift opens new challenges in terms of citizen engagement and participation, as well as in terms of their contributions. Furthermore, empathy and empathy-driven design are applied to engage stakeholders in co-designing solutions and ideas while at the same time contributing to design knowledge. The core contribution of this work is the development and implementation of a gamified web-based platform to

- collect ideas and thoughts by engaging (local) communities to partake in a web-based platform to brainstorm and co-create solutions to modern and future societal problems,
- enhance governments' foresight capabilities to achieve a meeting point between governments' actions and people's shared vision of the future,
- improve the communication between (local) governments and (local) communities, contributing to recovering trust in public institutions.

3.1.1 Context-specific Related Works & Background

Crowdsourcing in Public Engagement and Policy-making. The current lack of trust in public institutions has made engaging citizens in public decision-making a tough challenge. Recently, researchers cooperated with local administrations to develop methods and systems to achieve such a goal. Most solutions were implemented as digital tools, like web platforms and social media. "Love Your City" [419] allowed citizens to directly address fellow citizens or authorities, create solutions to proposed problems, and organize public events. "Decide" [97] empowered citizens by allowing them to propose new laws, to vote opinions about public proceedings, and to debate on how to redistribute the city's budget among projects. Similarly, "MindMixer" [165] is a platform through which citizens can express, support, and comment on public proposals as solutions to societal challenges. Researchers contributed towards these objectives as well. Bianchini et al. [42] promoted a gamified two-way exchange of proposals between politicians and citizens. Citizens create and vote for proposals, while politicians answer to these proposals. Structured and goal-oriented discussions are held while satisfying three levels of participation, *i.e.*, provision of information, engagement, and empowerment. All these platforms digitally involved citizens in partaking in the local policy-making scenario, proving their effectiveness in achieving citizens' involvement. Additionally, most solutions also apply Gamification to make the experience even more enjoyable.

3.1.2 COCTEAU - CO-Creating The European Union

The COCTEAU gamified web platform for crowdsourcing citizen feedback was designed, implemented, and finally tested in various events, reporting on its effectiveness and fundamental features.

This section describes COCTEAU (Co-Creating The European Union)[1], a web-based gamified application that enhances the interaction between citizens and policymakers, also exploring the role of empathy in engaging public conversations at scale. Citizens are involved through a set of gamified activities through which they share their thoughts with the community and debate about other members' opinions. Policymakers manage the platform by creating Scenarios, *i.e.*, summarized descriptions of real-life situations they would like citizens to relate to. A dashboard is implemented to provide insights into the content shared within the platform. Such a dashboard is available on PERSEUS[2], a complementary tool offering policymakers an environment to adopt evidence-based, foresight-based, and sustainability-oriented decisions.

The main actors involved in COCTEAU are citizens and policymakers. Citizens are engaged on the platform to share their thoughts and debate about shared content to let policymakers hear their voices. On the other hand, policy-making domain experts (including but not limited to politicians) and researchers manage the platform by creating, sharing, and maintaining context-specific content as they are interested in citizens' thoughts to guide their decision-making process. The key aspects on which our tool is based are summarized hereafter.

- **Engagement & Gamification**. The platform must keep users engaged as much as possible. Therefore, designing interesting and enjoyable activities for citizens and domain experts is necessary. One of the most used techniques to enhance and achieve engagement is Gamification. Its goal is to promote people's intrinsic motivation towards different activities by exploiting game elements and game design techniques. Intrinsic motivation factors (*e.g.*, self-improvement) were proven to be the most successful way to generate a greater feeling of engagement. However, extrinsic motivation elements (*e.g.*, points, leaderboards, etc.) may be employed to achieve a good initial level of engagement.

- **Community**. The platform aims to develop a community of proactive people. Therefore, a focus on collaborative and interactive activities is required since building a united community is a great way to increase the quality of the content provided by its users while maintaining high participation.

- **Feedback**. Even though policymakers receive feedback from the citizens through the content they share, providing citizens with feedback would also increase their satisfaction. Domain experts are prompted to share their thoughts with the community to let them know they are listened to. Moreover, other forms of feedback might motivate users, *e.g.*, showing the overall activity level of the platform encourages its users to keep using it.

- **Empathy**. The proposed approach is also based on engaging the user on an empathetic level. Citizens empathize with the thoughts shared by others to expand their opinions on the subject of discussion by observing multiple perspectives. This principle influences the design of most activities, establishing emotions as one of the most relevant elements to engage citizens.

[1]COCTEAU URL - https://www.cocteau.eu/ (Last accessed 14 October 2024)
[2]PERSEUS URL - https://perseus-platform.eu/ (Last Accessed 14 October 2024)

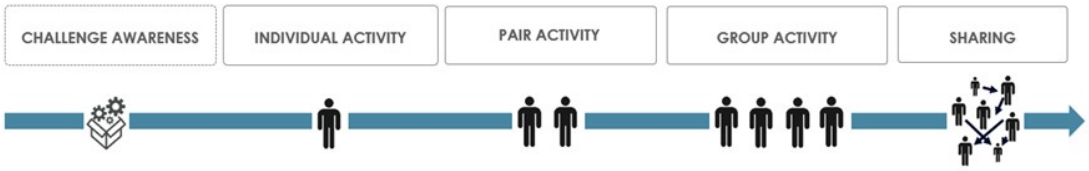


Figure 3.1: *A schematic representation of the five steps involved in the COCTEAU activity process during the preliminary validation at a workshop.*

Initial Validation. The interaction flow on which the COCTEAU platform is based was evaluated in an in-person workshop to gather feedback on its effectiveness in a co-operative and interactive environment. Furthermore, some core principles and design elements were validated too. In the end, the designed approaches were confirmed to be effective, although a few doubts arose. The chosen topic for the experiment was Artificial Intelligence. It was divided into five different phases representing the interaction flow to be implemented on the platform. Participants were given an envelope containing all the necessary material before the beginning of each step (as shown in Figure 3.1).

- **Challenge Awareness**. Participants were given an envelope with a set of pictures, descriptions, and examples to understand the scenario (*i.e.*, a matter for which policymakers must make a decision to tackle it).
- **Individual Activity**. In the second step, participants were provided with a sheet describing four different Artificial Intelligence application cases in the form of text printed on two tags each. Each participant had to choose only one of the proposed challenges, write their name on the corresponding sticky tag and stick it on a shared billboard representing a graph with two axes. The x-axis ranged between "Incremental" and "Disruptive", while the y-axis ranged between "Fear" and "Hope". Participants were also asked to stick the corresponding identical tag on a sheet. Then, they were asked to write the motivations for positioning the tag on the graph.
- **Pair Activity**. In the third step, couples were composed to make people discuss. Such matches were formed according to the distance between the tags on the billboard. For each couple, the members exchanged the sheet compiled during the previous step so that participants could understand the point of view of the person they were discussing with. After debating, each participant compiled a questionnaire explaining what they agreed or disagreed with during the interaction. Then, a picture-based activity was performed to have each couple converge on a single common opinion on the matter (so-called vision). Therefore, each couple was asked to pick one picture representing their vision about the future relationship between humanity and Artificial Intelligence and write three meaningful keywords to describe it. Finally, opinions were shared with other groups, and each couple voted on the ones they would like to be matched with for the final activity.
- **Group Activity & Sharing**. New groups based on the previous voting phase were created in the last step. Then, each group discussed and converged on a single

vision among theirs, for which they were asked to provide a title, their thoughts, and three new keywords. Then, all groups were asked to share their final opinion and thoughts with the community.

In the end, all the groups converged to a single vision. The participants stated the experience was properly engaging and even enjoyable. On the other hand, one of the most criticized aspects was the dimensions on the graph for classifying the Scenarios. In particular, "Fear" and "Hope" were perceived as too extreme to represent their feelings about the proposed cases, calling for a novel approach to represent their feelings. The meanings of the different axes were also difficult to grasp. Participants also claimed that the proposed challenge was too limited to evaluate and fully appreciate the method. A nice unexpected behaviour was that, while choosing the most representative image, some couples chose multiple pictures instead of one, cutting and merging them into a single image. They asserted their thoughts were too complex to be represented with a single picture since none of the pictures they could choose from perfectly matched their opinion. Such feedback contributed to the design and implementation of the platform.

User Interaction Flow. COCTEAU users are engaged in activities through which policymakers collect insights on their thoughts. As a community member interacts with the platform for the first time, a strict activity path is enforced to guide them through all the functionalities the platform implements (see Figure 3.2).

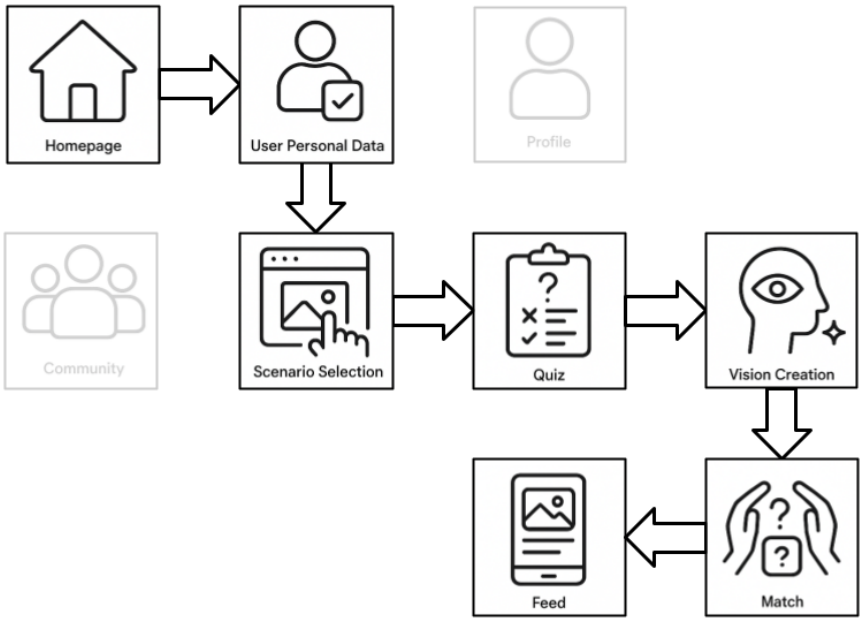


Figure 3.2: *The user interaction flow on COCTEAU the first time a user logs.*

After a member completes the first interaction pathway, the interaction flow slightly changes, allowing for higher flexibility (see Figure 3.3).

The activities included in the user interaction flow are the following.

- **Home Page**. The home page of the platform. It includes the login, a section with some information about the platform design and activities (*i.e.*, the classical

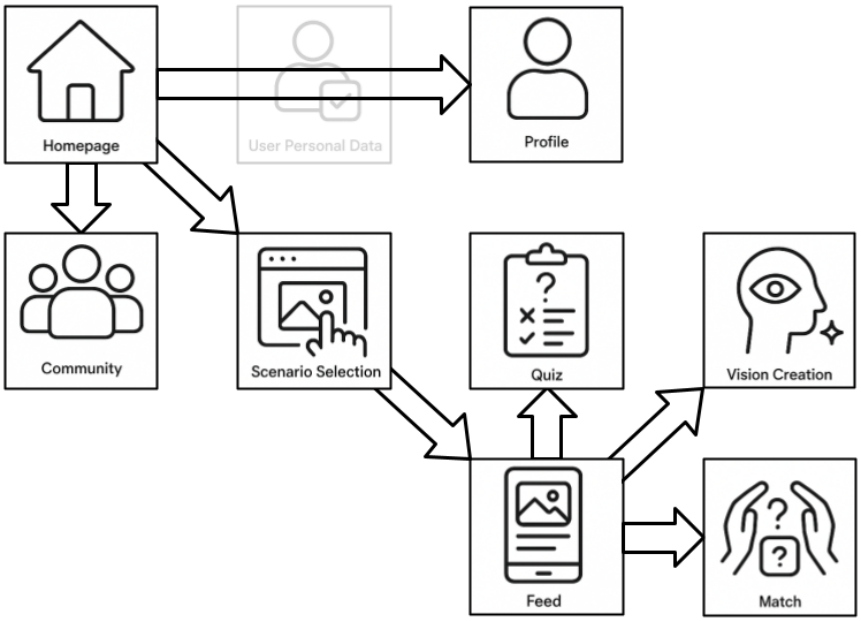


Figure 3.3: *The user interaction flow on COCTEAU. The User Personal Data step is the only step performed once.*

"about" section), the list of scenarios the user can interact with, and a set of languages the user can choose from (English and Italian as of this publication). The latter allows for language-specific scenarios within the platform. Users must accept privacy and cookies to join any scenario. These would allow tracking users' interactions on the platform by assigning them a unique identifier.

- **User Personal Data**. The only one-time activity across all scenarios is collecting users' data, *i.e.*, username, password, nationality, education, age, and gender. This process is not mandatory for a user to register on the platform. It involves a series of multiple-choice questions that allow profiling of users to carry out aggregated analyses. Most questions involve ranges (*e.g.*, age), making it not too personal but still useful for the analyses. The provided data are used to create an account for the user to keep track of their activities on the platform.

- **Scenario Selection**. In this section of the platform, users pick scenarios to explore. Each scenario comprises a picture and a description of a hypothetical scenario for which policymakers must make an informed decision. Scenarios are grouped into so-called narratives (*i.e.*, topics of interest). Scenarios are created and managed by policymakers.

- **Quiz**. A quiz involving a series of questions aimed at collecting structured opinions on matters of interest. Each question comprises a text and a fixed set of five answers (*i.e.*, strongly disagree, disagree, impartial, agree, and strongly agree). Each quiz is made of at least five questions. Users are asked to complete the quiz the first time they engage with the scenario. Domain experts may add additional questions for the users to answer at any time.

- **Vision Creation**. A section through which users can structure and share their thoughts with the community as so-called visions. A vision is made of a picture for which four different shapes are available, each involving a different number of picture slots to fill, a set of keywords and a textual description detailing the user's thoughts, and a graph based on Plutchik's model of human emotions [334] from which to choose the feeling associated with the vision. Keywords used when looking for figures through an integrated Unsplash API[3] are stored too. Users are prompted to create a vision the first time they explore a scenario. Users can create as many visions as they want after the first iteration.

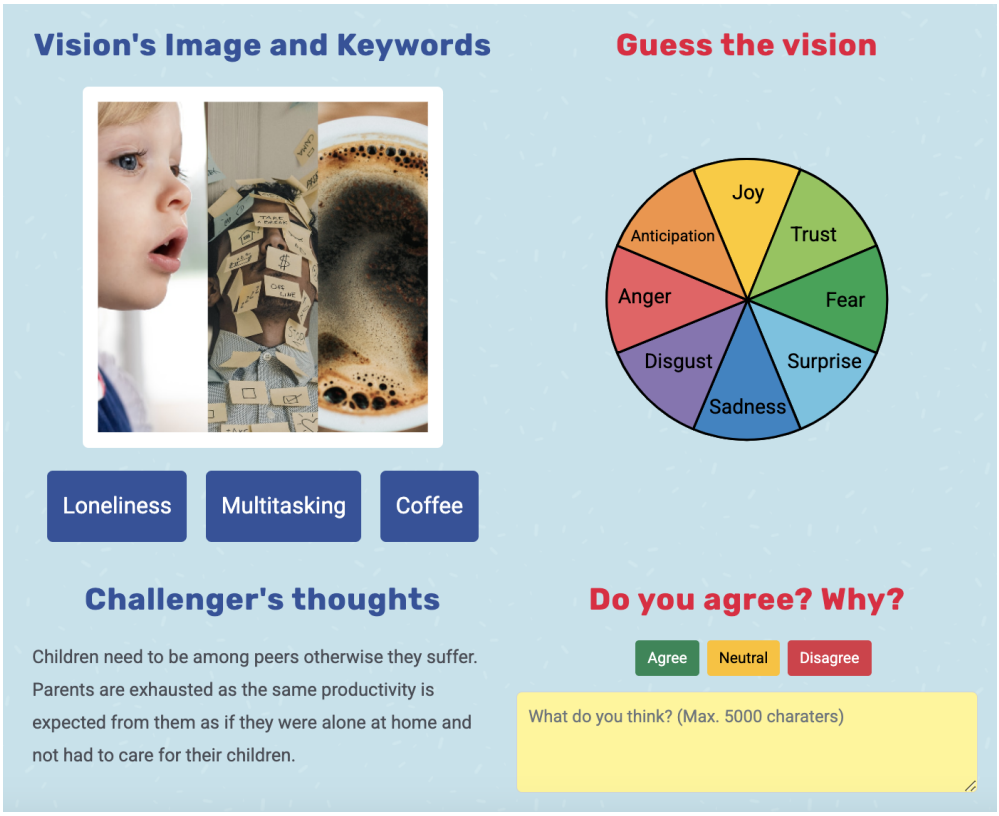


Figure 3.4: *A screenshot of an in-depth match. The vision, its keywords, and its description are displayed on the left, while the wheel of emotions, among which the player chooses, is displayed on the top right. Players can also state whether they agree or not and why.*

- **Match**. There are two types of matches: a quick one and an in-depth one (represented in Figure 3.4). Each match involves the current user and a vision made by another user. In a quick match, the user is provided with the picture and the keywords of the vision they are playing against. They are asked to guess the emotion of the vision by choosing one among these in Plutchik's model of human emotions. We approximate how empathetic a citizen is by measuring the distance between their guess and the emotion of the vision's creator. Consequently, the closer the user guesses, *i.e.*, the more empathetic they are, the more points they are awarded. An in-depth match is an extension of the quick match in which

[3] Unsplash https://unsplash.com/it (last accessed 21 October 2024)

Figure 3.5: *A screenshot of the part of the Community section displaying the most important users based on some criterion (e.g., the player with the highest score awarded with the Champion title)*

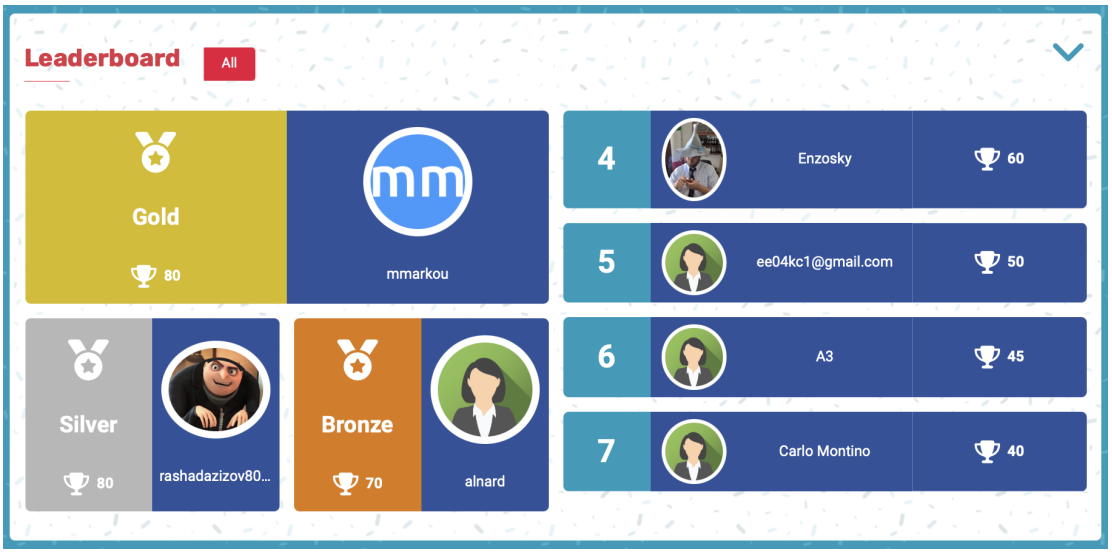


Figure 3.6: *A screenshot of the part of the Community section displaying the users with the highest score across the platform.*

the player is provided with the whole vision, and they are asked to provide their thoughts about it, *i.e.*, whether they agree with the vision and their thoughts as a textual description. Users are prompted to play five matches of any kind the first time they join any scenario. Users can play as many matches as they want after the first iteration.

- **Feed**. The feed displays the collection of all the visions shared in a chosen scenario by all the users who explored it. Each vision is displayed as a combination of its picture and its keywords. This section also allows the users to challenge a specific Vision shared by another user rather than a randomly chosen one when playing a match. When users reach this platform section, the first interaction flow is over.

- **Profile**. The representation of a user on the platform. It is made of various elements, *i.e.*, basic data (*e.g.*, nickname, description, etc.), avatar (*i.e.*, a customizable image identifying the user on the platform), activity frame (*i.e.*, a form of

"status as reward" awarded depending on the activities performed), most recently shared visions, and a list of achievements achieved.

- **Community**. An additional page displaying users and the visions awarded with status within the community. These include achievements (*e.g.*, the player who played the most matches) (see Figure 3.5), and a leaderboard displaying the players with the highest score (see Figure 3.6).

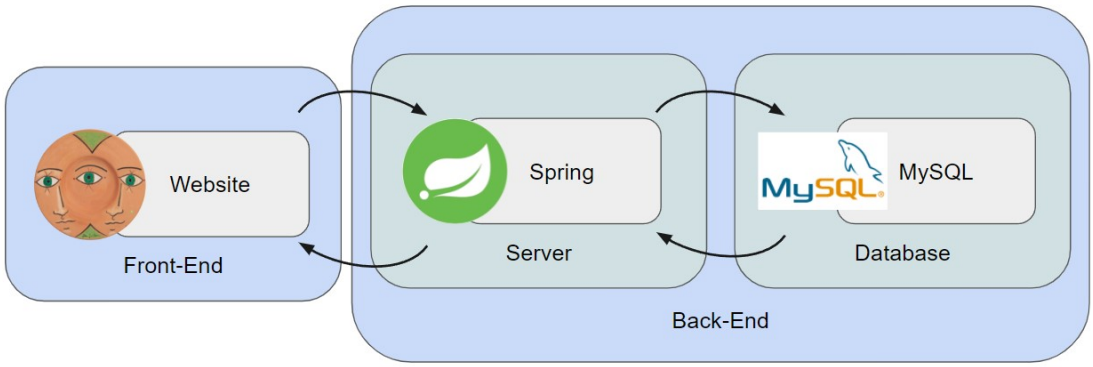


Figure 3.7: *A simplified view of the interactions between the various components included within COCTEAU. The front end interacts with the back end through the server, which collects or stores data from the database.*

Implementation. The online version of COCTEAU was implemented by adhering to the Web MVC (Model-View-Control) framework as a Spring Boot application. Figure 3.7 provides a simplified overview of the system. The platform's structure enables two key aspects from the developer and user perspective. The first aspect is the high code maintainability, as software components have clearly defined roles and responsibilities. The second aspect is a consistent and localized user interface. A summary of the navigation tree for the COCTEAU platform is shown in Figure 3.8, including all the sections implemented.

Use Cases & Analyses. The tool was openly released and employed in public events, collecting data about topics of interest (*e.g.*, the COVID-19 pandemic) and feedback about its usage and design. In particular, about 215 participants shared 85 visions and played more than 280 matches. These include policy-making- and healthcare-centred events in Brussels and workshops organized with master students (mainly aged between 24 and 25) at Politecnico di Milano. However, users' profile was collected through the platform, making it complex to outline the participant's features as these are optional. Hereafter, some analyses of the COVID-19 pandemic scenarios are reported, focusing on the shared and perceived feelings and their relationships with the most important keywords used to describe the shared visions and the pictures the users employed to describe them.

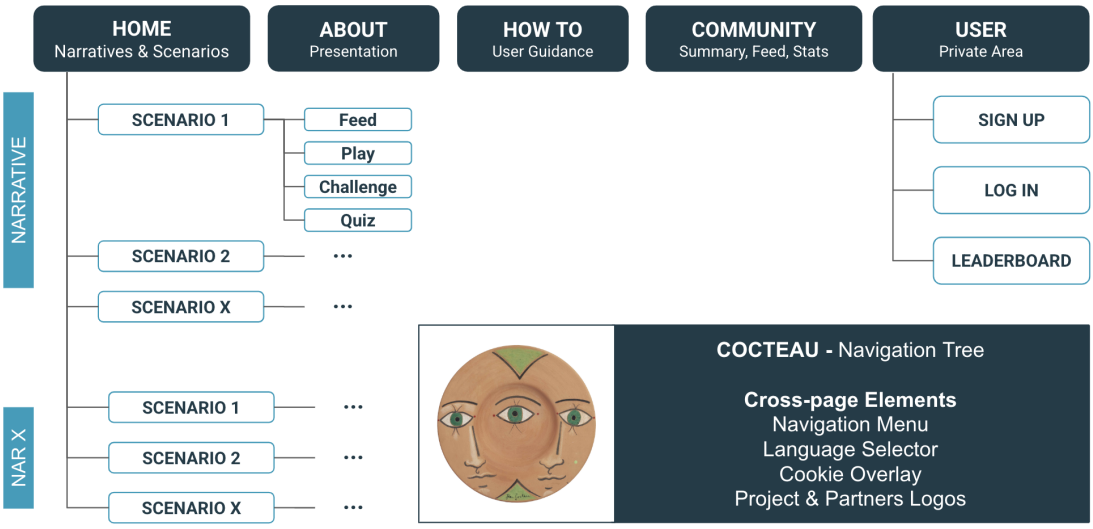


Figure 3.8: *The navigation tree for the COCTEAU platform, involving all the sections users can explore.*

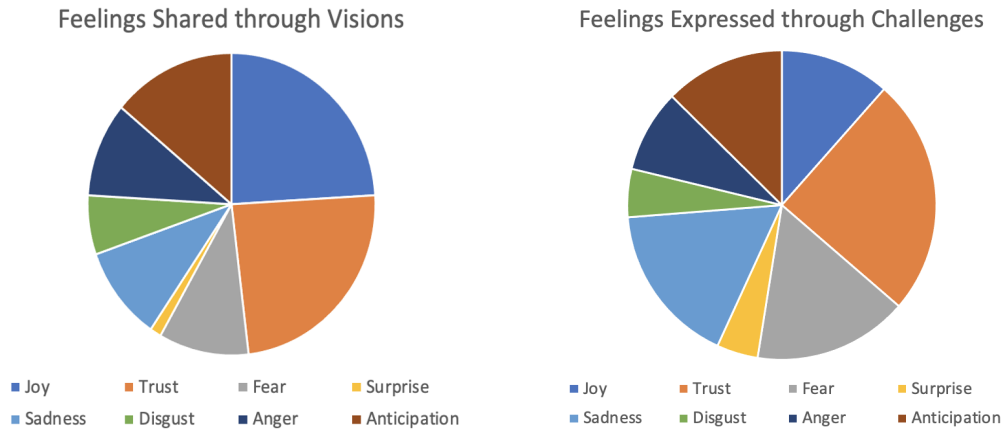


Figure 3.9: *On the left, a pie chart represents the distribution of feelings participants shared through their visions on the COCTEAU platform. On the right, a pie chart represents the distribution of feelings participants perceived when playing matches on the COCTEAU platform while participating in challenges with other players.*

A first analysis of the feelings shared on the platform through visions (Figure 3.9, on the left) displays that most people felt negative emotions, like sadness, anger, or fear. All such feelings are commonly associated with the pandemic and its consequences (*e.g.*, the lockdown measures enforced) that resulted in people experiencing loneliness, stress, and sadness. On the other hand, trust was also one of the most shared feelings, reporting that people still had faith the pandemic would have been over and trusted the solutions and changes proposed in the described scenarios on the platform. Similar to the feelings shared within the platform through visions, participants played in-depth matches, sharing their opinions and the feelings they perceived through others' visions. While most perceived feelings (Figure 3.9, on the right) are closely related to

the ones shared through visions, *i.e.*, most of the participants perceived the community mostly shared visions about trust and sadness, they perceived more visions shared fearful thoughts and ideas about the scenarios. On the other hand, some visions expressing joy were shared and were perceived as such by the community.

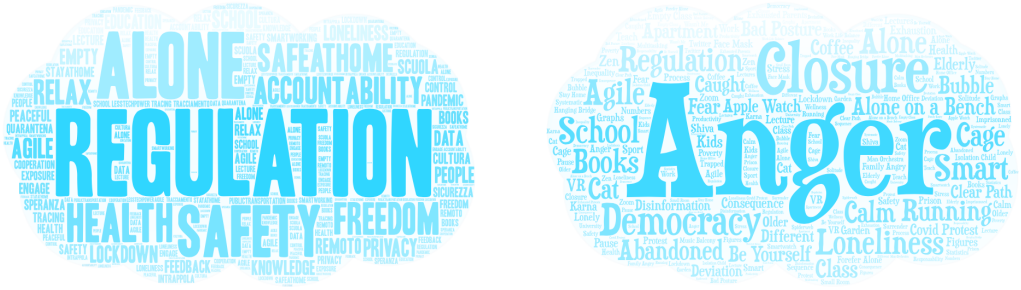


Figure 3.10: *On the left, a word cloud represents the keywords employed by participants to describe their visions. On the right, a word cloud represents the keywords employed by participants when looking for the figures to represent their visions.*

When sharing visions, participants are asked to provide three keywords to describe them. As it can be seen in Figure 3.10 on the left, while the most chosen keyword was "regulation", which is most likely associated with the need to enforce more regulations during the COVID-19 period, other keywords describe the feelings of people to stay safe at home during lockdown (*e.g.*, health, safety, and secure at home) while still highlighting the desire to escape the lockdown and the adverse effects it brought forth (e.g., alone and freedom). Participants were also asked to pick some pictures through an integrated Unsplash search engine integrated into the platform when creating a vision. Analysing the most used words shows a general trend of negativity (see Figure 3.10 on the right). For example, some of the most searched words were "anger", "closure", "loneliness", "cage", and "abandoned", highlighting the negativity associated with the lockdown enforced during the pandemic.

3.2 Main Takeaways for Driving User Involvement from COCTEAU

The development and testing of COCTEAU led to a series of fundamental conclusions that will contribute to the design of human-centred crowdsourcing applications.

The experience acquired in designing, developing, and testing the COCTEAU web application made it possible to identify important features to be considered when setting in place such systems.

- **Gamified Design**. Gamification was proven to be fundamental in engaging people while making usually tedious activities far more simple and enjoyable. When it comes to complex data collection processes (*e.g.*, the ones usually employed to collect citizen feedback), these might be even more tedious compared to the most common ones (*e.g.*, surveys involving multiple closed questions), potentially involving in-person meetings or discussions which might be complex to track. In

these cases, extracting useful data and insights might not be that easy. In the proposed application, Gamification made these activities enjoyable for users thanks to well-known design elements (*e.g.*, achievements, status, points, etc.). Another core objective of the applied gamified design is to make the data collection process simple, intuitive, and well-structured. In COCTEAU, custom processes for various kinds of data were developed to allow flawless and structured data collection, applying targeted design elements based on data complexity and further highlighting the need for ad-hoc design.

- **Feedback & Interaction**. Whenever an application with two actors benefiting each other is employed, it is essential to make it so both sides can interact. It can occur in an indirect and structured way (*e.g.*, by sharing content on the platform), or directly (*e.g.*, direct messages, forums, communities, etc.). These might also happen through other channels (*e.g.*, by applying the proposed feedback in local communities). These mechanisms allow users to get feedback about their performance and status as a reward whenever they are deemed worthy community members. On the other hand, initiators (*i.e.*, researchers, experts, etc.) collect content and thoughts about the activity and the process deployed to enhance it further. Furthermore, it is also essential to spark discussions and interactions between the platform members to keep users engaged and maintain a high activity level, fostering a dynamic and growing environment. In COCTEAU, users are shown a summary of the activities performed on the platform through a dashboard, a set of leaderboards, and personal achievements, making it possible for potential users to see the platform as a constantly growing environment. Furthermore, while feedback is provided through the platform by policy makers, interaction is mainly driven by custom gamified applications that drive structured opinion-sharing and communication.

3.3 Empathy-driven Gamified Crowdsourcing

Additional research involving crowdsourcing platforms was carried out to explore novel design patterns, considering the COVID-19 scenario in which effective data collection activities were proven challenging to develop as empathy plays a fundamental role.

The recent COVID-19 pandemic affected our lives unprecedentedly, changing our daily habits whilst disrupting our emotional and psychological health [219]. Among its consequences, the lockdown enforced by local governments affected people's lives the most, causing depression, anxiety, and stress across the population [21, 118, 129]. Although the pandemic's consequences are slowly fading, the research community's interest in understanding people's emotions over that period is still alive [436]. Improving the psychological understanding of such a complex scenario would contribute towards reducing the consequences of future occurrences and developing tailored approaches to tackle them. Furthermore, collecting accurate and complete data is essential as people's thoughts and feelings vary significantly. Researchers with various backgrounds shaped data collection processes and methodologies to involve people in sharing their feelings,

designing approaches combining Gamification with the most commonly employed survey approach [342, 448]. While questionnaires might be effective regardless of time, people are slowly starting to forget how they felt over these rough times, consequently hindering the effectiveness of such simple approaches. Hence, collecting people's feelings requires designing methods capable of sparking these emotions once again.

My Lockdown Escape is a gamified approach to collecting people's feelings during the COVID-19 pandemic. The tool makes people empathize with their past selves – a concept we call *self-empathy* – by combining various gamified techniques and design elements. The proposed design follows a hybrid approach involving a board game and a digital application. The digital application implements a storytelling-driven activity integrated with an escape room-like board game. The proposed methodology is validated through a series of experiments to assess its effectiveness in collecting people's feelings by sparking their empathy towards their past selves. Furthermore, the application's usability and the system's hybrid design are also assessed through well-known approaches The proposed methodology was implemented and shared on the COCTEAU platform.

3.3.1 Context-specific Related Works & Background

Empathy and Self-Empathy. Empathy can be described as *the capability of a human to put themselves in someone else's shoes* [170]. An extensive definition characterizes empathy as *an emotional response, dependent upon the interaction between trait capacities and state influences, whose resulting emotion is similar to one's perception (directly experienced or imagined) and understanding (cognitive empathy) of the stimulus emotion, with the recognition that the source of the emotion is not one's own* [94]. Regardless of their complexity, these definitions imply a social relation between two people, *i.e.*, the one who feels and expresses an emotion and the one who experiences the consequent emotional response. In this research, we stray away from such a standard model and focus on sparking and assessing people's empathy compared to their past selves rather than someone else, resulting in a so-called one-state model [154]. Such a perspective should drive people towards a better understanding of the emotions they experienced since they were the ones feeling them in the first place.

The research field on empathy found fertile ground in computer science [482], resulting in the development and assessment of various approaches leveraging gameful design elements to drive empathy [220, 259]. Such approaches make it necessary to highlight a fundamental difference between the empathy experienced by humans through their peers and the one conveyed through digital technologies. The first is a human reaction sparked by one's perception and understanding of the feelings of another human being through their senses. The second must leverage the features of digital technologies to spark it, relying on images, videos [134, 325], and sound [87] to convey emotions, feelings, and perceptions since digital technologies cannot convey them through the senses. Hence, it is necessary to design approaches that shorten such a gap while driving people to empathize when digital environments are employed.

Gamification. During the COVID-19 pandemic, understanding people's emotional and psychological state was fundamental to comprehending its impact. On the other hand, some researchers focused on tackling its consequences, demonstrating the effectiveness of gamified approaches in motivating and enhancing students' learning

[144, 460], approaching the elderly with healthcare initiatives [478], improving the population's awareness about disinformation [272, 413], and many more. In particular, escape room-like experiences were proven effective in springing cooperation [144] and motivation [144, 164, 431, 460] among participants, mainly when applied in remote digital environments. Similarly, digital storytelling (*e.g.*, narratives, interactive stories, etc.) was shown to improve the application's appealing [149], user engagement [289], and rising emotions and sparking imagination [69] as they engage users on a personal level in novel or familiar experiences. Despite Gamification's demonstrated effectiveness in this context, the research community acknowledged the need to apply these methods carefully (*e.g.*, to avoid biasing the user with the narrative [172], or to avoid using reward-based mechanisms in surveys [418]) to prevent undesired behaviours.

3.3.2 My Lockdown Escape

A hybrid application named *My Lockdown Escape* was developed and tested at some events, reporting its strengths and potential improvements. The application combines storytelling and escape room designs to spark self-empathy, finally collecting data about people's feelings during the pandemic.

A web-based application named *My Lockdown Escape* was designed to spark emotions and drive people to empathize with their past selves, finally moving them to describe and share the feelings they experienced during the COVID-19 pandemic. A hybrid approach was applied, *i.e.*, the proposed methodology includes both digital and physical entities, tackling the weaknesses of a fully digital design in sparking empathy whilst acknowledging its strengths (*e.g.*, ease of use, process automation, etc.). The process is implemented as a story-driven escape room experience through which players describe the room they spent their lockdown in and interact with pandemic-related items to escape the lockdown. An interactive digital application guides players throughout the activity using a storytelling-based approach that describes a series of common situations they (most likely) experienced during the pandemic. These elements drive players to describe the mental and emotional conditions experienced during the lockdown. In particular, we argue that remembering the room the player lived in during the lockdown would spark their memories and feelings. Furthermore, the similarity between the context and the elements considered in the approach's design (*i.e.*, the escape room design and the lockdown, and the storytelling-driven design and the people's experience) makes the process similar to the real experience, hence contributing to sparking emotions and stimulating self-empathy.

Game Assets. The hybrid methodology includes digital and physical assets. The web-based application is the only digital tool, hence requiring players to own a mobile phone with a camera to perform the activity. Physical assets include decks of cards and a board. Each deck is identified by a unique colour and name. Every card has a symbol printed on its front, and a unique QR code and the name of the deck it belongs to on its back. Examples of cards are depicted in Figure 3.12, on the left. Cards can be classified based on their role in the activity

- **Avatar Card**, *i.e.*, a card with a simple outline of a stylized person or a pre-made

avatar. Players can customize or pick their avatar before the beginning of the activity. The Avatar deck includes ten pre-made and one customizable card.

- **Decoration Card**, *i.e.*, a card representing furniture and items used to decorate the room. Each card is a phrase or word Players can use to complete a statement in the first part of the story. Each card is associated with one statement, and each statement is assigned multiple cards. Decoration decks allow for a variable number of cards.
- **Object Card**, *i.e.*, a card representing various pandemic-related items a person may have interacted with during the lockdown. Each card is associated with a question asked to players in the second part of the story. The Object deck includes 12 cards.
- **Container Card**, *i.e.*, a card representing furniture in which items may be stashed. Each card is associated with a question asked to players in the second part of the story. The Container deck includes six cards.

The questions, statements, and answers associated with the cards were predefined when designing the application. These can be customized based on the researchers' needs. The board includes two parts. The top part represents the lockdown room, *i.e.*, an abstraction of an actual room where players experienced the pandemic. It has a dedicated card slot for each Decoration Card deck involved (see Figure 3.13). The bottom part represents the spot (*e.g.*, a carpet) where players place the items they uncover when playing the second part of the activity (described in Figure 3.12 on the right). It includes three slots for the piles obtained from the Object and Container decks and a slot where to place the discovered items.

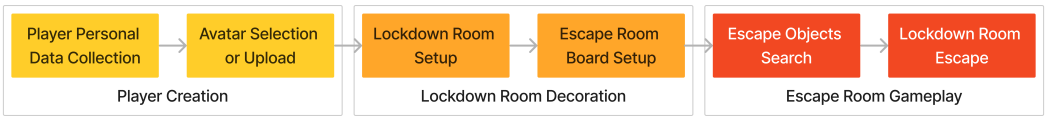


Figure 3.11: *A high-level representation of the three steps of the game and their sub-steps.*

The gamified process can be divided into three main steps (represented in Figure 3.11), each of which can be further divided into two sub-steps:

- **Player Creation**, *i.e.*, a step involving players' data collection and avatar customization.
- **Lockdown Room Decoration**, *i.e.*, a step involving the lockdown room decoration by placing Decoration Cards on the board and setting up the next step.
- **Escape Room Gameplay**, *i.e.*, a step involving escaping the lockdown room by uncovering Objects and Container Cards placed on the board.

Player Creation. Before playing the activity, players provide personal data through the digital application, *i.e.*, nickname, age, country, gender, ethnicity, and education level. They also create a physical avatar or pick a digital one. In the first case, they

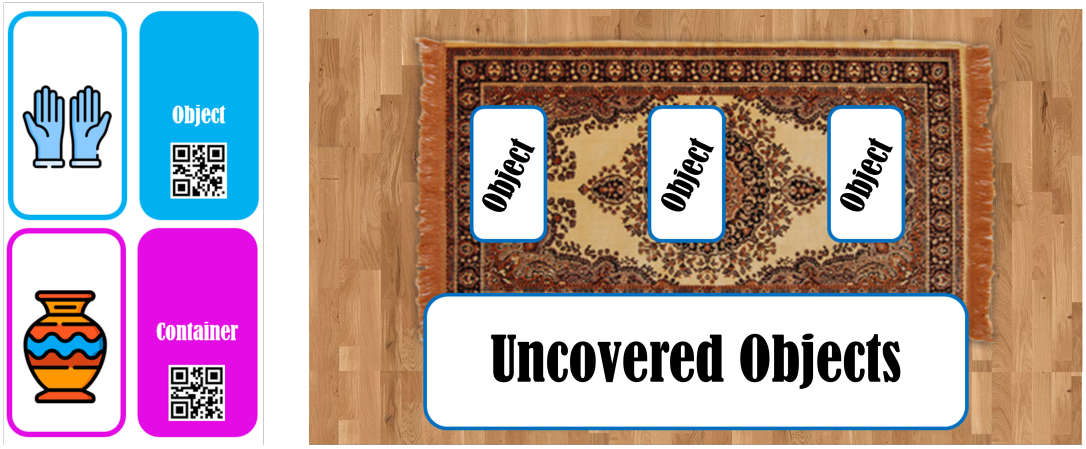


Figure 3.12: *On the left, examples of cards from the decks required for the Escape Room Gameplay step (i.e., a card from the Container deck on the left and from the Object one on the right). On the right, a representation of the part of the board that supports the Escape Room Gameplay step is provided. The three slots on top are dedicated to the three piles of cards that will be prepared from the Object and Container decks, while the cards uncovered by the player will be placed in the slot at the bottom.*

customize the Avatar Card by drawing on it using coloured markers, uploading it into the system by taking a picture, and placing it in the corresponding slot on the board. Alternatively, players pick one of the pre-made avatars through the digital application and position the corresponding pre-made Avatar Card on the board.

Lockdown Room Decoration. In the first part of the story, players decorate their lockdown room by completing statements in the story narrated through the digital application. For each statement, players inspect and pick a card of choice from the associated deck to complete the story, scan the QR code on its back using their mobile phone through the application's interface (which stores the choice in the system), and place it face-up in the corresponding slot. Whenever a statement is completed, the ongoing part of the story is updated and displayed alongside the next one. This process is repeated until a card is placed in each slot, hence having players answer all questions. An example of a statement and its list of completions are provided below.

> ***Statement:*** *Our story begins in early 2020, and not long ago, the COVID-19 pandemic broke out in your country and the whole world. You are at home watching the news. The titles are scary and doubtful. Take a look around you, and you will see that you are surrounded by ...*
>
> ***Possible Completions (Cards):*** *Family, Parents, Friends, Strangers, No one, Roommates, and Animals*

Then, players set the board for the second part of the story. They shuffle the Object deck and create three face-down piles in three dedicated slots by evenly distributing the cards. Then, one randomly selected Container card is placed atop each pile.

Escape Room Gameplay. In the second part of the story, players must find three Object cards, *i.e.*, the mask, hand sanitiser, and green pass cards to escape their lock-

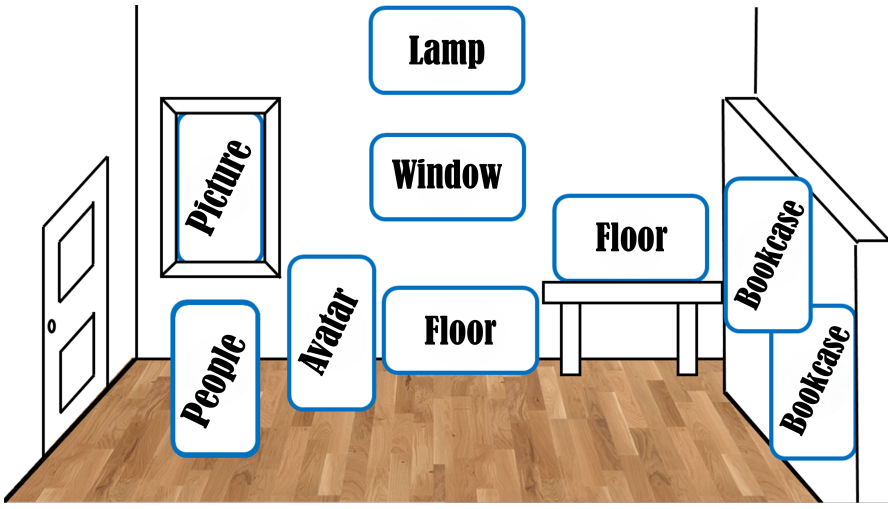


Figure 3.13: *A representation of the part of the board used in the Lockdown Room Decoration step. Each slot is assigned a name representing the deck to which the card to be placed there belongs. An avatar slot where players can place their Avatar card is also featured.*

down room, proceeding as follows. They scan the QR code on the back of the card on top of a pile of choices and answer the corresponding question on the digital application. An example of a question and the corresponding list of answers are provided below.

> ***Question (Cards):*** *Think back at your lockdown experience. If it was a movie, what title would it have?*
>
> ***Answers:*** *The Never-ending Story, The Social Network, Home Alone, Life is Beautiful, Back to the Future, Eat Pray Love, A Good Year, Cast Away*

Whenever a question is answered, players flip the card and uncover the item on its front. Then, the card is placed in the dedicated part of the board, and its corresponding digital icon is displayed in the application. The process is performed regardless of whether the item belongs to the Object or Container deck, and it is repeated until all the core Object cards are found. Then, players can escape the room or keep playing to uncover all the objects, potentially answering all the available questions. When they escape their lockdown room, players are shown their story, which can be shared with their peers. It includes their avatar, the textual description of the decorated room, and the items they uncovered.

Implementation. *My Lockdown Escape* implements physical and digital assets. The physical assets (*i.e.*, the cards and the board) were designed using digital tools, printed on cardboard, cut, and coated with plastic. The digital asset (*i.e.*, the web application) was developed as a three-layer architecture. The front end was implemented using HTML, CSS, Javascript, and Thymeleaf. The middle layer was developed using Java, Spring Boot, and the Model-View-Controller framework. The back end implements MySQL relational database. The application was deployed on a web server to make it accessible to multiple players simultaneously.

Methodology Validation. A series of experiments aimed at assessing different aspects of the proposed approach were performed. The first experiment involved 21 students and researchers (9 women and 12 men) from Politecnico di Milano, aged between 21 and 27 (26,7 on average) recruited through the university network, in individual experiments in Milan. The second one involved 28 people (17 women and 11 men) from various European organizations, mainly aged between 22 and 66 years old (28,4 years old on average), recruited through a presentation at a public event in Bruxelles. Participants were previously informed about the nature of the experience and voluntarily agreed to partake. Furthermore, they could opt-out at any time. Whilst the first experiment mainly collected feedback about the approach and the user experience, the second one contributed to testing the methodology in an open environment and collecting feedback about possible improvements. In these experiments, participants were given an initial description of the application. Then, they performed the activity without receiving any suggestions. Each participant brought their mobile phone to play, allowing testing on different mobile operating systems and web browsers.

The first experiment's participants were asked to answer a questionnaire including all the questions from the System Usability Scale (SUS) [50], *i.e.*, a questionnaire including questions on a 5-point Likert scale (ranging from 1 - "Strongly Disagree" to 5 - "Strongly Agree") to measure the system's usability [32], and nine 5-point Likert scale questions to evaluate the empathetic capabilities of participants, the tool's effectiveness in sparking self-empathy, and the hybrid approach. These were mainly inspired by the literature [194,206,331] as none of the questionnaires from the considered literature adequately addressed self-empathy and hybrid approaches. These questions are reported below.

- (Q1) I would describe myself as a pretty tender-hearted person.
- (Q2) When I think about sad past events of my life, I feel the same sadness.
- (Q3) I am often quite touched by things I see happen.
- (Q4) The game helped me empathize with my past self.
- (Q5) The game helped me remember my lockdown experience.
- (Q6) The escape-room style helped me remember my lockdown experience.
- (Q7) The storytelling style helped me remember my lockdown experience.
- (Q8) I found the hybrid method more engaging than digital-only methods.
- (Q9) I feel the hybrid method is better than full-digital or full-physical.

The collected answers are reported in Table 3.1. Moreover, we collected open feedback from participants regarding the gamified application design in both experiments.

RQ1 - Empathy. The method was assessed through a series of questions to assess participants' capability to empathize (Q1-Q3, Table 3.1) and the application's capability to make them empathize and remember their lockdown experience (Q4-Q7, Table 3.1). Overall, most participants (70%) describe themselves as capable of empathizing

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Strongly Disagree	0	1	1	0	0	0	0	0	0
Disagree	3	10	3	3	0	0	1	0	1
Neither Agree nor Disagree	8	2	2	6	3	6	4	1	1
Agree	7	8	13	9	13	10	12	11	13
Strongly Agree	3	0	2	3	5	5	4	9	6

Table 3.1: *A table representing the frequencies for each answer and each custom-made question.*

with others (Q3), even though only 50% state they are capable of feeling the same intense emotions from the past (Q2). Moreover, the method allowed most participants (60%) to empathize with their past selves (Q4). It was even more successful (85%) in making them remember their lockdown experience (Q5). The storytelling and escape room designs were proven effective towards such objectives (Q6-Q7), with the first being slightly better than the second (75% and 70%, respectively).

RQ2 - Hybrid Design. Participants were asked whether they enjoyed the hybrid system compared to a fully digital or physical experience (Q8-Q9, Table 3.1). More than 95% agreed that the hybrid design was more engaging than a potential full-digital or full-physical method (Q8). A similar statement is associated with the application design (Q9). We argue a hybrid approach is more effective as it allows better tracking of behaviours and answers in a data-driven fashion while allowing better scalability and empathy-sparking capabilities. These preliminary statements require further investigation by implementing fully digital and physical versions for better validation.

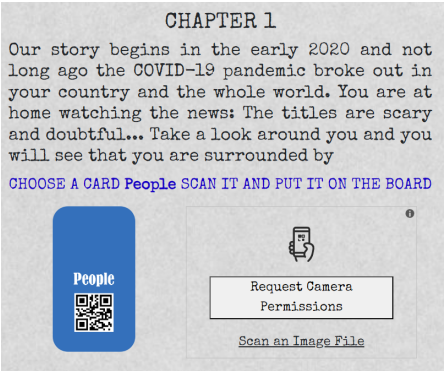


Figure 3.14: *A screenshot of the Lockdown Room Setup step.*

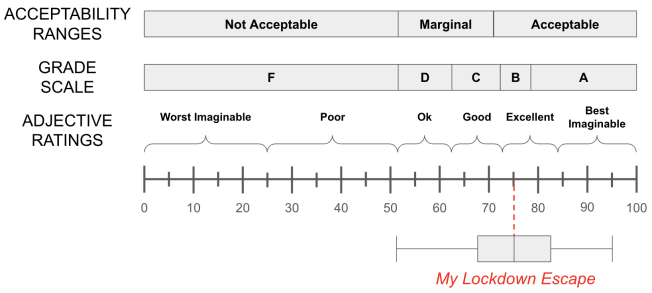


Figure 3.15: *A boxplot of the SUS scores assigned by the participants.*

RQ3 - System Usability. The first experiments contributed to validating the application's usability. The application (screenshot in Figure 3.14) achieved a median SUS score of 75, representing good usability compared to the average SUS score of 68 [374], with a first quartile of 67.5 and a third quartile of 82.5 (as displayed in Figure 3.15).

RQ4 - Gamification. Although most participants deemed the experience enjoyable and engaging, the feedback received and the behaviours observed outlined the improvements the application must undergo. First, the game's instructions may benefit from an

improved textual description and additional details. In the Lockdown Room Decoration step, some participants were misled to take their cards randomly instead of looking at and picking them. Consequently, this caused them to position their cards face-down on the board instead of face-up. Furthermore, participants preferred the Lockdown Room Decoration step, stating the Escape Room Gameplay may benefit from minor re-designs due to the randomness in finding the cards to escape their lockdown room. Additionally, most participants left the room when they met the conditions, resulting in nondeterministic data collection. Whilst it does not affect the assessment, such a design choice may hinder the amount of data collected. Future iterations of the application may require players to uncover a predefined number of items (potentially all) to leave their room to address such a flaw. Regarding the data collection, a few participants stated that the Lockdown Room Decoration step perfectly masked the data collection activity. On the other hand, they clearly felt they were completing a survey in the Escape Room Gameplay step. Such feedback highlights the need for improvements to better bind the approach with the data collection activity underneath, *e.g.*, by aligning the cards and the associated questions. Furthermore, people's motivation is fundamental to the system's usage. Filling out a form may be quite straightforward and requires no motivation other than the people's desire to contribute to their community or for research purposes. While such motivations still hold, the proposed approach introduces little intricacy as it applies Gamification to spark self-empathy and improve data quality, despite increasing the task's time complexity. Such an additional complication may hinder people's motivation towards the application, requiring further gamified expedients to keep the user motivated and engaged throughout the experience.

Limitations. While we deem the method effective overall, this research only focused on questions involving a predefined list of answers, requiring further research on its effectiveness when including open questions. Furthermore, a specific and complex scenario (*i.e.*, COVID-19) was addressed, making future research on the generalizability of the approach of fundamental interest. Validation could be extended by involving a more detailed list of questions, comparing the current implementation with a full-digital and a full-physical one, and involving a broader audience.

3.4 Main Takeaways from *My Lockdown Escape* for Driving User Involvement

Following the experience with the *My Lockdown Escape* gamified application, a series of fundamental features for the development of crowdsourcing applications were identified.

The experience acquired in designing, developing, and testing the *My Lockdown Escape* web application made it possible to identify important features to be considered when developing data collection processes.

- **Gamified Design**. Gamification was proven effective in involving people in the implemented data collection process. The proposed survey-based activity usually involves people filling out forms made of structured closed and open questions to

provide information. Such a process might take a lot of time, not only due to its potential length (considering that some questionnaires might require hours) but also because of its tiresomeness, which might discourage users from proceeding. In the *My Lockdown Escape* application, the gamified design made these activities enjoyable for users thanks to its storytelling and lockdown room design, which successfully masked (most of) the data collection process.

- **Empathy**. When it comes to specific gamified application designs, empathy might play a fundamental role in driving users and improving the outcome of the data collection process. In the design of *My Lockdown Escape*, the empathy-driven design contributed to make people concentrate on their past experiences, improving the accuracy of the collected data. Furthermore, users enjoyed the activity, demonstrating the effectiveness of empathy in driving their engagement.

3.5 EXP-Crowd: A Crowdsourcing Framework for Explainability

Most of the lessons learned from the previously described crowdsourcing applications were applied to designing an XAI- and human-centred crowdsourcing web platform to bridge the gap between humans and AI practitioners.

In the current XAI research scenario, most interactions between practitioners and the crowd are driven by a need for data from the research community. These are usually achieved through crowdsourcing platforms, awarding the crowd with monetary compensation for their work. Such an approach embraces extrinsic motivation, striving away from an intrinsic motivation-centred design. Indeed, workers are primarily motivated to perform their jobs quickly for higher remuneration, sometimes even misbehaving with the only purpose of achieving a reward. XAI crowdsourcing activities usually involve crowd workers in validating or providing knowledge as data for AI practitioners to train or validate models and explainability approaches. Activities might also use Gamification to enhance various aspects, making it more enjoyable and improving outcomes.

Regarding crowdsourcing and Gamification, the knowledge acquired from the aforementioned experiences gave us insights into the most critical elements and principles to be included in a potential crowdsourcing framework. Most of them were applied in the design of an XAI-centred crowdsourcing platform to connect AI experts, practitioners, and novices in sharing knowledge with each other. The latter will learn about Artificial Intelligence and explainability and share their expertise through crowdsourcing activities designed by AI practitioners that can potentially be shared on the platform. In this setting, intrinsic and extrinsic motivations will drive novices' engagement (as detailed in the following sections). Even though a few researchers explored the concept of empathy in AI [414], it is not strictly connected with XAI, making it more related to the design of context-specific gamified approaches or explainable empathy-driven applications [291, 383] rather than XAI itself, hence it will not be directly employed in the platform's design.

This research longs to envision an open, gamified crowdsourcing framework to

bridge the gap between the (X)AI research community and AI novices to

- share knowledge within the community, teaching novices about AI and its most recent developments while creating a network of AI practitioners to drive cooperation and share discussion on research topics of interest;
- improve the community's knowledge and understanding of AI models, their behaviour, and their explanations;
- engage the community members in providing helpful content to AI practitioners by employing intrinsic motivation factors and gamified approaches, striving away from the standard reward-centred pattern.

Ultimately, we strive to create an open community where common and context-specific knowledge might drive communication between various communities that benefit each other.

3.5.1 System Design

The main actors engaged in the proposed crowdsourcing framework fall into two categories: crowd or AI novices (from here on also referred to as users), who are involved by learning and partaking in activities, and AI practitioners, researchers, and experts who set up and share activities and knowledge about AI, ML, and explainability, acknowledged their level of understanding of these topics. Figure 3.16 provides a simple overview of the interaction flow proposed within the framework.

The following sections detail the framework and provide use cases to clarify its organization. These will mainly describe the researcher side since most of the activities described for the user side are trivial. A persona representing a researcher exploring and interacting with the final implementation of the proposed framework (*i.e.*, a web-based platform) is presented. This persona will be referred to as "Andrea".

Knowledge Assessment. As one of the main objectives of the proposed framework is to improve the capability of the crowd to understand black-box models' behaviour and explanations, educating users on (X)AI-related topics is essential. Therefore, the first step users go through is an assessment test to appraise their knowledge about AI, explainability, and similar issues. Users will be asked to complete a standard test defined by the research community with multiple-choice questions. They will be assigned a category representing their level of expertise based on the outcome. This category will describe the activities they can partake in and grant them status in the community. Users can improve their category by engaging with the proposed activities and enhancing their skills through education. On the other hand, the research community will build multiple-choice questions to be employed in the initial knowledge assessment test and the proposed activities. Each question is made of a text, a set of correct and incorrect answers, the explanations associated with each correct answer, a difficulty (*e.g.*, easy, medium, and hard), and a category (*e.g.*, AI, XAI, etc.). The community must approve questions to ensure high-quality content is provided. Therefore, each question must undergo an evaluation period during which community members can further improve it by suggesting updates, proposing new answers, and extending its explanations. At the end of this period, the question is approved if it receives enough positive evaluations.

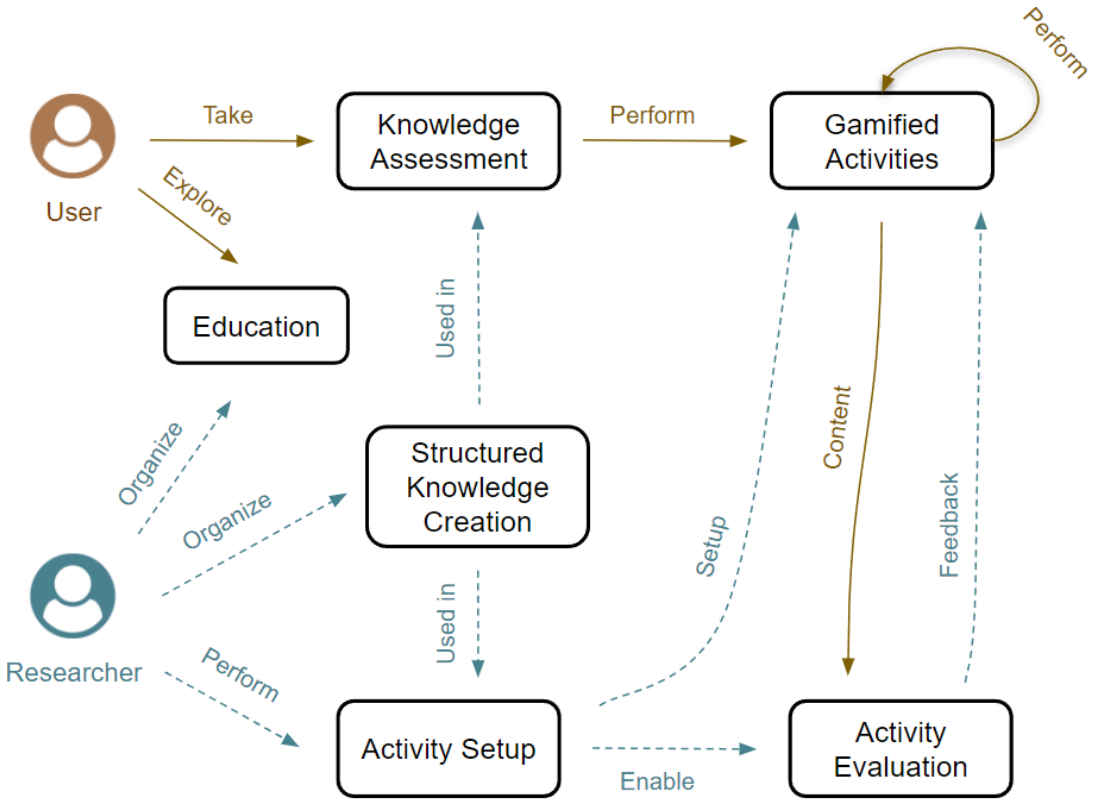


Figure 3.16: *Interaction flows of researchers (dashed cyan arrows) and users (orange plain arrows) across the activities devised within the proposed framework. Researchers organize users' knowledge and set up activities to collect data. As users engage with such activities, they provide content to researchers. In turn, researchers give users feedback about the activity they performed. Such feedback aims to improve users' understanding of the activity and the knowledge and context provided within it.*

Approved content will be openly available for the community to use as researchers may want to re-assess the users' knowledge as they engage with one of their activities. Researchers can still improve a question by providing new content after approval. Such a need makes it necessary to involve researchers and experts in advance. These can set up their activities to consider the members' knowledge level as a requirement to partake.

Use Case - Researcher. Andrea is a researcher in need of data for their research. When exploring the platform, Andrea discovers an activity that would fit their needs. Even though they would like to set it up immediately, they also want to evaluate the knowledge of the users who will perform the activity beforehand. Therefore, they explore the section dedicated to creating multiple-choice questions about AI, looking for questions that fit the context of their research. Unable to find them, they submit new questions to be assessed by the research community. A few days after their submission, they noticed some improvements were proposed (*e.g.*, new answers and explanations were provided). Andrea approves the ones they deem relevant. After a few more days, the questions are approved. Hence, Andrea sets up the activity by creating a prelimi-

nary assessment test, providing the data for the activity, and some additional parameters (*e.g.*, the number of iterations per data point, the maximum number of people that can partake, the level of knowledge each participant must have, etc.). Finally, the application is shared with the community, enabling users to join.

Education. Following the initial assessment test, users will be schooled in various ways while engaging with the framework. The following list describes how knowledge about topics of interest will be provided.

- **Preliminary Assessment Test.** Researchers may set up a small test of an arbitrary number of community-approved questions when setting up their activity. For each question, they choose its text and the list of answers. Such content would provide knowledge to users through explanations of the answers while allowing researchers to evaluate the level of education of the people playing the activity.
- **Knowledge Sharing.** Researchers can summarize, organize, and share knowledge by setting up tailored content for users to study (*i.e.*, a summary of an article, an outline of the knowledge related to a specific AI-related topic, etc.). Each publication includes a title, the topic it discusses, a summary of the content (*i.e.*, an abstract), and the textual content. Researchers can also share scientific articles for users to read if they have the rights. Only minimal information will be collected and shared to avoid copyright infringement, like its title, authors, and DOI or URL.
- **Open Discussions.** Researchers and users discuss subjects of interest in a forum-like fashion. We argue that debating with knowledgeable people would improve users' knowledge of a topic of interest. That would not only contribute to enhancing users' knowledge, but it will also be useful for the researchers to get feedback on the level of knowledge acquired by the community. Any content shared on the platform (*e.g.*, questions, paper summaries, etc.) can be referenced and discussed.

Users might comment on and report shared content whenever they deem it untrustworthy, wrong, or sensitive. Then, it is up to the platform's developers to check whether the report is reliable. If so, the content is removed from the platform.

Use Case - Researcher. Andrea would like users to understand how ML models learn so that the ones performing the activity they shared can provide better content. Therefore, they collect knowledge from scientific articles, summarize it, and share the summary in the "Education" section. Andrea achieved their first publication entitled "Understanding the way ML Models Learn: A Simplified Overview." They also provide a custom picture summarizing key concepts and a few references to the articles they used in the publication. Andrea reads an exciting article about their research topic a few days later. As it may further improve the users' knowledge, they share it by providing the necessary information.

(Gamified) Crowdsourcing Activities. Crowdsourcing activities are the core of the proposed framework as researchers seek contributions from users to get data for their research. These applications might use Gamification, acknowledging its effectiveness

in enhancing enjoyability and output quality. Moreover, the platform will push for gamified design, suggesting potential design patterns to researchers developing their applications. The steps researchers must accomplish to set up and evaluate the outcomes of an activity are summarized here.

Activity Setup. AI practitioners can pick between previously shared or predefined activities, or propose and share their implementation within the platform. Regardless, they will provide the data and the settings to set it up properly. Proposing a custom activity requires researchers to contact the platform's developers to provide and publish their implementation. The provided implementation should be compatible with the platform. Furthermore, the developers must thoroughly inspect the delivered code to avoid publishing unwanted content. This process might take a lot of effort and time from both sides, but it will benefit XAI research by providing open-source tools. Setting up an activity involves various steps, depending on the activity. Such processes generally involve a questionnaire setup step, a context setup step, and an activity setup step. In the first step, researchers decide whether to include a survey (as described in "Education") and potentially organize its questions. In the second step, researchers are asked to set up the content describing the context of their research, relevant concepts to know while carrying out the activity, etc. Finally, they must provide all the necessary material to set up the actual activity (*e.g.*, images, texts, etc.). Practitioners can include control questions (so-called attention checks) in the questionnaire and the activity to keep track of users' attention. Practitioners can also select an advised knowledge level to provide an overview of the complexity of the concepts presented within the activity or to potentially exclude users without a specific knowledge level.

Activity Assessment. On the proposed platform, users play gamified activities while researchers perform various activity-related tasks. In particular, they visualize relevant statistics about the users that partook in their activity, including the questionnaire's answers, the activity's outcome, the users' knowledge level, etc. The researcher also evaluates the users and provides feedback on their performance. In particular, they identify users who stood out, like those who answered many questions correctly (compared to their level of knowledge), those who provided high-quality content, etc. These users will be granted status-based awards that will make them distinguished community members. Similarly, users who supposedly cheated or performed badly might be reported to maintain a community of trustworthy users.

Use Case - Researcher. Andrea is finally ready to set up their activity. In the questionnaire setup step, they pick the questions (including the ones they got approved before) and their answers. In the context setup step, they provide the context of their research, describing what it consists of. Andrea also provides some of the previously shared knowledge summary content for those who did not read it. As the last step, they give the pictures, labels, and necessary content for the activity. The activity is then shared on the platform, allowing users to join. A few weeks after publishing their activity, Andrea overviews its outcomes. They noticed most users performed well while others did not. They pick the users who performed outstandingly and award them with a digital cockade. These users will be notified, and their profiles will exhibit the award. Furthermore, a few users did not pass the safety checks included by Andrea in the activity. Hence, their data is not considered, and the user might be reported.

Additional Considerations and Limitations. The platform design does not consider user engagement and profiling. Regarding engagement, sharing the platform within universities' and scientific communities' networks might strongly contribute towards involving researchers in the platform. Furthermore, extrinsic motivation might play a fundamental role in initially engaging new users. These might take the form of credits towards well-known digital platforms (*e.g.*, Google credits), provided to users after they complete a specific number of tasks. These should not be considered as the primary purpose of engaging with the platform but as a nice extra. These will either be asked as a small contribution to researchers for sharing and implementing their activities or through potential partnerships with the corresponding companies. Furthermore, it would be necessary to enforce checks (*e.g.*, personal data checks at registration) to prevent people from exploiting this system to earn free credits by creating multiple accounts and polluting the platform with bots or misbehaving users. Moreover, managing the platform and potential activities to be shared would require a lot of effort, making organizing a team of dedicated developers essential.

In conclusion, we argue that the proposed framework would drive and structure a knowledge flow between the research community and non-expert people, leading to an improvement in their level of understanding and the educational content openly available on topics of interest. Moreover, the presented crowdsourcing framework engages users on a different level than other platforms that mainly focus on extrinsic rewards (*i.e.*, monetary rewards). In particular, user education would drive user engagement and improve their awareness of AI systems and their behaviour. Such knowledge would also be enriched when a long-term engagement is achieved. Additionally, a case study on image classification and understanding is described as proof of concept of a gamified data collection activity.

3.5.2 Gamified Activity: a Case Study

A case study for a potential data collection application shared on the web platform is described, reporting its process.

In the context of image classification models, the crowd is usually engaged to describe pictures and their content. This is also true when it comes to model explanations. We assert the outcome of such labelling tasks is strictly tied to the images provided, *i.e.*, a person describing different pictures of the same entity may provide different details about the entity itself based on the provided representation. In particular, the description (regardless of its shape) might be limited to or by the features displayed in the figure. Therefore, we claim it would be possible to improve the collected features by unbinding the process from the entity's representations. In particular, the following research questions will be addressed

- **(Q1)** Is the picture displayed to the annotator causing bias when asked to describe the entity in the image?
- **(Q2)** Can we collect more features compared to standard annotation methods?

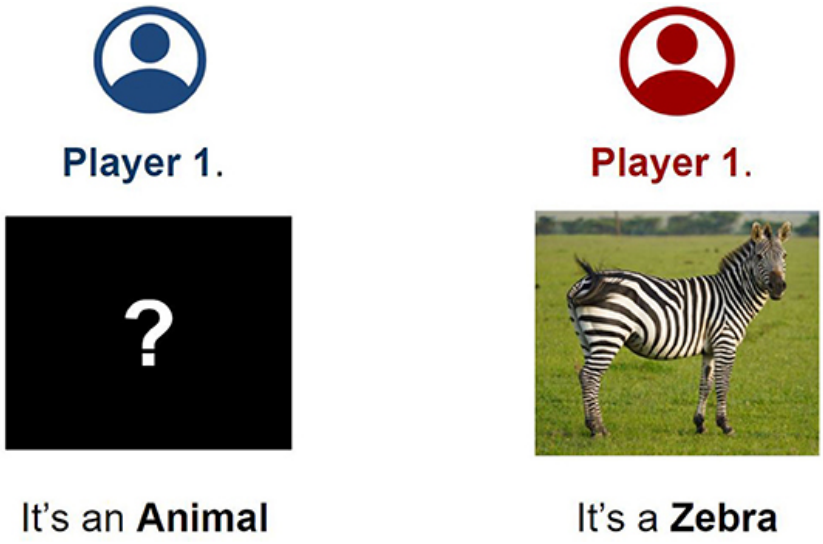


Figure 3.17: *In the **Setup Step**, Player 1 is provided with the category of the entity they have to guess (in this case, animal). Instead, Player 2 is supplied with a picture of the entity and its name (in this case, a picture of a zebra and zebra).*

The process is implemented as a Game With A Purpose (GWAP) to collect knowledge in terms of relevant features and descriptions of the considered content. In the presented use case, such features are organized into three categories, namely *Abstract* (*A*), *Not Represented* (*NR*) in the picture, and *Represented* (*R*) in the picture. *R* and *NR* features represent concrete features of the figure's content. The game involves the following steps.

- **Setup step**. *Player 1* is provided with the entity category they must guess. *Player 2* is given a representation of the entity (*i.e.*, a picture), its category and name (see Figure 3.17).
- **Basic Turn**. *Player 1* asks closed questions about the features of the entity to guess while *Player 2* answers them (see Figure 3.18 on the left). If needed, *Player 1* may fill in predefined templates (*i.e.*, "Does it have ...?," "Does it ...?," etc.). Whenever an affirmative answer is given, *Player 2* is asked to complete the Annotation Step.
- **Annotation Step**. In this step, *Player 2* is asked to perform a series of simple steps to identify the guessed feature in the picture they were provided with, if possible (see Figure 3.18 on the right). First, they are asked whether the feature is represented. If so, they are requested to outline it. Otherwise, they are asked whether the feature is abstract to properly classify the feature. Ultimately, the feature, its classification, and its potential outline are stored.
- **Hint Step**. If *Player 1* guessed no features in the last few questions, *Player 2* provides a bit of advice by providing a feature of the entity to *Player 1*. If possible, *Player 2* provides a feature that *Player 1* already tried to guess (*e.g.*, if they tried to guess something about the animal's skin, a feature associated with that aspect will be provided). Then, *Player 1* can proceed with the activity. *Player 2* is still

required to carry out the Annotation Step for the hinted feature, as it will still be considered in the final set of features.

- **Game Conclusion**. After *Player 1* has collected enough clues on the entity they are trying to guess, they can provide their final answer. If the answer is correct, the game is over; otherwise, the game moves on. When the game ends, *Player 1* is shown the original picture and the features outlined to check that *Player 2* performed their task correctly. If any element has been improperly outlined or any question has been incorrectly answered, *Player 1* can provide their version (*i.e.*, answer and annotation). Such an action generates a conflict the researcher will resolve when the activity outcomes are provided.

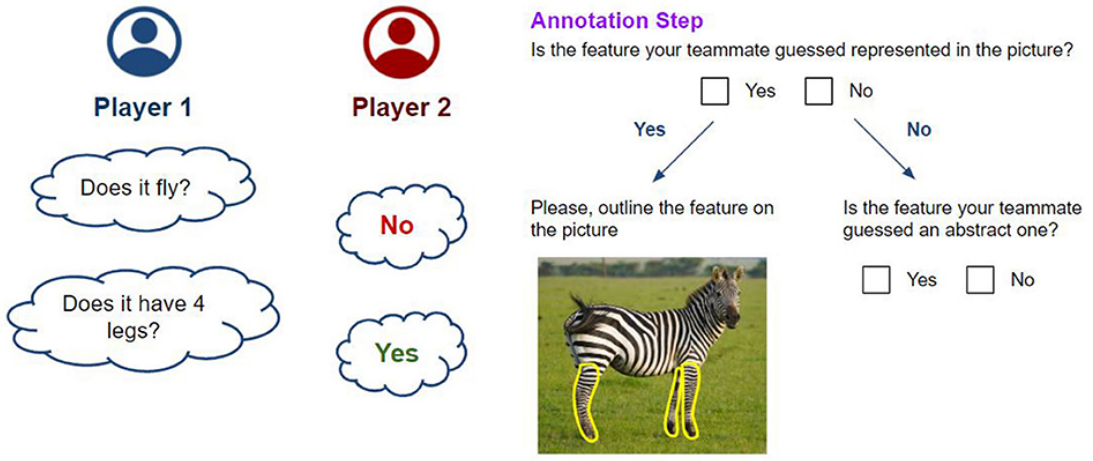


Figure 3.18: *On the left, the **Basic Turn** of the gamified activity is displayed. Player 1 asks closed questions about the entity and Player 2 answers such questions. On the right, the **Annotation Step** is summarized. Player 2 performs simple tasks to classify the guessed feature by answering questions and potentially annotating the picture.*

This activity can be customized so that players focus on specific types of features, *e.g.*, concrete features, abstract features, or both. Moreover, it could be extended by applying some changes and enhancing various steps, in particular:

- It would be possible to introduce an additional step at the end of the activity in which Player 2 provides more pictures of the same entity and outlines the features classified as *NR* on the new image, providing more data.
- It would also be possible to introduce an additional Annotation Step for Player 1 at the end of the game to improve the reliability of the results, allowing the comparison of both players' annotations to identify inconsistencies in the provided outcomes.

3.5.3 Preliminary Validation

The process was tested by involving participants organized in groups to cover three main feature types (*i.e.*, abstract, not represented, and represented)

This section describes a preliminary study on the proposed gamified approach's effectiveness and impact. The experiments were performed by selecting one category and asking participants to play the activity with some figures. In particular, the "animal" category was chosen. Images partially illustrating the body of crocodiles (*i.e.*, only its head was visible) and figures depicting the entire body of the parrot were collected. We engaged 30 people aged between 24 and 30 (with a mean of 26.4 and variance of 4.04), mostly (60%) employed in IT-related sectors. Most of them (75%) achieved an educational level superior or equal to a bachelor's degree. Participants were organized into three groups.

- **Standard Annotation Method**. This group was prompted to outline features on images using a standard approach. This group involved six people and will be referred to as *SAM*.
- **Gamified Activity (R and NR)**. This group was prompted to ask questions on concrete features (*i.e.*, *R* and *NR* features). This group involved 12 people and will be referred to as *GAR*.
- **Gamified Activity (A, R, and NR)**. This group was prompted to ask questions about any feature. This group involved 12 people and will be referred to as *GAG*.

Each person was provided with a document describing the activity based on the group they were assigned to. The members of each gamified activity group were organized in pairs, thus forming six pairs per group. Each player or couple of players was given one picture to play with. Players were asked to observe the procedure described in Section 3.5.2, depending on their assigned role and group. Each pair member alternately played both roles. Each group carried 12 matches (*i.e.*, six matches per picture). Additionally, participants were asked to keep track of questions and answers when playing as *Player 1* and the suggestions they provided as *Player 2*. On the other hand, each of the six members of the Standard Annotation Method group was given two documents containing the chosen figures. They were appointed to describe the represented animal by providing a clear and short description of their features, its potential outline on the image, and its category.

3.5.4 Results and Discussion

The collected data was analyzed to discuss the considered research questions.

Following the preliminary experiment, the research questions are answered through the outcomes and the feedback collected.

Concerning **(Q1)**, which aims at assessing whether pictures generate bias in the player describing the displayed object, most of the *R* features reported by each *SAM* group member were represented in the picture, 73% for the crocodile and 97% for the parrot (Table 3.2). Within the same group, a clear tendency to report *R* features first and forget about features not represented within the picture was identified. Indeed, 50% of the participants provided no *NR* features for the image partially representing a crocodile. These observations are aligned with our initial thoughts and expectations, *i.e.*, when a person is asked to describe an entity, they mainly attain to the particular representation provided rather than the actual entity, even when it is well-known.

[SAM] Standard Annotation Method			
Picture	*R* Features	*NR* Features	*A* Features
Crocodile	$3,67 \pm 0,51$	$1,33 \pm 1,63$	$1,5 \pm 1,38$
Parrot	5 ± 2	$0,17 \pm 0,48$	$2,17 \pm 1,72$
[GAR] Gamified Activity (R and NR)			
Picture	*R* Features	*NR* Features	*A* Features
Crocodile	$3,83 \pm 1,94$	$2 \pm 0,63$	$0,17 \pm 0,41$
Parrot	$6 \pm 0,89$	0 ± 0	$0,83 \pm 0,41$
[GAG] Gamified Activity (A, R, and NR)			
Picture	*R* Features	*NR* Features	*A* Features
Crocodile	$0,5 \pm 0,55$	$1,17 \pm 0,75$	$3,33 \pm 0,81$
Parrot	$1,5 \pm 0,55$	0 ± 0	$2,67 \pm 1,51$

Table 3.2: *The average and the sample m.s.e. per participant for each feature type and each picture considered in the described experiment*

Moreover, a significant difference in the ratio between the amount of *NR* and *R* features collected was observed for the crocodile picture across different experiments. In particular, the *NR* feature ratio grew from 27% in the *SAM* group to 34% in the *GAR* group. Such a difference is even more emphasized in the *GAG* group. An increase in the amount of *NR* features collected was also identified with a 50% increase comparing the *SAM* and the *GAR* groups. In conclusion, this provides initial proof that creating a sharp separation of roles and hiding the picture from the gamified activity reduces the bias it might induce.

Regarding **(Q2)**, the proposed methodology can collect more features w.r.t. a standard annotation method. Indeed, participants in the *GAR* group collected 20% more *R* features for the parrot picture and 33% more *NR* features for the crocodile one, compared with the *SAM* group. When analyzing the outcomes of the *GAG* group, a clear tendency to ask questions about abstract features was identified (*e.g.*, "Is it carnivorous?," "Is it oviparous?," "Does it live in the Jungle?," "Is it able to speak?," etc.), resulting in a 55% increase in *A* features collected concerning the *SAM* group. Such behaviour might be strictly related to humans' capability to abstract concrete concepts and distinguish similar entities through peculiar, selective, and (sometimes) abstract features. Questions on such selective features even played a fundamental role in the *GAR* group. Indeed, most people who had already collected a lot of concrete features expressed the need to ask a few abstract questions to consolidate and finalize their guesses. Furthermore, several descriptive dimensions, *e.g.*, the features' selectivity and the entity's category, affect such behaviours.

Participants' feedback was also collected to improve the gamified activity. Among the collected suggestions, the following ones might be meaningful to apply.

- *Player 2* will not provide annotations for the collected features during the activity

but only at the end. Such a change would smooth the process, making it quicker and more enjoyable.

- At the end of the activity, both players will perform the Annotation Step, improving the consistency of the results and the amount of data collected. Additionally, both players will be shown the picture of the entity to further enrich the collection of features by describing those represented in the image.

In conclusion, this use case described a methodology that enhances gamified visual annotation and labelling methods, mitigating the bias caused by pictures by unbinding them from the activity and allowing an even more complete collection of features. This activity can collect data about what an image classification model should have learned about the entity (*R* and *NR* features). Such knowledge can be compared with the outcomes of other explainability methods to evaluate the difference between what the model has learned and what it should have learned. Such a comparison can be carried out for models learning from pictures of the entity by comparing the heat maps derived from the model and the annotated *R* (and optionally *NR*) features and textual descriptions of the entity by comparing the outcomes of saliency-based analyses and the collected features. Moreover, the collected knowledge could be further combined to enhance the outcomes of non-textual, local explainability methods or improve textual descriptions of textual ones. In particular, non-abstract features annotated by the crowd would be helpful to describe pictures in which the same feature is detected by other methods (*e.g.*, heat maps, etc.). At the same time, abstract details would be helpful to complete textual descriptions, making them more human-understandable and human-like, or check the knowledge from complex state-of-the-art Generative AI models.

CHAPTER 4

Human Knowledge for Explainability and Robustness of the ML Pipeline

This chapter discusses the research describing human knowledge and involvement in the ML and XAI pipeline through a state-of-the-art analysis. Furthermore, a human-centred perspective on robustness is presented. This chapter is mainly built upon the articles.

1. Andrea Tocchetti, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie Yang. 2024. A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities. ACM Comput. Surv. https://doi.org/10.1145/3665926

2. Tocchetti, A.; Brambilla, M. The Role of Human Knowledge in Explainable AI. Data 2022, 7, 93. https://doi.org/10.3390/data7070093

The PhD candidate contributed to defining and performing the data collection methodologies, collecting, filtering, and organizing the considered papers. Furthermore, they redacted the content, debating the role of human knowledge in XAI (1) while contributing to the content by discussing human perspectives concerning robustness in AI (2) in cooperation with their colleagues from TU Delft.

4.1 Introduction

As humans and AI models represent two sides of the same coin in XAI, the importance of building trustworthy, robust, and human-understandable AI systems is reported.

Acknowledged the potential of combining human and non-human skills in real-life scenarios, there are still several aspects to be addressed, *e.g.*, managing potential trade-offs and ways to validate and deem AI systems accountable [130], as well as many aspects associated with the concept of Trustworthy AI [35, 54, 197]. As AI is more than ever applied in safety-critical areas (*e.g.*, self-driving cars [347]), it is essential to develop reliable systems. Towards this objective, one of the most critical properties of Trustworthy AI is Robustness, defied as *the insensitivity of a model's performance to miscalculations of its parameters* [309,513]. Recently developed robust approaches encompass the whole ML pipeline, from data collection and extraction to model training and prediction [485], and have been applied in a wide variety of contexts (*e.g.*, Computer Vision [76] and Natural Language Processing [228]). Considering the growing importance of such a topic, an analysis of the state-of-the-art and the current suggested solutions is proposed, highlighting research gaps and future research directions. A human-centred perspective is applied, focusing but not limiting the scope to human-driven approaches while emphasizing the potential of human-in-the-loop approaches in improving AI robustness.

While (human-in-the-loop) robustness is essential towards achieving Trustworthy AI, a much broader perspective on the role of humans in the field of Explainable AI is also of fundamental interest. Recently, the XAI community has put a lot of effort into understanding practices for employing data to enhance AI models and their explainability. Several researchers collected and organized articles describing XAI methods, definitions, and much more, while a few focused on approaches applying crowdsourcing and human knowledge. Acknowledged the importance of engaging humans in Explainable AI, it is fundamental to understand how they have been involved. Hence, this section organizes state-of-the-art research on the role and contribution of human knowledge in the context of the explainability of ML models and explainable AI. In particular, methods for collecting and employing human knowledge to create, improve, and evaluate the explainability of black-box models in AI are considered.

4.2 A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities

As model robustness is one of the most important features of Trustworthy AI, several research has been performed. The latter is collected, organized, and discussed. Finally, a human-centred perspective is given.

Research Methodology. Key definitions in robustness and robust AI [67, 160, 341] in the context of Computer Science were inspected, organized, and further enhanced,

including aspects related to Trustworthy AI and Human-centeredness (*e.g.*, involving human knowledge). The final list of keywords is available in Table 4.1

Group Name	Keywords
Fundamental	Robustness, Robust
Scope	Artificial Intelligence, Machine Learning, Neural Network
Context	Explainability, Explanation, Human Computation, Design, Adversarial, Transparency, Unknowns, Interpretable, Reasoning, Human Knowledge, Confidence, Stability, Resilience, Accuracy, Reliability, Interpretability, Accountability, Noise, Reproducibility, Trustworthy, Performance, Knowledge, Knowledge Elicitation, Knowledge Base, Human Interpretation, Human-in-the-loop

Table 4.1: *The three groups of keywords considered in the data collection process and the corresponding keywords.*

Multiple bibliographical databases were queried by generating all the triples of keywords combining one keyword from each group, *e.g.*, "Robustness" AND "Artificial Intelligence" AND "Explainability", finally generating 156 unique search queries. Articles were collected in July 2022 through the Publish or Perish[1] software by querying Google Scholar, Scopus, Semantic Scholar, and Web of Science and focusing on the literature between January 2012 and July 2022. The data collection returned about 100,000 papers: 31,000 from Google Scholar, 18,450 from Scopus, 30,800 from Semantic Scholar, and 19,400 from Web of Science. The collected papers underwent data cleaning. Duplicates (*i.e.*, papers with the same title and author) and papers published outside the period of interest were removed, resulting in 35'800 unique articles. Papers were then manually inspected to exclude the ones whose context or content require domain-specific knowledge (*e.g.*, healthcare and medicine), or whose notion of robustness is not associated with ML (*e.g.*, signal processing). Furthermore, the inspected articles were labelled based on their content, finally excluding the ones with the least frequent keywords (*i.e.*, appearing only once). Throughout this process, only significant or late-breaking articles were kept. These include 94.1% papers published in peer-reviewed venues, 1.9% non-archived peer-reviewed papers (*i.e.*, accepted in workshops with no proceedings and published on arXiv), and 4% non-peer-reviewed papers (*i.e.*, only published on arXiv). Further conditions about the number of citations of non-peer-reviewed papers were enforced, keeping only those with at least 50 citations if published before or in 2019 and at least 15 citations if written after 2019. In the end, 560 were systematically analyzed, 370 of which were systematically summarized and discussed. The list of collected, filtered, and analyzed papers can be found on GitHub[2].

4.2.1 Main Concepts surrounding Robustness

The collected articles revealed a heterogeneous scenario about the term Robustness, showing the term is ill-defined. Indeed, various and different viewpoints were observed (the main ones are discussed in subsection 4.2.2), and a wide variety of robustness-related concepts (see Figure 4.1) are employed in the literature.

[1]Harzing, A.W. (2007) Publish or Perish, available from https://harzing.com/resources/publish-or-perish, accessed on 25 November 2024)

[2]Inspected Articles - https://github.com/AndreaTocchetti/ACMReviewPaperPolimiDelft.git (Last Accessed 25 November 2024)

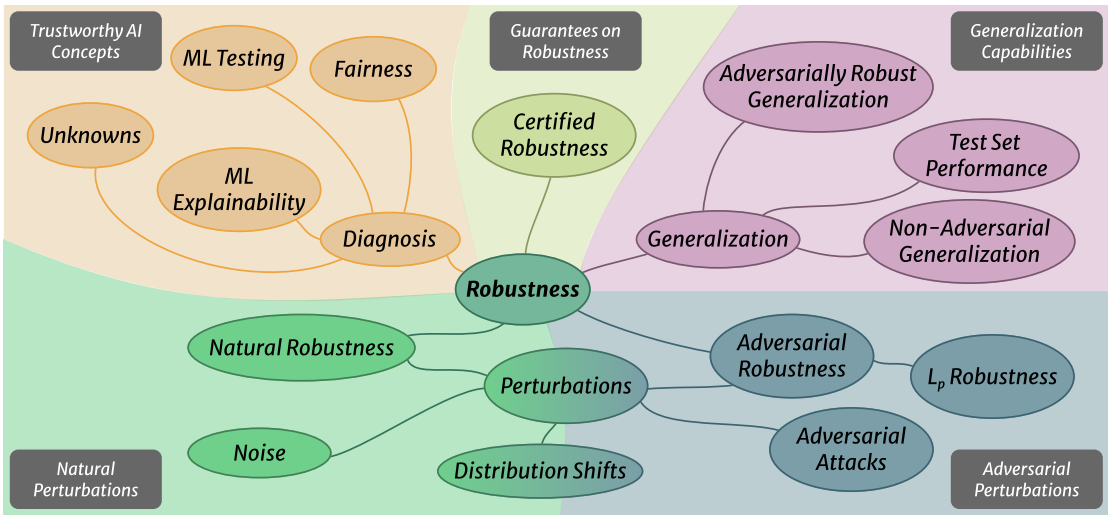


Figure 4.1: *Main concepts found through our analysis of the literature on Robust AI.*

4.2.2 Robustness Definition

> As different kinds of robustness are described throughout the literature (*i.e.*, against adversarial or natural perturbations), these are reported and described.

Given the broadness of the inspected robustness-related literature, a common ground about its definition is provided. Robustness is generally described as *the insensitivity of a model's performance to miscalculations of its parameters* [309, 513]. Inspecting the literature revealed two main branches of robustness: robustness to adversarial attacks (or perturbations) and robustness to natural perturbation.

Robustness to Adversarial Perturbations. The capability of a model to maintain its performance under potential adversarial perturbations is referred to as Adversarial Robustness. Adversarial perturbations are imperceptible, non-random modifications of a model's input to change its prediction to maximize its error [429]. Such a process generates adversarial examples, *i.e.*, an input x' close to a valid input x according to some distance metric, with different outputs [64]. The generated data might be employed to perform adversarial attacks, finding any x' according to a given maximum attack distance [76]. Adversarial attacks can be classified as targeted or untargeted [75], and white-, grey-, or black-box [292]. In untargeted attacks, adversarial examples are generated to cause misclassifications, while targeted attacks only cause misclassifications for specific classes. The main difference between white-, grey-, and black-box adversarial attacks is the attacker's knowledge about the model or the defence mechanism.

Robustness to Natural Perturbations. The capability of a model to preserve its performance under potential naturally-induced image corruption or alterations is referred to as Natural Robustness [108]. Natural perturbations (also referred to as common corruptions [176] or degradations [148]) represent conditions more likely occurring in real-world scenarios, *e.g.*, natural noise [176, 468]. In the same perturbation family,

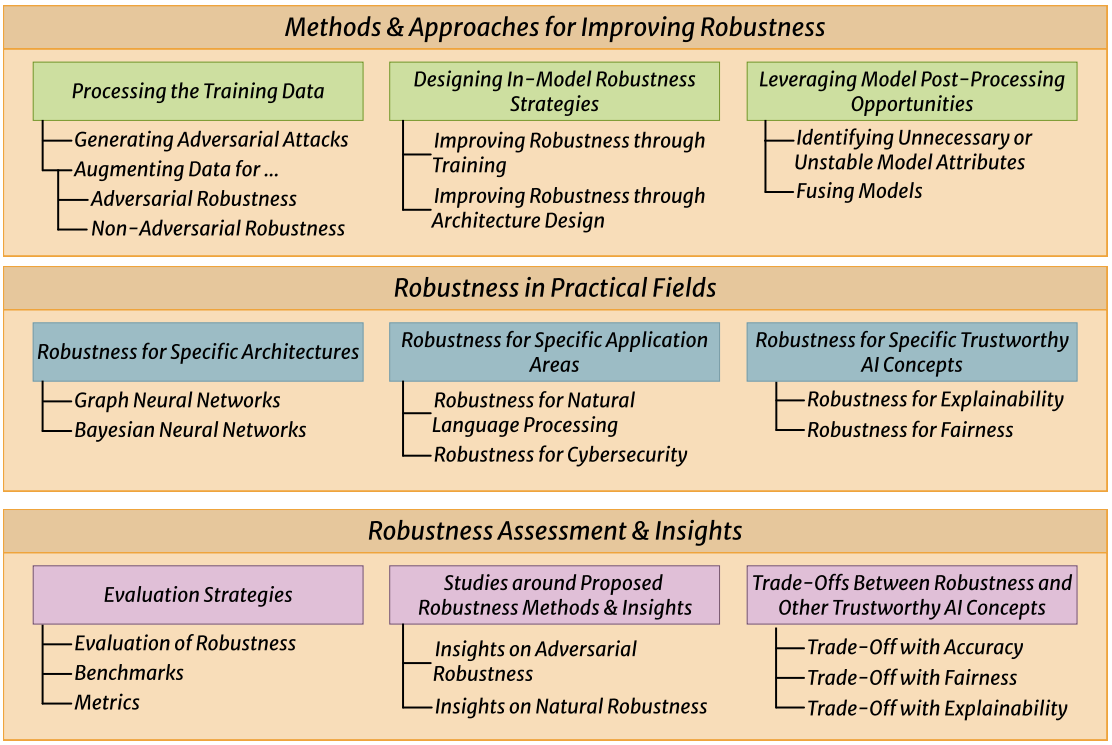


Figure 4.2: *The three identified themes and their sub-categories.*

temporal perturbations hinder the capability of a model to detect objects in perceptually similar, nearby frames in videos [385]. Natural perturbations cause the test set distribution to differ from the training set one [218] (so-called distribution-shift [104, 438], Out-of-Distribution data (OOD) [153, 391], and data outside the training set [329]).

4.2.3 Robustness-related Themes

The inspected literature was organized into three main categories, *i.e.*, methods and approaches for improving robustness, robustness in practical fields, and robustness assessment and insights.

A thematic analysis approach [48] identified three primary themes, further specialized into three recurring categories each (see Figure 4.2).

Methods and Approaches for Improving Robustness. This category includes all the methods to achieve robustness based on the stage of the ML pipeline to which they apply, *i.e.*, training set pre-processing, model creation, and trained model post-processing. The identified approaches were further organized based on the type of robustness (*e.g.*, adversarial and natural perturbations) and the ML components they apply to (*e.g.*, training procedure or model architecture). Sub-categories based on the transformations (*e.g.*, loss functions or regularizers) applied to the ML component were identified, and the main similarities and differences were described.

Robustness in Practical Fields. This category includes papers worth investigating and organized based on their research field. For each category, further sub-categories were defined, *i.e.*, based on model type (*i.e.*, Graph Neural Networks and Bayesian Learning), application area (*i.e.*, Natural Language Processing and Cybersecurity), as well as the specific concept in the Trustworthy AI domain they address (*i.e.*, Explainability and Fairness). The latter has different objectives than other works as it investigates the evolution of fairness and explanations under the effects of perturbations rather than its performance.

Robustness Assessment & Insights. This category involves articles covering robustness assessment, including but not limited to methodologies, benchmarks, and metrics. A broad range of processes was identified in the considered literature, regardless of their objective (*e.g.*, defining an assessment process or validating a robustness approach). Furthermore, researchers also performed studies to assess existing robustness methods, further collecting insights to define the optimal performing conditions. Finally, many articles propose or assess robustness methods to tackle trade-offs in model performance or related to trustworthy AI concepts.

4.2.4 Methods and Approaches for Improving Robustness

The literature on enhancing model robustness was further organized based on the type of approach, *i.e.*, through data processing, in-model robustness strategies, and model post-processing.

The collected literature focuses on improving model robustness across their lifecycle by developing various approaches, *i.e.*, data augmentation techniques to include malicious samples in the training set, ad-hoc training procedures and architectures, post-training pruning and model fusion methods.

Training Data Processing. Data augmentation approaches were developed to generate perturbation and enhance model robustness. The first path includes techniques generating adversarial attacks, varying based on their objectives, *i.e.*, the type of task of interest (*e.g.*, NLP [74, 202], image classification [64], or object detection [77]), the type of attack constraint (*e.g.*, on the physical space before capturing the data sample [468], general or component-specific attack [74], attacks preserving some input properties [202], etc.), and the brittleness type of interest (*e.g.*, only the prediction is incorrect or its explanations are too [481]). Furthermore, these approaches involve different objective functions [64, 77] or methods to generate adversarial data points (*e.g.*, GANs [1] or rule-based approaches [74, 202]).

Secondly, data augmentation approaches involving data transformations (*e.g.*, rotation [440], background-removal [454]) or generation [8,78,236,426,466] and ready-to-use data [98] are employed to enhance model robustness [71,98,236,426,440,466,528] and adversarial accuracy [8], sometimes even reducing time costs [71] and adversarial attacks success rate [41]. GAN-based techniques [1, 139, 426, 466] are broadly employed to generate adversarial samples [1], perturbations [466], and boundary data

points [426]. Additionally, ad-hoc training procedures might be needed to select [78] and adapt [71] to the optimal data to achieve model robustness.

At last, not all approaches focus on adversarial robustness. Indeed, approaches tackling noise [340], non-adversarial perturbations [139, 236, 307, 528], spurious correlations [71, 474], and distribution shifts [329, 516] were also studied. Human rationale was demonstrated effective in generating new datasets [329] and counterfactual-augmented data [71] and in defining perturbation levels [307] to improve performance [71], model robustness [307], and distributional shift robustness [329].

Designing In-model Robustness Strategies. In such a context, adversarial training was revealed as the most employed approach to achieve model robustness [163, 267, 422, 439, 473], mainly by borrowing ML paradigms like self-supervised and unsupervised techniques [303, 452]. These methodologies complement natural data with perturbed ones to allow models to incorporate information better representing real scenarios, finally driving the model to learn robust features. Projected Gradient Descent (PGD) [267] is a well-known white-box algorithm for adversarial training for which various extensions were also provided [163, 439]. Furthermore, dynamic perturbations [81,266] and human-inspired attention mechanisms [534] were also proven effective. Approaches leveraging outputs and misclassifications were also proposed, mainly focusing on rejecting low-confidence predictions [422,473] and tuning labels [80,166]. While adversarial training was proven to be the de facto standard to achieve model robustness, other methods to learn robust features and train models through regularization [70, 247], adapted regularizers [25, 221, 498], and loss functions were developed too. The latter include triplet loss [274], minimizing the distance between true and false labels [250], mutual information [489], consistency across data augmentation strategy [433], perturbation regularizers [492], adding maximal class separation constraints [296], and combining [216] or approximating [124] loss functions. Besides adjusting such functions, alternative training procedures were also proposed [249, 282, 326, 363, 416, 531] by combining several techniques (*e.g.*, distillation [53, 178]).

Model robustness can also be achieved by making a network architecture more robust through layer tweaking or employing Spiking Neural Networks (SNN) or Neural Architecture Search (NAS). Concerning the first, most of the work was performed in the context of Computer Vision applications, developing methods that take advantage of the noise injected in malicious inputs to adjust the network's layers [203, 290, 486]. A similar approach was applied to generative models [207]. Concerning noise-based attacks, SNNs [265] were proven inherently robust against them because of their functioning (*i.e.*, neurons transmit information only after they surpass a threshold) [82,388]. Concerning network crafting, NAS was applied to discover architectures with the highest accuracy in specific use cases [100, 251]. A combination of NAS and regularization was employed to drive better search performance [106, 186, 288].

Leveraging Model Post-Processing. Post-model training approaches might also be helpful towards achieving robustness. These mainly include pruning-inspired approaches to replace unstable neurons [74] or remove unnecessary classification features [138] that might lead to adversarial attacks.

Model fusion is another post-processing approach that might contribute towards

model robustness by combining additional models into a trained one to identify and deal with problematic data instances (*e.g.*, out-of-distribution, mistaken [339], noisy [362], or adversarially modified [503] samples). GAN-based approaches were the most promising ones, dealing with input data [426] and models [83]. Moreover, these approaches can also be applied to infected models to compare their robustness against compression techniques [480].

4.2.5 Robustness in Practical Fields

> The literature describing robustness when specific architectures and tasks are set in place was further organized into architecture-, domain-, and trustworthy AI-specific approaches.

Researchers also focused on improving robustness when specific architectures, tasks, and systems are applied. Architecture- and domain-specific approaches were identified. Furthermore, methods bridging the gap with Fairness and Explainability of non-functional requirements were also the subject of study.

Architecture-specific Robustness Approaches. Research trends involving specific architectures were identified in the literature, mainly related to Graph Neural Networks (GNNs) and Bayesian Neural Networks (BNNs). Given their susceptibility to adversarial perturbations, GNNs were analyzed under several aspects, like link prediction [330], controllability and connectivity robustness [260], as well as inspecting the effect of noise [131] and potential defence mechanisms [142]. Approaches for GNNs robustness certification [44,464] and designing inherently robust networks [79,204,515] were also developed to guarantee and enhance model robustness, respectively. As many adversarial attacks focus on identifying directions of high variability, which in turn is mainly linked to prediction uncertainty, BNNs are of fundamental interest in robustness. In particular, researchers demonstrated the effectiveness of such networks to adversarial attacks [63], especially when specific approaches are applied [280,453].

Domain-specific Robustness Approaches. Research on robustness spans various application areas. A clear trend can be identified among those in Natural Language Processing and Cybersecurity. NLP systems have been actively researched over the last years to develop approaches to enhance their robustness. Weaknesses of various granularity were addressed, *i.e.*, sentence-level [527], word-level [110, 499]. Misspelling- [339] and noise-based attacks [74,248,530] were also studied to reduce their effectiveness on NLP models. Acknowledging the increasing interest in AI in Cybersecurity, it is essential to achieve high robustness in systems to resist intelligent attacks. Robustness against malware was achieved through data augmentation [2] and studying the properties of operating systems and platforms to pick the most robust ones [19]. Enhancing robustness against Distributed Denial of Service (DDoS) was also of interest [1, 14], using specific approaches to generate attacks [1] and train the anomaly detection network [14].

Trustworthy AI-specific Concept Robustness. Robustness has also been discussed

alongside Explainability (mainly in the context of counterfactual explanations) and Fairness. Concerning the Explainability context [105, 232], it has been found that explainers are as fragile as the models they strive to describe [402], revealing a fundamental weakness. Hence, several approaches have been proposed to discover which inputs can be consistently explained [517], assess the quality of explanations concerning robustness [299], and evaluate explainability approaches [11,24]. When addressing counterfactual explanations' robustness, researchers focused on improving robustness [462], exploring [327] and generating [26, 474] such explanations. Another important topic is the relation between fairness and robustness and how they contribute to each other. Researchers achieved robust fairness [354] and bounds on fairness violations [472], developed adversarial approaches through fairness using a robust approach [507], and studied the connection between counterfactual fairness and graph stability [6].

4.2.6 Robustness Assessment & Insights

The literature describes methods for evaluating model robustness mainly through procedures or benchmarks. Trade-offs between robustness and trustworthy AI are discussed as well.

While implementing methods to improve model robustness has been a core research topic, researchers also developed assessment procedures, benchmarks, and studies to assess model robustness. Such approaches and the most relevant findings obtained by intersecting Robustness, Fairness, and Explainability in the context of interest are discussed here.

Evaluation Strategies. Most methodologies were found to either compute a safe radius [228,364] or region [157] where the model performs robustly, or a complementary error region [519]. A first approach to robustness evaluation is made by combining Abstract Interpretation (*i.e.*, a theory which dictates how to obtain sound, computable, and precise finite approximations of potentially infinite sets of behaviours [141]) with constraint-solving [496] or importance sampling [273]. Other formulations were also considered, *e.g.*, as a mixed integer linear program [445], or as a proportion of the inputs for which an adversarial property is not satisfied [476]. Such a robustness re-framing broadens the solutions to assess robustness, improving their scalability [445,476], computational performance [445, 502], as well as enabling pre-existing tools [177]. Certified robustness evaluation was also broadly studied [31, 109, 199, 241, 245, 398, 400, 514,520] as researchers focused on computing robustness bounds [109,241,520] while improving the training process to achieve certifiable [520] or ready-to-certify [199] networks. Similarly, L_1 and L_2 robustness were computed using deterministic [245] and random [111] smoothing. Instead, certifiable bounds computation was enhanced through overapproximation [398], orthogonalization relaxation [400], and regularization [199].

Comprehensive benchmarks were also proposed. In Computer Vision, robustness against various types of adversarial attacks [107, 156, 332] and common corruptions [176,275,526] (*e.g.*, noise) were evaluated through benchmarking on datasets [176,275, 332], using custom measures [107,176] or comprehensive frameworks [437]. When as-

sessing adversarial attacks, figures are corrupted using adversarial or common perturbations [176, 275, 437]. They are then employed to evaluate model behaviour [107, 176] and its generalizability (*i.e.*, the model's capability to classify new instances) [275]. Benchmarks on graph networks were developed too [526]. While most benchmarks focus on assessing defence mechanisms [107, 526], approaches evaluating architecture robustness were also developed [93, 437]. These concentrate on validating architecture design and training techniques against perturbations [437] and resource availability [93]. Such approaches usually consider the data they employ reliable (mainly when well-known datasets are used). Such an assumption should not be taken lightly as it might impact the benchmarking process and its outcomes [310].

When assessing model robustness, picking the correct method or benchmark is as crucial as choosing metrics representing it. Most researchers focused on metrics describing model robustness against adversarial attacks [477, 505] through a local Lipschitz constant estimation problem [477] or loss visualization [505]. The collected literature shows that most articles focus on Computer Vision, while only a tiny fraction discusses metrics in other contexts. For example, linguistic fidelity was employed to extend robustness metrics in NLP [227], while a lack of metrics for tree-based classifier was identified [60], finally highlighting the need for sound and robust metrics in less covered contexts. Computational aspects like precision in computing robustness bounds [398, 496], reducing computational complexity [427], or execution time [483] are also in need of further exploration. Metrics for other aspects of adversarial attacks were also studied, like metrics for convergence stability [238], adversarial attack comparison [36], and inspecting the relationship between robustness and adversarial samples [20], and accuracy [324].

Robustness Methods & Insights. Model robustness assessment encompasses adversarial and natural robustness methods alike. Regarding adversarial robustness assessment, comparison [200, 381], activation function and weight perturbation [386, 410, 449], and language perturbation [228, 293, 371, 470] methods were mainly studied. Comparison studies examined the generalization capabilities [200] and robustness transferability of models based on training data [381], discovering insights concerning the features considered by the network and model robustness, respectively. Novel activation functions for generalizable [449] and robust networks [386, 410, 449] against weight [449] and adversarial perturbations [386, 410] were proposed by inspecting model robustness against both white-box and black-box attacks. Considering NLP models, methods tackling synthetic character- [293], word-level [228, 293], and semantic [371] perturbations were performed, revealing the frailty of NLP models against these attacks. Furthermore, the effectiveness of human-generated annotations in enhancing robustness was also proven [470].

Considering natural robustness, researchers studied common real-world settings involving noise [33, 533] and out-of-distribution data [55, 104]. It was proven that models trained using noisy data resulted in better accuracy and generalizability [533] and robustness [33, 533]. When studying out-of-distribution data, researchers reported that batch normalization (*i.e.*, an approach aligning means and variance of each channel in a CNN across different distributions) might result in accuracy loss [55]. Moreover, per-

formance on in-distribution data is not affected by a single out-of-distribution element in specific contexts [104].

Trade-offs Between Robustness and Trustworthy AI. When it comes to model accuracy, it is essential to study how it is affected by potential robustness-enhancing changes. Several trade-offs were found, discovering how an increase in robustness might result in a decrease in accuracy and vice-versa [40, 281, 343, 423, 450, 488]. In particular, such trade-offs are mainly caused by the network architecture [423]. Other studied aspects include the effects of adversarial training [40, 343, 488] and the relationship between accuracy and in- and out-of-distribution data points [281]. Fairness and robustness are also entwined, as researchers demonstrated the need for mitigating unfairness in adversarial defences [488] as well as the effectiveness of certified robustness and bias mitigation methods in improving fairness [338]. At last, few works research the impact of robustness-improving approaches on prediction-relevant features and their meaningfulness concerning human judgment [313, 481]. Findings discuss the negative effect of adversarial attacks [481] and out-of-distribution examples on such aspects [313], hence highlighting the complexity between robustness, explainability, and feature robustness.

4.2.7 Analyzing Trends and Gaps in Robustness

While the literature covers various perspectives about robustness, trends and gaps were found, including but not limited to explainability, accuracy, human-in-the-loop approaches, and improving and assessing robustness.

Several trends regarding AI and ML robustness can be identified from the literature. Although researchers from diverse domains performed in-depth studies on the impact of data-centred and natural perturbations and how to enhance model architectures and train models, several gaps are yet to be addressed. This section reports on these aspects, highlighting the most relevant gaps and trends in the literature.

Natural Brittleness. An analysis of the collected literature revealed the marginal emphasis on defining natural perturbations and attacks, as most works focused on defining synthetic attacks and potential defence mechanisms against them in CV. Despite the effort, such approaches might not enhance or provide robustness in real-world scenarios. The lack of model-agnostic adversaries underlines another interesting research area, as existing approaches target specific AI systems. Generating model-agnostic attacks would provide a common baseline for evaluating and comparing the robustness of AI systems. Furthermore, model- and perturbation-agnostic robustness assessment would allow disentangling the relationship between scenario-specific aspects and the actual robustness of a model, finally leading to an unbiased analysis of its robustness [477].

Computer Vision. Regarding model robustness, the most researched context is Computer Vision, especially when Convolutional Neural Networks (CNN) are applied regardless of the subtopic of interest. An explanation for such a focus might be the inherent complexity in defining perturbations and attacks in specific data manifolds (*e.g.*,

word embeddings in NLP) or the lack of alignment between ML and context-specific robustness (*e.g.*, in signal processing). Furthermore, pictures have a different complexity in perturbable features and available approaches to compute distances compared to other data types, affecting the research's broadness.

Robustness and Explainability. Acknowledged the brittleness and opaqueness of existing AI models, explainability is essential to generate faithful and trustworthy explanations [159, 160]. Despite their importance, only a few works discuss the robustness of XAI methods and their outcomes. Moreover, explainability might also contribute towards implementing approaches to enhance model robustness. The little work performed relies on the assumption that an alignment between the extracted features and human reasoning (*i.e.*, the features are meaningful towards a prediction for a data point) results in higher robustness. A few examples of works combining explainability and robustness are found in the literature [74, 132, 221, 246, 299].

Accuracy, Robustness, Fairness, and Explainability. Inspecting the literature revealed a strong interplay between various aspects of trustworthy AI. Scholars focused on enhancing robustness at the expense of accuracy, a trade-off similar to the one between accuracy and explainability. The relationship between robustness and fairness and its possible issues were also of interest. These aspects are of equal importance when building trustworthy and fair AI systems. In that regard, data-driven approaches were proven limited, finally pointing towards integrating symbolic knowledge (although they were not extensively discussed in the literature).

Human-in-the-loop approaches were also found in the literature, involving humans in assessing or improving model robustness. Yet, potential challenges in accomplishing such tasks are not discussed in detail, potentially threatening the effectiveness of methods and frameworks for robustness. Further investigation on such a topic is needed in areas where human involvement is essential.

Enhancing Robustness. When improving robustness, humans can be involved in various aspects of the process, *e.g.*, in collecting and refining adversarial examples [202]. Similarly, researchers demonstrated that human knowledge contributes to generating more robust models, *e.g.*, by leveraging human uncertainty on sample labels [329], by integrating human rationales into the training process [71, 299], or by actively querying the most relevant perturbations from an expert during training [307]. Furthermore, the proposed approaches might be enhanced by leveraging existing research on human computation to assess or enhance the quality of the crowdsourced outputs [128, 195] to understand the nature of uncertainties and define useful human rationales, especially when subjective tasks are involved.

Assessing Robustness. When assessing model robustness, human knowledge might potentially contribute to the design of suitable perturbations or attacks. Researchers defined some constraints on this topic that must be verified to be considered realistic [202, 227]. While this approach represents a first step towards designing proper samples, to what extent such constraints and the samples they transform align with the

human concept of "realistic" has to be properly addressed. Similarly, works on robustness to natural perturbations should define a comprehensive set of domain-specific perturbations relevant to the problem at hand and its context. On such a topic, research developing benchmarks or robustness-enhancing methods are yet to achieve comprehensiveness [176, 218], developing tools to support defining relevant perturbations (*e.g.*, collaborative documentation of domain-specific perturbations, libraries to generate such perturbations semi-automatically, and many more).

4.2.8 Involving Practitioners in ML Robustness

> The most important identified gap was the lack of human-in-the-loop approaches, human engagement, and support for practitioners in dealing with robustness.

Among the identified gaps, the lack of human-centred approaches and workflows to support ML practitioners in handling robustness were two of the most important.

Robustness by Human Knowledge Diagnosis. In the literature, most robustness-related research focuses on generating out-of-distribution (OOD) data to enhance model robustness at training time [41, 71, 139, 329]. Regarding natural perturbation, it is very challenging to characterize the data types a model might be provided as input before generating the actual data [104, 108, 153]. The motivations behind such challenges might be various. In some cases, practitioners might not generate out-of-distribution-data for privacy or contractual reasons [184, 457], or changes to the deployed model's context or goal might happen, requiring a constant model evolution and dynamic out-of-distribution data generation [369]. Collaboration with domain experts might also be complex [297, 512], especially when discussions on which data should be considered out-of-distribution are made or when evaluating the meaningfulness of model features to estimate robustness [132, 299]. Furthermore, practitioners might lack budget, time, or training resources when dealing with model trustworthiness [346]. While these challenges might hinder the process of achieving model robustness, an interesting research direction has yet to be explored. The development of complementary, hybrid human-machine approaches leveraging explainability, crowdsourcing, human-in-the-loop ML, and knowledge-based systems might contribute to estimating model performance on more realistic data distributions without requiring such distributions.

Existing Approaches. Only a narrow part of the literature leverages human capabilities to identify and mitigate potential model failures. Practitioners might employ dataset explanations [370] to identify data skews that might impact model performance. On this topic, humans were involved in identifying patterns among unknown unknowns to train classifiers to automatically detect such patterns on new data [256]. Aside from datasets, models trained using features aligned with human reasoning were developed to achieve high robustness [22, 415]. These approaches leverage human explanations of inference tasks to control and align the features learned by the model.

Envisioned Research Opportunities. Besides a model's prediction and its confidence, the approaches collected from the literature highlighted potential elements that

can be leveraged to estimate robustness, like model features or the training dataset. Assessing model features and their human alignment contributes to assessing model robustness to OOD data points, as features might not be meaningful even when a correct prediction is made. Moreover, understanding the link between such features and the dataset might help mitigate unaligned features.

Surfacing Model Features using Explainability and Human Computation. Various explainability methods might be employed to extract model features [370]. In particular, while some models are built to be explainable by design [428, 518], others require post-hoc interpretability methods [30, 355, 421] with different properties (*e.g.*, nature of explanations being correlation- or causation-based, different scopes be it local or global, different mediums be it visual or textual, etc.) [253, 399, 407]. However, existing explanations often allow for various interpretations, and practitioners may lack the necessary domain expertise to disambiguate the highest-fidelity features. For example, saliency maps [396] or image patches [147, 212] do not represent the actual human-interpretable features the model has learned. Indeed, achieving human and model features alignment might require clear human concepts [28], highlighting the need to extend the research on semantic and concept-based explanations acquired via human computation [30, 180].

Leveraging Knowledge Acquisition for Identifying Expected Features. While very few works focused on understanding a model's expected features to achieve feature alignment [387], research on commonsense-knowledge acquisition could be helpful [509]. In this research field, methods harvesting knowledge from existing sources (*e.g.*, libraries) or through human involvement (*e.g.*, GWAPs [27, 361, 463] or crowdsourcing tasks [189, 373]) were proposed. These techniques should be adapted to collect knowledge of interest, focusing on translating such knowledge into significant feature-based information.

Comparing Features via Reasoning Frameworks and Interactive Tools. The alignment between the model and its expected features is also of fundamental interest to researchers. Interactive frameworks and user interfaces represent the first step [28], enabling feature exploration with various degrees of automation for expected feature comparison. AI diagnosis [91, 359] (*e.g.*, abductive reasoning or automated feature-reasoning) approaches might also contribute towards speeding up the feature comparison process whilst making it more reliable.

Involving Humans in the ML Lifecycle. Most approaches proposed in the literature involve humans to increase model performance, while none consider human involvement to enhance model robustness. Hence, it is fundamental to investigate how existing approaches could be shaped to improve model robustness.

ML with a Reject Option. ML systems are usually designed to provide an output for all input samples. When critical decisions are involved, always providing a prediction might be harmful, especially when they are likely incorrect. Accordingly, researchers developed approaches capable of detecting when human agents must be involved as the model's outcome might not be correct [175]. Rejectors can be integrated at dif-

ferent stages of the ML pipeline, *i.e.*, before the predictor filters input samples, after the predictor to exclude improper predictions, or alongside the predictor to predict the output's correctness. These approaches bear advantages and disadvantages based on the context in which they are applied. Moreover, adapting these approaches would strongly benefit the research on robustness as only a few works discussing them were identified [324, 422].

Human-in-the-loop ML Pipelines. Recently, the interest of the ML research community shifted towards developing human-in-the-loop frameworks. These systems are employed to account for noisy crowdsourced labels [349] by defining models of the annotation process (*e.g.*, task difficulty, task subjectivity, annotator expertise, etc.) and often rely on active learning to reduce costs [491, 494]. Novel human-in-the-loop approaches focus on building model pipelines through human involvement to achieve various objectives, *e.g.*, to identify model weaknesses [314], to identify noise or biases in the training data [188, 495], or to propose potential explanations to incorrect predictions [57]. Only a few research articles in the inspected literature discuss the intersection between active learning and adversarial training [278, 279, 394, 401]. On the other hand, no work analyzes robustness types and strategies to involve humans in the ML pipeline when robustness is desired, making them an interesting research topic. Furthermore, such research subjects are promising and realistic scenarios of ML systems that succeed in making the model more accurate.

Supporting ML Practitioners in Handling Robustness. The problem of enhancing model robustness should not only be tackled theoretically. As practical solutions must be established, it is crucial to understand the barriers practitioners encounter when making their systems robust. Such a topic has yet to be addressed, and very few articles are close to such topics [385].

Understanding Practices Around Robustness. In the HCI community, semi-structured interviews with ML practitioners have been extensively applied. Such literature covers various topics, like stakeholder collaboration [225, 333], debugging practices [29], tools for explainability [181, 185, 253] or fairness [239, 358], and resulted in frameworks modelling practitioner's process and challenges. Applying a similar approach to robustness would contribute to defining a better research direction. Furthermore, defining a robustness question bank (similar to what was done in Explainability [253]) would provide a structured understanding of potential research gaps. Besides, the literature discussing fairness demonstrated the gap concerning the guidance in choosing metrics and mitigation methods. Acknowledging the plethora of robustness metrics and techniques, user studies around robustness would reveal a similar gap that could be bridged by taking inspiration from the fairness literature.

Integrating Robustness into ML Workflows. When building models, workflows [405], analysis tools (*e.g.*, user interfaces to investigate a model's details [28, 302]), and documentations or checklists [18, 140, 284] might be of help to researchers. Although few speculations on developing similar approaches were proposed [390], designing such supportive tools and integrating them into existing solutions might enhance the robust-

ness of the research field.

4.3 The Role of Human Knowledge in XAI

Human understandability is another relevant focus for the XAI community, as humans are involved in multiple stages of the XAI cycle. Literature covering each role is collected, organized and discussed.

Research Methodology. This research focuses on articles and papers published from 2017 to 2022. Articles were collected from bibliographic databases, combining input from open-access (*i.e.*, Google Scholar) and subscribe-only (*i.e.*, Scopus) sources in the field of computer science. A strategy aligned with the PRISMA methodology [322] for literature reviews was implemented. A search strategy to collect articles that include any pair of concepts created by combining the keywords listed in Table 4.2 was defined, finally resulting in 48 different combinations of keywords. In particular, pairs of keywords were generated by concatenating one explainability keyword (those in the table's left column) with one knowledge-related keyword (those in the right column of the table).

Table 4.2: *The list of keywords used to generate the couples used to search for papers.*

Explainability-Related Keywords	Knowledge-Related Keywords
Interpretable Machine Learning	Knowledge Extraction
Explainable Machine Learning	Knowledge Elicitation
Explainable Artificial Intelligence	Crowdsourcing
Explainable AI	Human-in-the-Loop
Explainability	Human-centred Computing
Interpretability	Human-centred Computing
	Human Computation
	Concept Extraction

All articles whose titles contained the words "survey" and "review" were excluded. When querying Google Scholar, only the first 100 articles ranked by relevance for each query were considered since a drop in pertinence to the topics of interest was identified after the 80th position in the ranking. The performed queries collected

- 3718 non-unique articles from Google Scholar, queried using the Publish or Perish software, resulting in 2056 unique papers;
- 327 non-unique articles from Scopus, queried using the Scopus web interface, resulting in 216 unique articles.

Most queries performed on the Scopus bibliographic database returned very few results, as the scope of each was quite narrow. A final integrated set of 2197 unique articles was achieved by combining the resulting sets of articles and removing duplicates. These were manually inspected, considering only the ones attaining the scope of this research. In particular, all and only the articles in which humans and human knowledge played a fundamental role concerning the explainability of an AI or ML system

were considered. Finally, the bibliographic references cited in the collected documents were also inspected, extending the literature.

4.3.1 Human Knowledge and Explainability

> Humans are involved in various stages of the XAI cycle, *i.e.*, for collecting knowledge, assessing, improving, or interacting with explanations.

Bridging the understandability gap between humans and black-box models requires the development of techniques able to answer the many-faceted problem of explainability, addressing the faithfulness and completeness of the explanations representing the model's behaviour while also accounting for the capability of the human interpreter to understand them. In ML, humans are commonly employed to collect or label data, debug and evaluate the outcomes of machine learning models, and many more [52]. Due to the recent enthusiasm for XAI, researchers' data interests shifted towards collecting human knowledge in the form of human rationale [417], *i.e.*, the reasoning humans apply to perform ML tasks. Such valuable [420] information is at the centre of many explainability-related tasks and it can be employed in various steps of the XAI cycle, *i.e.*, the process revolving around developing and explaining an AI model with human involvement. In a broader sense, human knowledge is also applied in most human-in-the-loop approaches in which explanations are used to explore [155], evaluate, or improve the explainability and the performance of models. Furthermore, humans are directly involved in the creation [270], assessment, or improvement [469] of such explanations or the model itself [231]. Given the critical role of human knowledge in such processes, human-in-the-loop approaches are essential to achieve interpretable and explainable AI [68, 121]. Approaches using humans and their knowledge to achieve such objectives are reported, organized in categories, and finally discussed. These categories include various parts of the XAI cycle (represented in Figure 4.3) and can be classified as follows.

- **Knowledge Collection**, *i.e.*, approaches collecting human knowledge in various contexts (*e.g.*, computer vision and NLP) and shapes by involving novices and experts alike.
- **Explainability Evaluation**, *i.e.*, approaches assessing explanations through metrics and benchmarks involving human knowledge by directly involving humans in assessing model interpretability and trustworthiness.
- **Human Understanding**, *i.e.*, approaches understanding how humans perceive, understand, and interact with explanations and whether these can benefit human interpreters.
- **Explainability Improvement**, *i.e.*, approaches employing explanations and humans to improve models and their explanations and perform debugging.

Figure 4.3 represents the different steps of the XAI cycle involving humans, the way the latter are engaged, and the aforementioned categories. These are described in the following sections.

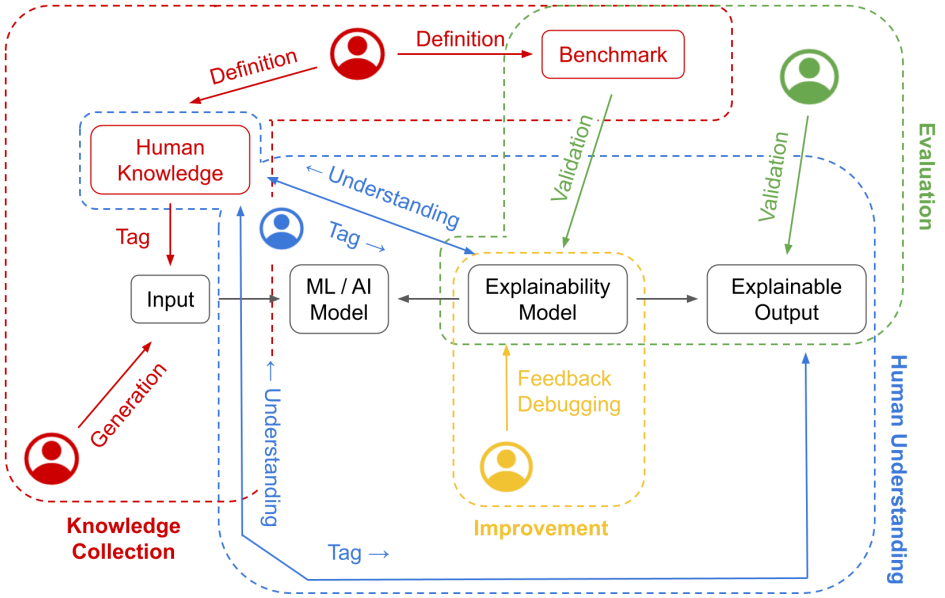


Figure 4.3: *The figure represents the four main areas in which human knowledge is employed in XAI, i.e., knowledge collection (red), explainability evaluation (green), understanding human perspective (blue), and improving model explainability (yellow). In the schema, the human icons represent the steps in which human actors are involved in the XAI cycle.*

4.3.2 Explainability and Human Knowledge Collection

The literature describing tasks in which humans are involved in collecting data for XAI approaches through crowdsourcing is organized and discussed.

In computer science, crowdsourcing is a well-known practice widely employed to collect a large amount of human-generated data by engaging heterogeneous groups of people with varying features and knowledge in undertaking a task [120]. Given the fundamental role of humans in XAI, crowd knowledge collection is essential to leverage human intelligence at scale to achieve robust, interpretable, and hence trustworthy AI systems [136]. When addressing the explainability of black-box models, many different factors influence such an approach. Depending on the system's complexity [229], the model's purpose, the task's complexity, and its goal, it might be necessary to involve individuals with specific knowledge or features [235]. Indeed, complex explainability-related tasks may need preliminary expertise, requiring the involvement of expert users [191]. For example, collecting and employing human knowledge to label [30, 387] and evaluate visual explanations extracted from an image classification model (*e.g.*, heatmaps) could be trivial as users may be asked just to highlight parts of the picture they deem to be important [283]. On the other hand, editing attention maps to successfully improve the explainability of a system [285] or providing domain-specific knowledge [417] are not tasks that non-expert users can easily accomplish. Hence, interactive approaches have been developed to employ human knowledge at its best while accounting for such complexities.

In the context of QA systems, Li et al. [252] collected a dataset by engaging crowd

workers in interacting with an initial model and providing feedback on the quality of the answers in a structured and unstructured way. The collected data was then employed to train a new model, extending the original with re-scoring and explanation capabilities. In the context of image classification tasks, Mishra et al. [283] designed a concept elicitation pipeline to gather high-level concepts to build explanations for image classification datasets. Data was collected as mask-label pairs by showing the picture's true label to users and having them outline the entity and the features they used to identify it. Per-image and per-class aggregations were employed to build a variety of concept-driven explanations. Similarly, Uchida et al. [451] proposed a human-in-the-loop approach to collecting human knowledge to generate logical decision rules to explain the output of classification models. They explained the outcome of the original model by collecting human-interpretable features of pictures as text to create rule tables associating the classes and the collected features. Balayn et al. [27, 30] proposed a GWAP to collect high-quality discriminative and negative knowledge. Inspired by the popular game *GuessWho?*, users are engaged in a competitive, two-player game in which each user guesses the card assigned to the challenger by asking questions about the represented entity. The answers represent structured knowledge about the entity. Such a particular kind of knowledge can be helpful to improve the trustworthiness and robustness of AI systems. Zhao et al. [525] designed ConceptExtract, a system implementing a human-in-the-loop approach to generate user-defined concepts for DNN interpretation. Users can overview and filter image patches extracted from input pictures, provide new visual concepts, and overview the performance and the interpretation of the target model. Attempting to achieve a similar objective, Lage et al. [230] proposed a human-in-the-loop approach to learn a set of transparent concept definitions relying on the labelling of concept features. Users were engaged in providing their understanding of the domain of interest, making the collected concepts intuitive and interpretable. They were asked to define the associations between a series of features and concepts and provide feedback on whether the function learned by the model satisfies the conditions. Similarly to Zhao et al. [525], Zhang et al. [521] developed FaxPlainAC, a tool to collect user feedback on the outcome of explainable fact-checking models. When the system receives a query, its decision, *i.e.*, the truthfulness of the input fact, and the considered evidence are displayed. Users are asked whether the documents employed to generate such content support or refute the input by highlighting the most relevant parts of the text or whether they are misleading or irrelevant. Sevastjanova et al. [384] extended the usage of explainability to support interactive data labelling of complex classification tasks by applying visual-interactive labelling and gamification. Such an approach is implemented in QuestionComb, a rule-based learning model that presents explanations as rules, supporting iterative and interactive data optimization.

These methods demonstrate data collection processes may employ explanations and model details to improve the accuracy of the collected knowledge. Furthermore, the strategy to apply is also influenced by the kind of data desired, *i.e.*, task-specific or generic, resulting in the design of a variety of data collection techniques.

4.3.3 Evaluation of Explainability Methods using Human Knowledge

In the literature, several human-in-the-loop approaches employ human knowledge to assess explanations' features (*e.g.*, their understandability). Similarly, metrics and benchmarks using human knowledge were built.

The design and implementation of approaches to choose the best explainability method or explainable model have been at the centre of discussion of the research community for years. Consequently, recent research efforts have focused on collecting and developing benchmarks as they enable, organize and standardize the evaluation and comparison of multiple models through explainability-related measures. Mohseni et al. [287] developed a benchmark for quantitative evaluation of saliency map explanations of images and text tasks through multilayer, aggregated human attention masks. They collected human annotations of salient features by asking users to highlight the most representative parts of documents or images. The efficacy of their approach was validated through a series of experiments, demonstrating its capabilities to evaluate the completeness and correctness of model saliency maps. De Young et al. [101] proposed the Evaluating Rationales And Simple English Reasoning (ERASER) benchmark, comprising various datasets and tasks extended with human annotations of rationale. Such datasets cover various NLP tasks, such as question answering, sentiment analysis, and many more. They evaluated their benchmark on baseline models using metrics designed to measure faithfulness and the agreement between human annotations and the model's extracted rationales. While benchmarks provide fixed datasets to evaluate model explainability, Schuessler et al. [376] developed a library that allows researchers to create customized datasets for human-subject and algorithmic evaluations of explanation techniques for image classification.

The employment of automatic metrics to evaluate and compare model explainability is still debated in the XAI literature. In particular, it is argued that the metrics used to evaluate explainability methods must be carefully chosen, while significant room for improvement exists for such assessment approaches [173]. Moreover, exploring the relation between human-based and automatic evaluations is another aspect researched in the XAI community [305], as various evaluation methods and approaches have been proposed [304]. On such a topic, it is still argued that the best way to assess the interpretability of black-box models is through user experiments and user-centred evaluations, as there is no guarantee for the correctness of automated metrics in evaluating explainability [261] and high explainability metric scores do not necessarily reflect high human interpretability in real-world scenarios [122, 261]. The same is true for well-known metrics (*e.g.*, F1-score) [377]. Supporting such claims, Fel et al. [122] conducted experiments to evaluate the capability of human participants to leverage representative attribution methods to learn to predict the decision of various image classifiers. Such a process aimed to assess the usefulness of explainability methods and the capability of existing theoretical measures in predicting their usefulness in practice. The designed framework can be employed to perform such evaluation given a black-box model, an explanation method, and a human subject to predict the predictor (*i.e.*, so-called meta-predictor). A two-phase procedure is applied. In the learning phase,

the human meta-predictor is trained using triples made of an input sample, the model's prediction, and its explanation to uncover rules describing the model's functioning. In the evaluation phase, the accuracy of the meta-predictor (*i.e.*, the relevance of the rules they learned) is tested on new samples by comparing their predictions with the ones provided by the model. In their conclusions, the authors argue that faithfulness evaluations are poor substitutes for utility and that putting humans in the loop is necessary. Moreover, they discuss that such metrics do not account for the usefulness of the explanation to humans as, in some cases, they can either be useless or generate ambiguity. We argue the main problem is not related to the application of automatic evaluations and metrics but to the interpretation of the computed (faithfulness) scores. Faithfulness is just one side of the coin, *i.e.*, the model's side, as it measures the closeness of the derived explanation concerning the true reasoning process of the model. The other side of the coin is represented by interpretability, *i.e.*, explanations must be made so that human interpreters can properly understand it. Misunderstandings occur when there is confusion between these two aspects. Indeed, model faithfulness and interpretability are not to be considered equivalent when it comes to the evaluation of the explainability of models.

An explanation's interpretability is usually evaluated by involving users in manually interpreting the explanations generated by the model or derived through explainability methods. The same approach applies to evaluating black-box models' interpretability, *i.e.*, directly understanding the intrinsic explainability of a model [133]. Such evaluations are usually achieved through user questionnaires [174, 377, 465, 506] whose questions vary depending on the nature of the experiment, model, etc. On the other hand, comparing the interpretability of different explainability methods to choose the best-suited one requires the design and implementation of ad hoc human-in-the-loop approaches. Soltani et al. [409] improved the existing XAI algorithm by employing cognitive theory principles to provide explanations similar to domain experts. Humans were involved in experiments to evaluate the novel and basic approaches to understanding, which led to the best explanations. Lu et al. [261] designed a novel human-based evaluation approach using crowdsourcing to assess saliency-based XAI methods, mainly focusing on methods that explain the prediction of picture-based models, *e.g.*, Grad-CAM [382], SmoothGrad [403], etc. through a human computation game named "Peek-a-boom". Their human-centred approach compares different explainable AI methods to identify the one that yields the best interpretations. In the proposed GWAP, the XAI method plays the role of Boom, revealing parts of an image as the game progresses. The player plays the role of Peek, guessing the entity in the picture from the parts displayed. In summary, evaluating the explainability of black-box models requires assessing human interpretability and faithfulness while not misunderstanding these two concepts and consequently generating unmotivated trust.

More commonly, humans evaluate the effectiveness of methods in generating explanations and their usefulness in real scenarios [13, 348, 350, 523]. Zhao et al. [523] employed Generative Adversarial Networks (GANs) to generate counterfactual visual explanations. Crowd workers were recruited to evaluate their effectiveness in classification tasks. In Visual Question Answering, Arijit et al. [348] involved users in a

collaborative image retrieval game named Explanation-assisted Guess Which (ExAG) to evaluate the efficacy of explanations, finally demonstrating their usefulness. Alvarez-Melis et al. [13] implemented a method to generate explanations based on the weight of evidence from information theory. User experiments demonstrated the methodology's effectiveness in generating accurate and robust explanations, even in high-dimensional, multi-class settings. Zeng et al. [511] present a human-in-the-loop approach to explain ML models using verbatim neighbourhood manifestation. A three-stage process is employed to (i) generate instances based on the chosen sample, (ii) classify the generated instances to define local decision boundaries and delineate the model behaviour, and (iii) involve users in refining and exploring the neighbourhood of interest. A series of experiments revealed the effectiveness of the implemented tool in improving human understanding of model behaviour. Baur et al. [39] presented NOVA, a human-in-the-loop annotation tool to interactively train classification models from annotated data. The tool employs semi-supervised active learning to pre-label data automatically. Moreover, it implements recent XAI techniques to give users a confidence score about the predicted annotations and visual explanations. Heimerl et al. [174] employed NOVA in emotional behaviour analysis. They engaged non-expert users and evaluated the impact and the quality of the extracted explanations, revealing their effectiveness while getting useful insights on the employment of visual explanations. Steging et al. [417] proposed a knowledge-driven method for model-agnostic rationale evaluation employing human-in-the-loop to collect dedicated test sets to assess targeted rationale elements based on expert knowledge of the domain.

Finally, while part of the XAI research community focused on designing and implementing methods to generate explanations, developing techniques to generate trust in models is another fundamental aspect of interest. Zöller et al. [535] implemented XAutoML, an interactive visual analytic tool to establish trust in AutoML-generated models. The user-centred experiments revealed the tool's effectiveness in generating trust while addressing the explainability needs of various user groups (*i.e.*, domain experts, data scientists, and AutoML researchers). De Bie et al. [96] proposed and evaluated RETRO-VIZ, a method to estimate and assess the trustworthiness of regression prediction. The system comprises RETRO, a method to estimate the trustworthiness of the prediction quantitatively, and VIZ, a visualization provided to users to identify the reasons for the estimated trustworthiness. Although they demonstrated the effectiveness of their methodology, the authors remark it must be used with caution so as not to generate unguided trust.

4.3.4 Understanding Human Perspective in Explainable AI

> Humans have been involved in understanding the effectiveness, level of detail, and other features of explanations.

An explanation that cannot be properly understood has no value and may potentially mislead humans. Indeed, providing accurate and understandable explanations is essential as poor explanations can sometimes be even worse than no explanation at all [312] and may also generate undesired bias in users [38, 375]. Consequently,

properly structuring [205] and evaluating the interpretability and effectiveness of explanations requires a deep understanding of how humans interpret and understand them while also accounting for the relationship between human understanding and model explanations [73, 522]. For such reasons, the explainable AI research field spreads from IT-related fields, such as computer science and machine learning, to a variety of human-centred disciplines, such as psychology, philosophy, and decision-making [17]. Therefore, recent studies evaluating human behaviours when exploring, interpreting and using explanations have been conducted [301, 404, 487]. Moreover, Gamification and Games With a Purpose have been proven quite effective when designing methods to assess how humans interpret XAI explanations [135]. Feng et al. [125] evaluated how humans employ model interpretations and their effectiveness, measured in terms of improvement in human performance. They designed Quizbowl, a human-computer cooperative setting for answering questions, supporting various interpretations, and guiding users to trust the model's prediction. Questions are displayed word-by-word, and players are asked to stop the display when the model's interpretations are enough to answer the question correctly but before it is completely revealed. They discovered that interpretations help non-expert users and experts in different ways. Additionally, while expert users could mentally tune out bad suggestions, novice users trusted the model too much, consequently choosing an incorrect answer. Such a result demonstrates that even though one of the objectives of explainability is to improve users' trust in the model, it is necessary to organize the content provided to avoid generating a sense of overconfidence in the system. A similar result was achieved by Ghai et al. [146], who combined XAI techniques in the context of Active Learning. They analyzed the impact of the proposed approach while also researching human-related aspects. Their findings revealed explanations successfully supported users with high task knowledge while impairing those with low task knowledge. Indeed, users with low knowledge were more prone to agreeing with the model, even when it misbehaved. On the other hand, they demonstrated the effectiveness of explanations in calibrating user trust and evaluating the model's maturity. In conclusion, achieving high transparency is not always beneficial to improving the user's understanding [335, 375]. Indeed, providing complex or a large number of explanations would generate a trade-off between their understandability and the time required by human interpreters to interpret them [229, 254].

Consequently, it is necessary to comprehend the proper level of transparency, explanation complexity and quantity, even in simple cases [114]. Regarding such an aspect, Mishra et al. [283] performed user studies to understand the proper level of conceptual mapping using granularity and context of the data to generate explanations. The authors discovered that balancing coarse and fine-grained explanations helps users understand and predict the model's behaviour. On the contrary, using structured coarse-grained explanations negatively impacted user's trust and performance. While Mishra et al. [283] focused on understanding the granularity of the explanations, Kumar et al. [226] compared the visual explanations provided by the proposed visualization framework concerning two text-based baselines, revealing the effectiveness of their approach in the context of interest through user experiments. In conclusion, engaging humans in XAI is fundamental as primary targets of explanations. Additionally, improving their understanding of explanations and models is beneficial to enhancing the design and development of explanations.

Moreover, it is desirable to design flexible explanation approaches and explainability methods that properly convey model behaviour depending on "who" the human is [56, 114, 357]. Turró [357] categorizes the main user groups. Depending on their *goals, background and relationship with the product*, users are grouped into three categories: developers and AI researchers, domain experts, and lay users. The author discusses the importance of approaching explainable AI in a user-centred manner, providing tailored explanations based on the needs and characteristics of the targeted group of users, finally improving affordability and user satisfaction, and easing the explanation evaluation process. Striving to understand how and why such groups employ explanations and behave, several researchers have carried out experiments by engaging specific user groups. Hohman et al. [181] involved professional data scientists in exploring how and why they interpret ML models and how explanations can support answering interpretability-related questions. Users can be generally classified as domain or expert users and non-expert users. Nourani et al. [311] inspected the behaviour of such user groups on their first impression of an image classification model based on the correctness of its predictions. They discovered that providing early errors to domain experts decreases their trust, while early correct predictions help them adjust their trust based on their observations of the system performance. On the other hand, non-expert users relied too much on the predictions made by the model due to their lack of knowledge. Such over-reliance on the ML system [125, 146, 311] highlights how it is always necessary to account for the users engaged in the system. Moreover, while it is necessary to engage non-expert and end users in evaluating such a system, it is also recommended to consider their features, preliminary knowledge, and understanding of the system of interest.

Finally, while explanations were proven to be effective in leading users in achieving a task and improving their trust and understanding of the model, it has also been demonstrated that sometimes they are either not able to improve [86, 465] or, worse, they reduce human accuracy and trust [389]. A similar result in a different context was found by Dinu et al. [103]. They focused on post hoc feature attribution explanations. They discovered that such explanations *provide marginal utility in our task for a human decision maker and, in some instances, result in worse decisions due to cognitive and contextual confounders*. Such findings bring forth a fundamental conclusion. Even though explanations and explainability methods may improve users' understanding, accuracy and trust [493], it is still necessary to investigate how humans perceive such content concerning the context, the model, and the performed task.

4.3.5 Human Knowledge as a Mean to Improve Explanations

Human knowledge has been employed to improve and debug explanations, involving humans in inspecting AI explanations and systems.

As faithful explanations provide meaningful insights into the behaviour of models, researchers have designed novel and effective methods to employ such content to improve the explainability and performance of models. Such human-in-the-loop approaches mainly display a model's explanations and outcomes to humans, who are

then asked to discover undesired behaviours (*i.e.*, debugging the model) and to provide possible corrections. The effectiveness of such explainability-focused approaches is discussed by Ferguson et al. [126]. They report on the usefulness of explanations for human-machine interaction while stating that augmenting explanations to support human interaction enhances their utility, creating a common ground for meaningful collaboration. They experienced the effectiveness of editable explanations, modifying the machine learning system to adapt its behaviour to produce interpretable interfaces. Many examples of approaches that use such a strategy can be found in the literature. Mitsuhara et al. [285] propose a novel framework to optimize and improve the explainability of models utilizing a fine-tuning method to embed human knowledge collected as single-channel attention maps manually edited by human experts. They reveal that improving the model's explainability also contributes to performance improvement. Coma et al. [89] designed an iterative, human-in-the-loop approach to improve the performance and explainability of a supervised model detecting non-technical losses. In particular, each iteration improves the performance and reduces the complexity of the model to improve its interpretability. Kouvela et al. [224] implemented Bot-Detective, a novel explainable bot-detection service offering interpretable, responsible AI-driven bot identification focused on efficient results detection and interpretability. Users can provide feedback on the estimated score and the interpretation quality while specifying their agreement and describing potential explanations' improvements provided through LIME [355]. Such an approach not only improves the explainability of the model but also contributes to the model's performance. Collaris et al. [88] introduced an interactive explanation system to explore, tune and improve model explanations. The tool allows stakeholders to tune explanation-related parameters to meet their preferences while they employ such evidence to diagnose the model and discover potential model or explanation improvements. Yang et al. [500] addresses the problem of generalisability by allowing users to co-create and interact with the model. The authors introduced RulesLearner, a tool that can express ML models as rules while allowing users to interact with and update the patterns learned. Their studies demonstrated the proposed approach's effectiveness in improving the analyzed system's generalisability and the quality of the explanations employed in the process. In the presented systems, users directly interact with the explanations of the model to improve their explainability. Other studies collect and employ human rationales [22] or domain knowledge [90, 116] to achieve the same goal. Arous et al. [22] introduced MARTA, a Bayesian framework for explainable text classification. Such a system integrates human rationales into attention-based models to improve their explainability. Confalonieri et al. [90] evaluated how ontologies can improve the human understandability of global post hoc explanations, presented as decision trees. The proposed algorithm enhances extracted explanations using domain knowledge modelled as ontologies. While sometimes increasing the model's performance is a side effect of improving its explainability [88, 89, 224, 285], a few researchers employed explanations to improve model performance [92, 252, 411]. Li et al. [252] collected human feedback, made of a rating label and a textual explanation describing the quality of the answer, to improve a BERT-based Question Answering model's performance and capability of explaining the outcome's correctness. While Li et al. [252] employed human feedback, Spinner et al. [411] engaged humans in a conceptual framework focused on practicability, completeness and full coverage to op-

erationalize interactive and explainable machine learning. The most relevant element of the system is the Explainable AI pipeline, which maps the explainability process to an iterative workflow that allows users to understand and diagnose the system to refine and optimize the model.

Another process benefiting from faithful explanations is model debugging. Such an activity employs human knowledge and expertise to identify errors, bias and improper behaviours in models with the final objective of correcting them, consequently improving them [243] or their explanations. The central concept on which model debugging is based is the interactive exploration of models [162, 182, 312] using an interface to summarise its behaviour. Moreover, allowing users to interact with explanations produces an even deeper understanding of the model behaviour, consequently improving their capability to identify potential bugs. In this scenario, providing faithful, complete, and understandable explanations is essential, as they influence the ability of users to identify such errors and the soundness of the results. With the final aim of understanding the model's failures, Nushi et al. [315] implemented Pandora, a system leveraging human and system-generated observations to describe and explain when and how ML systems fail. The tool employs content-based views (*i.e.*, views creating a mapping between input and the overall system's failure) to explain when the system fails, while component-based views (*i.e.*, views modelling how internal model dynamics lead to errors) explain how the system fails. Crowdsourced human knowledge is employed for various purposes, such as system evaluation, content data collection, and component quality features data collection. Liu et al. [258] describe an error detection framework for sentiment analysis models based on explainable features employing a variety of explanations. Their approach is organized into four different units, namely, a "local-level feature contributions" module extracting unigram features through LIME [355], a "global-level feature contributions" module performing perturbation-based analyses by masking individual features of the training samples, a "human assessment" module asking humans to assess the most relevant globally contributing features learned from the previous step, and a "global-local integration" module that quantifies the erroneous probabilities of instance-level predictions made by the model. Even though providing a wide variety of interactive explanations may improve the debugging of ML systems, it is still unclear which ones are the most useful. Seeking to answer such a question, Balayn et al. [28] developed an interactive design probe that provides various explainability functionalities in the context of image classification models. They discovered that common explanations are primarily used due to their simplicity and familiarity. In contrast, other types of explanation, *e.g.*, domain knowledge, global, textual, active, interactive, and binary explanations, are still helpful in achieving various objectives. Such conclusions support and highlight the importance of presenting diverse explanations. Using explanations to debug models could also benefit the explanations themselves. For example, Afzal et al. [5] described a human-in-the-loop explainability framework to debug data issues to enhance interpretability and facilitate informed remediation actions. In conclusion, the variety of human-in-the-loop approaches described demonstrates that human knowledge can be a valuable asset even for tasks that do not employ it as structured data and directly engage humans in understanding, fixing and optimizing ML models.

4.4 Final Remarks

In the first part of this chapter, the literature related to robustness in AI systems was collected, organized, and discussed, highlighting various concepts, definitions and domains. Three main themes were thoroughly addressed, *i.e.*, approaches to improve model robustness against adversarial and non-adversarial perturbations, approaches to enhance robustness in different application areas, and evaluation approaches and insights. In the end, research gaps and a lack of human-centred solutions concerning robustness were emphasized, while future research might strongly benefit from involving human actors in robustness-centred methods. On a similar topic, the second part of the chapter focuses on human involvement and human-in-the-loop approaches, highlighting the various roles humans might have in the ML cycle. These mainly cover four main aspects, *i.e.*, human knowledge collection and structuring, XAI methods assessment by means of human knowledge, human understanding of explanations, and human knowledge applied to enhance explanations. In the end, human-driven approaches are growing in relevance in modern research as the ever-increasing complexity of models makes it necessary to make them trustworthy and interpretable in the eyes of humans, hence requiring human knowledge, reasoning, and involvement.

CHAPTER 5

Explainable AI in Natural Language Processing

This chapter discusses the research performed on explainable AI in the context of natural language processing (NLP). In particular, the theoretical formalization of *rationale mappings* and *trees* and their collection process are described. This chapter is mainly built upon the article

1. Andrea Tocchetti, Jie Yang, and Marco Brambilla. “Rationale Trees: Towards a Formalization of Human Knowledge for Explainable Natural Language Processing". In: Proceedings of the 4th Italian Workshop on Explainable Artificial Intelligence colocated with 22nd International Conference of the Italian Association for Artificial Intelligence(AIxIA 2023), Roma, Italy, November 8, 2023. Vol. 3518. CEUR Workshop Proceedings. CEUR-WS.org, 2023, pp. 29-46. url: `https://ceur-ws.org/Vol-3518/paper3.pdf`.

and it is extended by the subsequent research carried out in cooperation with MsC. Valentina Naldi, involving designing, implementing, and assessing a human-in-the-loop approach to collect instances of the defined data structures. The PhD candidate contributed by examining the literature on the context of interest, finally leading to the design of *rationale mappings* and *trees*. Furthermore, the candidate strongly contributed to designing and structuring the application, the data collection process, and the experiments. Part of this chapter will be soon published as follows

1. A. Tocchetti, V. Naldi, and M. Brambilla. "Web-based Human-centered Explainability of NLP Tasks with Rationale Mapping Theory". In: Proceedings of the 16th International Conference on Applied Human Factors and Ergonomics. July 26-30, 2025, Florida, US.

5.1 Introduction

As the thought process applied to NLP tasks by humans is usually neglected when building explanations, a structure for organizing such knowledge and a process to collect it are described.

In the context of NLP, researchers have defined explanations by identifying the most critical words in the input(s) [240], providing the most influential training examples affecting an outcome [169], or generating textual explanations [15]. One might think such explanations are interpretable for humans as they rely on their ability to understand and reason using natural language. Moreover, given their simplicity and (apparent) intuitiveness, one might not question such explanations. However, they may sometimes be too complex [95, 123], or their structure may not be intuitive or representative enough for a human interpreter to understand the model's behaviour promptly. For example, saliency map scores assigned to the words of a sentence in a sentiment analysis task may not be fully understood, as these scores represent which parts of the input were deemed essential and do not represent the actual model's reasoning [378]. Similar representations (*i.e.*, highlights and textual explanations) are employed when collecting human knowledge to train or improve models and evaluate their explanations [479] as these are pretty simple for humans to describe. Crowd-based methods usually focus only on labelling results with relevant human-generated tags, explicitly identifying objects, actions, or other elements in the output. Therefore, potential explanations only refer to the data elements and the model parts that produce them. The cognitive process applied by the human to perform the task is completely neglected. This means there is no alignment or reconciliation between how a person would solve a task and how the machine learning model generates the results.

Striving to provide a complete and structured representation of human rationale, we propose ad hoc formalizations to organize human knowledge by drawing inspiration from Argumentation Mining [286, 323] and the recent literature in Data Structuring in XAI for a set of NLP tasks of interest, *i.e.*, Sentiment Analysis, Text Summarization, Natural Language Inference, Claim Verification, and Question Answering. We analyzed and organized these tasks based on their type (*i.e.*, text classification or generation) and the number of inputs (*i.e.*, single or multiple inputs) as we identified them as fundamental discerning factors. We inspected the processes and the nature of the considered NLP tasks from the human and model perspectives and designed the formalizations. These are referred to as *rationale mappings*. A standard structure, referred to as *Rationale Mapping*, is described and further characterized for each task to reduce complexity and enhance expressiveness. These are hierarchically organized in tree structures, referred to as *Rationale Trees*. Such representations organize and detail the reasoning steps humans apply when identifying and reasoning on the essential parts of the texts they are provided with.

Building on this characterization, we present a web-based, human-centred approach to collect rationale mappings for a subset of the considered NLP tasks, *i.e.*, Sentiment Analysis, Text Summarization, and Question Answering. The literature about these tasks was analyzed in detail and reported. Although a few tasks were not consid-

ered, we deem the performed research to be representative of the complexity of the considered spectrum of NLP tasks. We describe the design of the human-computer interaction paradigm, the data collection process specification, its implementation as a crowdsourcing web application, and its validation with experimental studies ascertaining its reliability and effectiveness. User feedback on the process and its complexity is collected. The usability and workload of the application are assessed through standardized user questionnaires. A preliminary dataset is built and openly shared with the research community.

5.2 Context-specific Related Works & Background

Rationale Mappings and Trees are applied to three NLP tasks of interest (*i.e.*, Sentiment Analysis, Text Summarization, and Question Answering) to organize human rationale for XAI approaches.

5.2.1 Human Knowledge and Reasoning in NLP and XAI

Natural Language Processing (NLP) is a research field aimed at interpreting, analyzing, and manipulating natural language data to learn, understand and produce human language content [179, 211]. NLP include a broad variety of tasks, some aimed at human language understanding (*e.g.*, Coreference Resolution, Natural Language Parsing, etc.), while more complex tasks focus on classifying (*e.g.*, Sentiment Analysis, Natural Language Inference, etc.) or generating (*e.g.*, Text Summarization, Question Answering, etc.) text. In language-based tasks, humans can reason on and understand the provided text(s) through their inherent linguistic knowledge to generate a desired output. Such data are usually collected as couples of input-output texts and employed to train models capable of achieving specific tasks [366]. However, such a data collection approach does not include how humans performed the task and reasoned on the provided text(s). In particular, whenever model explainability is desired, human actors are also requested to describe or provide evidence of the thought process they applied. These are usually collected as free text or highlights of the input's words and sentences. In particular, while the first is more expressive and readable, the latter provides a compact, sufficient, and comprehensive representation [479]. Such information can be used to train so-called self-explainable models [12], *i.e.*, models capable of providing explanations for their outputs [479], or to assess the explanations extracted through other XAI techniques [23, 479]. Even though explanations are mainly collected through crowdsourcing approaches using the representations mentioned above, a wider variety of formats is available whenever an explanation is provided to a human interpreter. In the context of Explainable AI, NLP tasks' explanations are represented as saliency maps [397], declarative representations (*i.e.*, trees and rules) [337], examples [161], or machine-generated natural language [255]. While lay users can easily understand the latter [95], the others may not be directly understandable to human interpreters since a deep understanding of XAI may be required [95, 123]. Despite the similarities in the shape explanations provided by and provided to humans, there is a significant gap regarding their interpretability. Furthermore, although some explanations may be humanly understandable, they are not structured to match human reasoning. Hence, a misalignment

between how humans think when performing a natural language task and how explanations provide evidence for the model's reasoning can be identified. We argue such a difference must be addressed across the whole explainability process, starting from the data collection and structuring steps. While several techniques exist to provide explanations for NLP models, *e.g.*, saliency-based approaches, declarative representations, and natural language, they mostly rely on the (sometimes improper) assumption the human receiving the explanation will interpret it as intended [95]. However, this assumption has been proven not to hold in several cases [378].

The following NLP tasks are analyzed in depth as they were involved in the data collection step. Their features, how their explanations are provided, and how humans are involved and perform the task are described.

Sentiment Analysis (SA) determines whether subjective text (*e.g.*, people's opinions, thoughts, etc.) conveys a positive, negative, or neutral view [432, 475]. It can be applied at different levels of granularity (*i.e.*, document-level, sentence-level, or aspect-based) [277]. When performing the task, human interpreters would identify all the portions of the text expressing subjective opinions and subsequently assess and combine their views to derive the overall sentiment of the text. Human rationale describing the sentiment attribution can be represented by associating the output sentiment with the parts of the input text that most influenced the output label. Concerning datasets, highlights [102, 508], text snippets [102, 471, 508], and structured content [406, 471] were employed to organize human rationale for explainable NLP. The data collection process mainly involves crowd workers picking the sentences or words that contribute to the final sentiment and sometimes explaining their choices.

Text Summarization (TS) generates a summarized version of a given text [117]. Two main approaches exist, *i.e.*, extractive and abstractive [117, 490]. In extractive approaches, the most important sentences are directly reported in the summary. In abstractive approaches, essential pieces of information are syntactically reshaped into semantically equivalent sentences to produce a summary. While the first can be easily achieved from a model (and human) perspective, the latter requires complex text processing and understanding capabilities [137]. When performing the task, human interpreters would identify and summarise the most critical information in the text. Rationale can be represented by mapping its informative content to where it is extracted from the input text. Concerning datasets, Kim et al. [214] collected annotations by matching input sentences with their corresponding output and assigning labels to highlight their importance.

Question Answering (QA) provides a relevant answer to a question given a paragraph or a set of documents containing relevant information for answering it [58]. Answers can be provided in different forms, *e.g.*, an extract from an input text or a newly generated one. When performing the task, human interpreters would understand the type of information they must find in the paragraph, subsequently developing the answer by inspecting the provided content. Rationale can be represented by extracting which paragraph(s), sentence(s), or sub-sentence(s) are meaningful to

the question or contain the answer. Concerning datasets, rationales have been collected by having crowd workers highlight sentences in documents to answer questions [145, 497, 501]. Furthermore, they were sometimes associated with textual explanations [344, 351, 504]. More expressive approaches involve structured content like graphs [198, 484] or other complex structures [196, 234]. Common data collection approaches involve question [501] and answer generation [497], highlights [234, 504], rationale description [344,504], or complex guided procedures [145,196,198,351,484].

Coreference Resolution (CR) determines all the expressions that refer to the same real-world entity in a text. Coreference resolution enables or improves the performance of other tasks, such as Text Summarization, Question Answering, Information Retrieval, and Language Translation [425].

5.2.2 Data Structuring in NLP and XAI

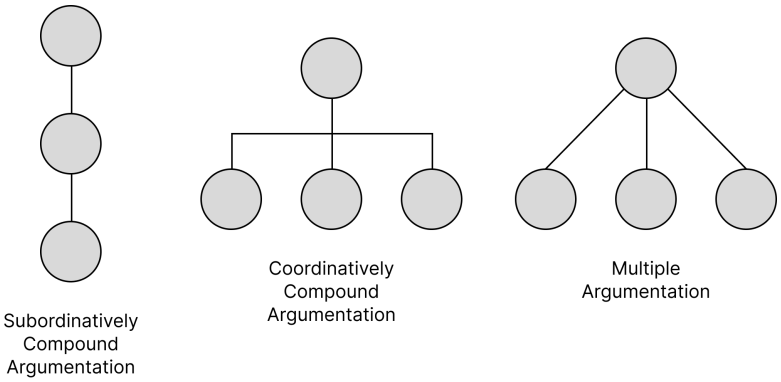
In the literature, several approaches for structuring NLP-related data were identified, *e.g.*, labelled relationships between words.

Over the last few years, various datasets organized human knowledge applied to the explainability of NLP tasks in the form of free-text [223], highlights of the most important words or sentences [102, 406], or a combination of both [61]. Although such simple structures were proven effective, researchers demonstrated that enhanced detail improves models' performances [210, 504] and understandability [234]. Most structures have been designed in the context of Question Answering as it is one of the most complex NLP tasks. Lamm et al. [234] defined annotation triples for Question Answering tasks by identifying relationships between the question and the provided passage. The annotator selects the passage entailing the answer, then chooses a short text span with the answer within the entailed text and marks the equivalent noun phrases in the question and the answer. Finally, entailment patterns are extracted. In the context of machine reading comprehension, Ye et al. [504] defined quadruples of question, paragraph, answer, and a textual explanation that motivates the human reasoning applied to build the annotation. WorldTree [198] and WorldTree V2 [484] are explanation graphs that motivate answers to science questions. They are built by defining and labelling relationships between words in the question, answer, and explanations generated through domain and world knowledge. Although the described processes and structures significantly advance the state-of-the-art in Question Answering in the corresponding contexts, these are task-specific, and their alignment with human reasoning has yet to be proven.

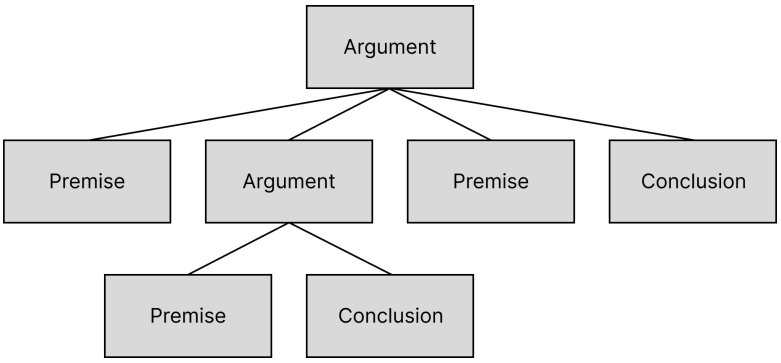
5.2.3 Argumentation Theory & Mining

Argumentation Theory describes tree-like structures for organizing texts as arguments, while Argumentation Mining describes their collection and structuring following the theory.

Argumentation Theory is a research field examining the way arguments are pro-



(a) *Pragma-Dialect Theory*



(b) *Argumentation Tree Example*

Figure 5.1: *(a) The structures proposed described by the Pragma-Dialect theory for argumentation. (b) An argumentation tree structure proposed by Mochales et al. [286]. Each argument is supported by one or more premises and a conclusion and can be premises for other arguments.*

duced, analyzed, and evaluated. While several approaches aim at establishing criteria for defining reasonable argumentations [34], our research mainly focuses on argumentation structuring, *i.e.*, how an argument can be represented regarding the relationship between its premises and conclusions. Argumentation structures allow for identifying argument schemes, *i.e.*, stereotypical patterns in human reasoning, to determine typical critical questions that could be used to counter such an argument. The basic argumentation unit is an *argument* whose structure involves implicit or explicit premises and a conclusion or, more generally, a set of at least two propositions [286]. For each *argument*, a schema defining relations between prepositions following human reasoning patterns is defined. Typical argument structures identified in Informal Logic are serial, linked, convergent, and divergent structures [328]. In linked structures, multiple premises are necessary to support a conclusion effectively. On the other hand, in convergent structures, multiple premises independently support a conclusion. In divergent structures, a single premise leads to more than one conclusion. Finally, serial structures aim to further develop an argument by supporting one of the premises. Likewise, Pragma-Dialects theory [456] describes argumentation structures that represent the relation between arguments through coordination, subordination, or forming multiple arguments, as depicted in Figure 5.1(a).

Argumentation Mining is the process of detecting arguments in a textual document,

their relationships, and their internal structure [65, 323]. Typically, Argumentation Mining involves text segmentation, argument/non-argument classification, and defining simple and refined structures for the identified argumentation. The first step consists of breaking down the text into atomic units, usually clauses or sentences [237], which are then classified as argumentative or non-argumentative. Then, argumentation structures are employed to define formalisms and organize the retrieved argumentative snippets as premises or conclusions. While a more complex graph structure is usually employed to represent the relationships among these elements [65], Mochales et al. [286] applied the Pragma-Dialects theory to define a tree-structure representation in which every tree and sub-tree represents a single argumentation structure. In such a setting, all arguments are uniquely related to another argument of a tree representing a premise. Figure 5.1(b) represents and examples of such a structure.

5.3 Formalization

5.3.1 NLP Task Classification

NLP tasks of interest were classified based on their features (*e.g.*, task type, number of inputs, and more) to refine Rationale Mappings and Trees.

This research maps and generalizes the concepts described to fit various NLP tasks, structuring representations capable of representing human knowledge and reasoning. We consider five different Natural Language Processing tasks: Sentiment Analysis, Text Summarization, Natural Language Inference, Claim Verification, and Question Answering. We identified which features make these tasks substantially different (*e.g.*, objective, solving process, number of inputs, type of output, etc.). Considering such differences, our research acknowledged the similarity in the nature of the inputs (*i.e.*, all these tasks accept free-text inputs) and the number of outputs (*i.e.*, all these tasks take a single output) while identifying significant differences in the process, the type of task (*i.e.*, whether the task generates or classifies text), and the number of inputs (*i.e.*, whether the task takes one or multiple inputs). While the process is unique for each considered NLP task, the type and number of inputs can define a categorization. Table 5.1 reports the outcome of this classification.

Task	Task Type	N Inputs	Input(s) Type	Output Type
Sentiment Analysis	Classification	Single	Free-text	Discrete
Text Summarization	Generation	Single	Free-text	Free-Text
Natural Language Inference	Classification	Multiple	Free-text	Discrete
Claim Verification	Classification	Multiple	Free-text	Discrete
Question Answering	Generation	Multiple	Free-text	Free-text

Table 5.1: *A tabular representation classifying each NLP task of interest based on the features.*

5.3.2 Rationale Mappings

Rationale Mappings are defined as triples of < text, text, label >. These are further specialized and refined based on the thought process applied and the NLP tasks.

When reasoning on text, humans are so used to finding logical, syntactical, and semantical connections between words that they are unaware of such behaviour. A simple example is the capability of humans to find all the expressions that refer to the same entity in a text (so-called Coreference Resolution in NLP). Such a task rarely requires complex human reasoning as it can be promptly achieved thanks to the linguistic flexibility and knowledge we have developed. On the other hand, extensive human reasoning may be necessary for complex language-based activities, like Question Answering, in which a human interpreter must understand the paragraph, the question, and the relations between their content to perform the task. Such reasoning is a fundamental building block for defining and structuring human rationale in Natural Language Processing tasks. We refer to them as *rationale mappings*, *i.e.*, representations that organize humans' analytical reasoning steps when identifying and associating the essential parts of the texts involved in a language-based task leading to its output. In particular, we characterize three types of mappings standard to the considered NLP tasks:

- *External mappings* represent the reasoning a human interpreter applies between two terms and/or parts of the text belonging to **different** texts in an NLP task (*e.g.*, the question and the paragraph in a QA task).
- *Internal mappings* represent the reasoning a human interpreter applies between two different terms and/or parts of text in the **same** text in an NLP task (*e.g.*, the paragraph in a QA task).
- *Resolution mappings* are *internal mappings* representing anaphora or coreference resolution reasoning between two terms and/or parts of text in the **same** text in an NLP task.

We define the structure of *rationale mappings* by combining the literature about data structuring in XAI and the argument structure described by the Pragma-Dialects theory [286]. In our definition, we constrained the number of propositions and extended it with their relationship, finally merging them into a single representation. Hence, we define *rationale mappings* as triples

$$\langle \ \textit{text}, \textit{text}, \textit{label} \ \rangle$$

where *text* is a word or a set of consecutive words from any text involved in the human reasoning applied to the language-based task and *label* is a term that defines the relationship between the *texts*. The latter is defined based on the type of mapping. In *external mappings*, they are specific to the NLP task to which the mappings are applied, *i.e.*, when the task involves a discrete output (*i.e.*, a finite and well-defined set of outputs is possible) or specific terms that describe the applied approach, these are employed as labels as they represent both human- and model-understandable concepts. Otherwise, more generic linguistic labels are considered, *i.e.*, *semantic* or *syntactic*, respectfully representing the semantic or syntactic similarity between texts. Whenever a *semantic*

label is applied, mappings can be extended to include a textual description of the rationale a human interpreter applies, enhancing the level of detail. Such generic labels are also applied to *internal mappings* as they define a syntactic or semantic relationship between the texts involved.

When designing *rationale mappings*, some elements were repeated when some conditions were verified, finally leading to a simplified notation. Indeed, *External mappings* can be simplified in specific cases, hence defining these *mappings* as couples

$$\langle \ \textit{text}, \textit{label} \ \rangle$$

where *text* is a word or a set of consecutive words from any text involved in the human reasoning applied to the language-based task and *label* is a term that specifies the *text*. In particular, we consider simplifications only when the nature of the task and the labels allow for them. The fact that the two texts in a mapping coincide is not considered a simplification, even though it might be helpful for what concerns data storage. *Internal mappings* are not subject to any simplifications as there is no meaningful overlap between *texts* and *label*. However, they may be subject to slight changes to improve their expressiveness when applied to specific tasks. *Resolution mappings* can not be simplified as it is necessary to specify the type of resolution used and the parts of text involved. Further clarifications will be made for each considered NLP task in their dedicated sections.

5.3.3 Rationale Trees

> Rationale Trees are defined as tree structures built from Rationale Mappings following a set of rules ensuring that the deeper the node in the tree, the more detailed the rationale they describe.

While defining *rationale mappings* is useful to understand the human reasoning applied to a task, these can be further hierarchically structured to describe the complete rationale involved in a specific task instance. Hence, *rationale mappings* are organized in a tree structure in which each mapping is a tree node whose meaning and constraints depend on its type. We refer to these structures as *rationale trees*.

In these structures, the root node represents the (generic) relationship between the input(s) and the output, *i.e.*, a standard input-output representation of the task. Each other node (*i.e.*, internal nodes and leaves) further details the mapping between the texts in its parent node. In particular, considering a parent node p and its child node c defined as

$$p \quad \langle \ p\_text\_I, p\_text\_II, p\_label \ \rangle$$
$$c \quad \langle \ c\_text\_I, c\_text\_II, c\_label \ \rangle$$

and assuming that their corresponding *texts* (*i.e.*, *p_text_I* and *c_text_I*, and *p_text_II* and *c_text_II*) are extracted from the same text, either one of the following constraints is enforced.

- **Paired Text Detailing.** *c_text_I* $\subset$ ***p_text_I***, *i.e.*, *c_text_I* is a word or a set of consecutive words that are a subset of *p_text_I*, **and** *c_text_II* $\subset$ ***p_text_II***, *i.e.*, *c_text_II* is a word or a set of consecutive words that are a subset of *p_text_II*.

- **Individual Text Detailing.** *c_text_I* $\subset$ ***p_text_I***, *i.e.*, *c_text_I* is a word or a set of consecutive words that are a subset of *p_text_I*, **and** *c_text_II* $\subset$ ***p_text_I***, *i.e.*, *c_text_II* is a word or a set of consecutive words that are a subset of *p_text_I*. The same can be applied considering *p_text_II*.

Such conditions define a structure in which the deeper the node, the more specific the rationale it describes. Furthermore, while *external* and *internal mappings* can either be internal nodes or leaves that detail the parent node's rationale, *resolution mappings* can only be leaf nodes and define rationale to be applied to their parent and sibling nodes whenever meaningful. Child nodes are considered to be in a coordinative relationship towards their parent node, simultaneously contributing to specifying the parent's node mapping. Moreover, while *external mappings* can have both *internal* and *external mappings* as child nodes, *internal mappings* can only have other *internal mappings* as child nodes since they are mainly employed to detail the rationale applied in *external mappings* and not vice-versa. Additionally, *resolution mappings* can be child nodes for both *internal* and *external mappings*. A generic example of a *rationale tree* is depicted in Figure 5.2.

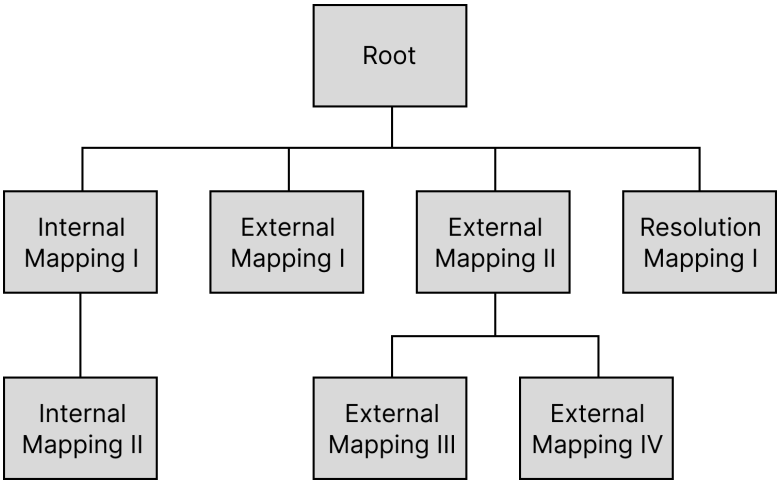


Figure 5.2: *A generic rationale tree structuring the rationale mappings following the described rules.*

The only condition enforced between sibling nodes is that their *text_I* and *text_II* should not completely overlap simultaneously, *i.e.*, considering any pair of sibling nodes *s1* and *s2* defined as

$$s1 \quad \langle \; s1\_text\_I, s1\_text\_II, s1\_label \; \rangle$$
$$s2 \quad \langle \; s2\_text\_I, s2\_text\_II, s2\_label \; \rangle$$

and assuming that their corresponding *texts* (*e.g.*, *s1_text_I* and *s2_text_I*, and *s1_text_II* and *s2_text_II*) are extracted from the same text, the following constraints are enforced.

- **Non-Overlapping Text II.** If $s1\_text\_I \equiv s2\_text\_I \Rightarrow s1\_text\_II \neq s2\_text\_II$.
- **Non-Overlapping Text I.** If $s1\_text\_II \equiv s2\_text\_II \Rightarrow s1\_text\_I \neq s2\_text\_I$.

Such conditions define a structure where the same mapping can not be duplicated, although they still allow a fine granularity in the differences between the mappings associated with a parent node.

Task	Labels	Simplification
Sentiment Analysis	Positive, Negative	Yes
Text Summarization	Extractive, Abstractive	Yes
Natural Language Inference	Neutral, Contradiction, Entailment	No
Claim Verification	Support, Refute	No
Question Answering	Syntactic, Semantic	No

Table 5.2: *A tabular representation summarizing some of the features of each NLP task of interest.*

The following sections describe the formalizations, detailing a set of features of interest. In particular, the simplest ones are summarized in Table 5.1, while the most complex ones are outlined in Table 5.2 and further detailed in the corresponding sections.

- **Labels**, *i.e.*, the concepts applied as labels when defining the mappings. These are mainly employed in *external mappings*, although *internal mappings* may sometimes benefit from such labels too.
- **Mappings Interpretation**, *i.e.*, a task-specific description for *internal* and *external mappings*, if needed. Whenever no specific interpretation is provided, we consider them aligned with their general description.
- **Simplifications**, *i.e.*, whether any simplification can be applied to the mappings, their description and structure.
- **Mapping Guidelines**, *i.e.*, the process a human interpreter applies to define *mappings* and a *rationale tree* for a task of interest. We consider human interpreters to be performing the task themselves, although we do not include details of such a process in the guidelines. The same approach can be applied even when the interpreter is provided with all the texts involved in the task.
- **Example**, *i.e.*, an example of a *rationale tree* collected by applying the process to a task entry from a specified NLP dataset. Each mapping type is identified by its starting letters (*e.g.*, EM stands for *external mapping*).

5.3.4 Sentiment Analysis

Sentiment Analysis is an NLP task in which a human interpreter defines the output by assigning a sentiment (either *positive*, *negative*, or sometimes *neutral*) to an input sentence or text. For such a task, generic *internal mappings* are defined as

$$\langle \ \textit{input\_text}, \textit{input\_text}, \textit{label} \ \rangle$$

where *input_text* is a word or a set of consecutive words from the input text and *label* is the sentiment between the *texts* (*i.e.*, *positive* or *negative*, in our case). On the other hand, *external mappings* are defined as

$$\langle \ \textit{input\_text}, \textit{output\_text}, \textit{label} \ \rangle$$

where *input_text* is a word or a set of consecutive words from the input text, and *output_text* and *label* represent the sentiment associated with the *input_text*. Acknowledged the overlapping between *output_text* and *label*, *external mappings* are applied a simplification. Hence, they are defined as

$$\langle \; input\_text, label \; \rangle$$

where *input_text* is a word or a set of consecutive words from the input text and *label* is the sentiment associated with *input_text*.
A human interpreter providing *rationale mappings* for a Sentiment Analysis task performs the following assignments.

- They define *external mappings* between the input and the output texts.
- For each of the previously defined *external mapping*, they recursively define *internal* and *external mappings* detailing the texts involved until a desired level of detail is achieved. The same process is applied to the newly found *mappings*.
- For each of the previously defined *internal* and *external mappings*, they define whether any *resolution mapping* was applied. These are defined as child nodes or sibling nodes based on where they are used.

We picked a data point from the Large Movie Review Dataset [264] and built its *rationale tree* as an example, represented in Figure 5.3.

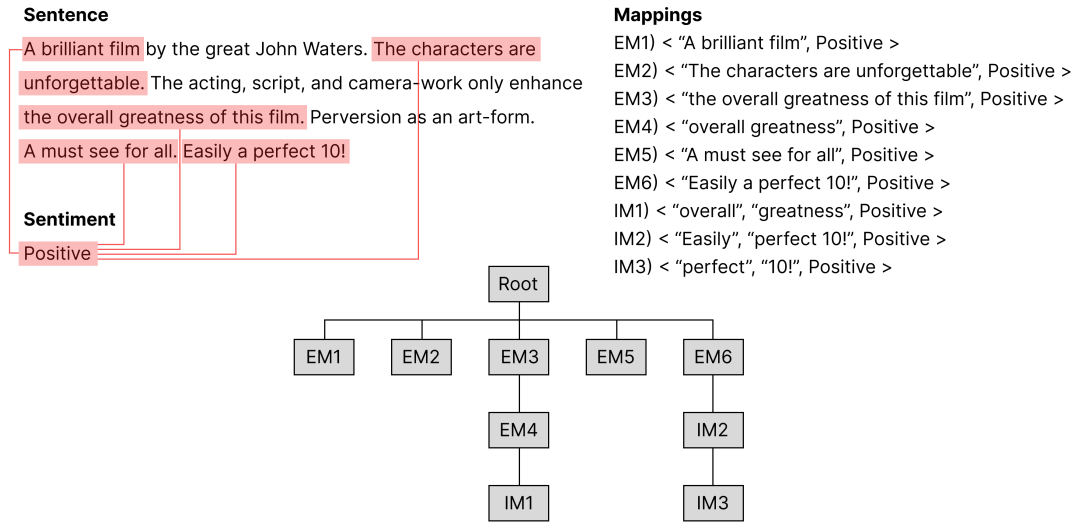


Figure 5.3: *A rationale tree structuring the mappings of a chosen data point from the Large Movie Review Dataset. Internal mappings were omitted for clarity purposes.*

5.3.5 Text Summarization

Text Summarization is an NLP task in which a human interpreter is provided with an input text, and they give a summarized output text. Two different approaches can be applied and combined. An *extractive* approach reports parts of the input text into the output text, maintaining the same syntax. An *abstractive* approach formulates the output text to have the same semantics as parts of the input text while using a different syntax. For such a task, generic *external mappings* are detailed as

$$\langle \; input\_text, output\_text, label \; \rangle$$

where *input_text* is a word or a set of consecutive words from the input text, *output_text* is a word or a set of consecutive words from the output text, and *label* is the summarization approach (*i.e.*, *abstractive* or *extractive*) applied to *input_text* to generate *output_text*.
Whenever an extractive approach is applied, *external mappings* can be simplified as such an approach involves reporting the exact text from the input in the output text. Hence, they are defined as couples.

$$\langle \ \textit{input\_text}, \textit{“extractive”} \ \rangle$$

where *input_text* is a word or a set of consecutive words from the input text.
A human interpreter providing *rationale mappings* for a Text Summarization task performs the following assignments.

- They define *external mappings* between the input and the output texts.
- For each of the previously defined *external mapping* that is assigned the *“extractive”* label, they recursively define *internal mappings* detailing the texts involved until a desired level of detail is achieved. Instead, for each of the previously described *external mapping* that is assigned the *“abstractive”* label, they recursively define *internal* and *external mappings* detailing the texts involved until a desired level of detail is achieved. The same approach is applied to the newly found *mappings*.
- For each of the previously defined *internal* and *external mappings*, they define whether any *resolution mapping* was applied. These are defined as child nodes or sibling nodes based on where they are used.

We picked a data point from the CNN/Daily Mail Dataset [298] and built its *rationale tree* as an example, represented in Figure 5.4.

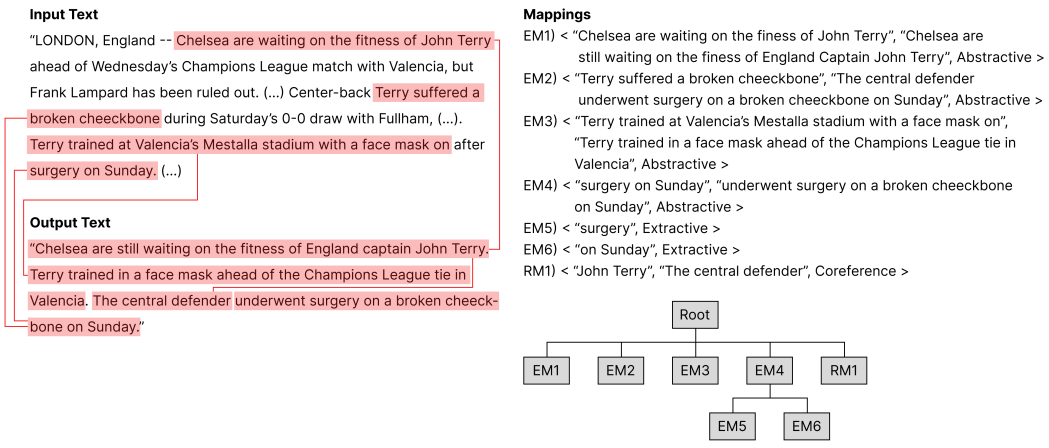


Figure 5.4: *A rationale tree structuring the mappings of a chosen data point from the CNN/Daily Mail Dataset Dataset. Only one external mapping was refined for clarity purposes. Similarly, part of the input text that was not deemed useful was omitted.*

5.3.6 Natural Language Inference

Natural Language Inference is an NLP task in which a human interpreter is provided with two texts, an *hypothesis* and a *premise*, and they define whether they are in an *entailment*, *contradiction*, or *neutral* relationship. For such a task, generic *external mappings* are defined as

$$\langle \; premise\_text, hypothesis\_text, label \; \rangle$$

where *premise_text* is a word or a set of consecutive words from the premise, *hypothesis_text* is a word or a set of consecutive words from the hypothesis, and *label* is the relationship (*i.e.*, *entailment*, *contradiction*, or *neutral*) between *premise_text* and *hypothesis_text*.
A human interpreter providing *rationale mappings* for a Natural Language Inference task performs the following assignments.

- They identify *external mappings* between the premise and the hypothesis texts.
- For each of the previously defined *external mappings*, they recursively define *internal* and *external mappings* detailing the texts involved until a desired level of detail is achieved. The same approach is applied to the newly found *mappings*.
- For each of the previously defined *internal* and *external mappings*, they define whether any *resolution mapping* was applied. These are defined as child nodes or sibling nodes based on where they are used.

We picked a data point from the e-SNLI Dataset [61] and built its *rationale tree* as an example, represented in Figure 5.5.

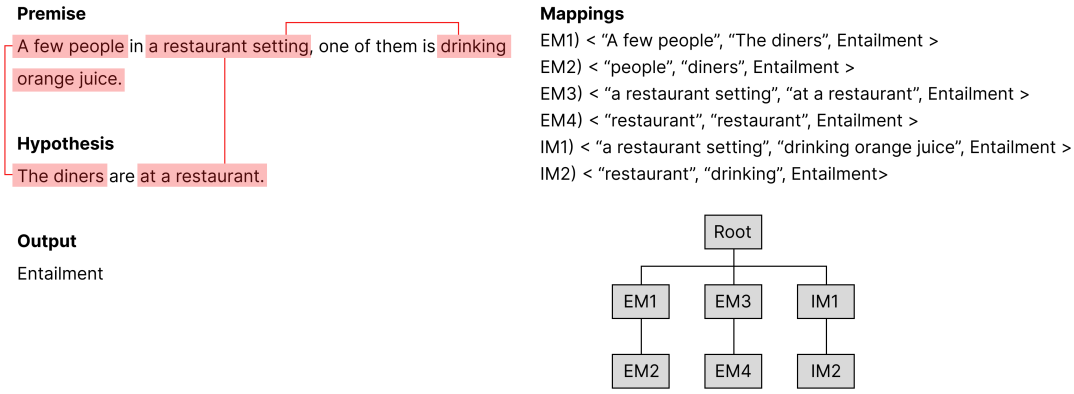


Figure 5.5: *A rationale tree organizing the mappings of a chosen data point from the e-SNLI Dataset.*

5.3.7 Claim Verification

Claim Verification is an NLP task in which a human interpreter is provided with two texts, *i.e.*, a *claim* and an *evidence*, and they define whether the *evidence supports* or *refutes* the *claim*. For such a task, generic *external mappings* are defined as

$$\langle \; claim\_text, evidence\_text, label \; \rangle$$

where *claim_text* is a word or a set of consecutive words from the claim, *evidence_text* is a word or a set of consecutive words from the evidence, and *label* is the relationship (*i.e.*, *support* or *refute*) between *claim_text* and *evidence_text*.
A human interpreter providing *rationale mappings* for a Claim Verification task performs the following assignments.

- They identify *external mappings* between the claim and the evidence.
- For each of the previously defined *external mappings*, they recursively define *internal* and *external mappings* detailing the texts involved until a desired level of detail is achieved. The same approach is applied to the newly found *mappings*.
- For each of the previously defined *internal* and *external mappings*, they define whether any *resolution mapping* was applied. These are defined as child nodes or sibling nodes based on where they are used.

We picked a data point from the FEVER Dataset [444] and built its *rationale tree* as an example, represented in Figure 5.6.

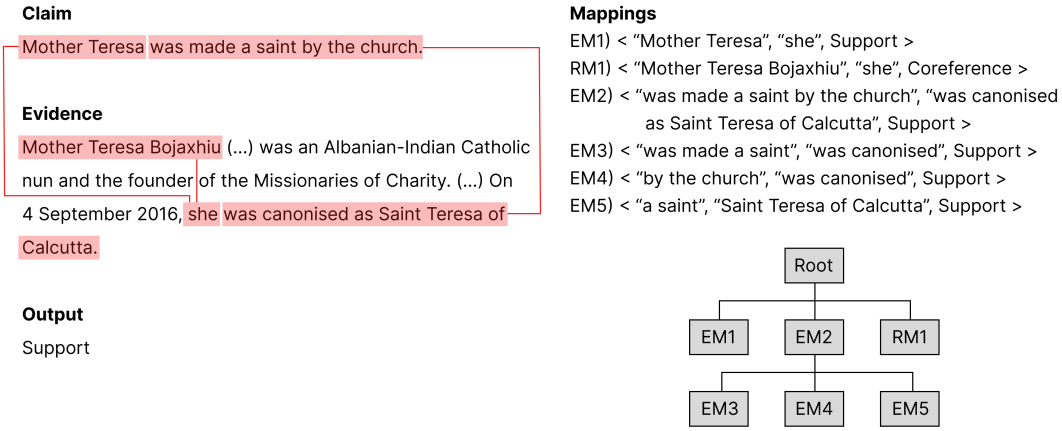


Figure 5.6: *A rationale tree structuring the mappings of a chosen data point from the FEVER Dataset. The evidence defined in the dataset and collected from Wikipedia was represented, removing the text that was not deemed useful for clarity purposes.*

5.3.8 Question Answering

Question Answering is an NLP task in which a human interpreter is provided with a *question* and a *paragraph*, and they give an *answer* to the *question* through the *paragraph*. For such a task, generic *mappings* are defined as

$$\langle \; text, text, label \; \rangle$$

where *text* is a word or a set of consecutive words from the same (in *internal mappings*) or different (in *external mappings*) texts, *i.e.*, the *question*, the *paragraph*, or the *answer*, and *label* describes whether there is a *semantic* or *syntactic* relationship between the *texts*. Similarly to *internal mappings*, whenever a *semantic* label is applied, the mapping is further detailed by collecting comments detailing the relationship between

the *texts*.
Rationale trees increase complexity in Question Answering tasks as the process is more convoluted than the other considered NLP tasks. First of all, a new type of mapping has to be defined. *Abstractive mappings* define which word or set of consecutive words of the question contributed to defining its class among the following question types.

- Yes/No Question, *i.e.*, questions looking for confirmation in the paragraph.
- Wh-Question, *i.e.*, questions looking for the answer based on the type of wh-question (*e.g.*, Who, What, etc.).
- Choice Question, *i.e.*, questions looking for the answer among the ones proposed in the question based on the paragraph.
- Disjunctive Questions, *i.e.*, questions looking for confirmation in the paragraph.

Such mappings are introduced to be aligned with the question-answering process in which a human interpreter identifies which information they should look for to answer the question before reading the paragraph [59, 209, 360]. *Abstractive mappings* are defined as couples

$$\langle \ \textit{question\_text}, \textit{question\_class} \ \rangle$$

where *question_text* is a word or a set of consecutive words from the question and the *question_class* describes the question class chosen from a list of values defined from the question types described, *i.e.*, *yes/no question*, *disjunctive question*, *choice question*, and *wh-question*. The latter is further detailed based on the type of wh-question, defining a *specialization* (described in Table 5.3). Moreover, each *rationale tree* can only have one *abstractive mapping* and must be a child node of the root node.

Specialization	Wh-Question Keywords
Person	Who, Whose, Whom
Information	What, How
Location	Where
Time	When
Reason	Why, What for, How come, Why do not
Quantity	How many, How much, How far, How long, etc.
Choice	Which, Whom

Table 5.3: *A table summarizing the specializations for the class of wh-questions. For each specialization, a list of keywords identifying the wh-question is provided.*

A human interpreter providing *rationale mappings* for a Question Answering task performs the following assignments.

- They define an *abstractive mapping* associated with the question.
- They define *external mappings* between the question and the paragraph. The same is done for *internal* and *resolution mappings* in these texts. These are recursively refined until the desired level of detail is achieved.
- They define *external mappings* between the paragraph and the answer. The same is done for *internal* and *resolution mappings* in these texts. These are recursively refined until the desired level of detail is achieved.

- They detail the previously defined *abstractive mapping* by defining *external mappings* between the question and the answer. These are recursively refined until the desired level of detail is achieved.

We picked a data point from the SQuAD 2.0 Dataset [345] and built its *rationale tree* as an example, represented in Figure 5.7.

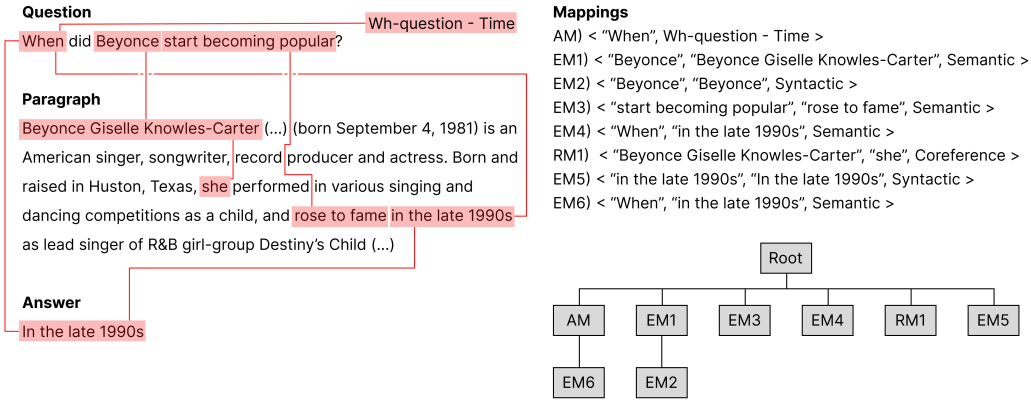


Figure 5.7: *A rationale tree structuring the mappings of a chosen data point from the SQuAD 2.0 Dataset.*

5.4 Rationale Trees Data Collection Approach

An approach for collecting Rationale Trees is described, involving a three-step process including sentence-level, sub-sentence-level, and word-level steps to generate the triples. These are further combined into Individual Rationale Trees which in turn are merged into Complete Rationale Trees.

This section introduces an approach to collecting human rationale to build *rationale trees*. These are collected for the three NLP tasks of interest, *i.e.*, Sentiment Analysis, Text Summarization, and Question Answering, based on the tasks' characteristics and intended human meaning. The data collection step involves human actors in creating Rationale Mappings. These will then be organized into *individual rationale trees*, *i.e.*, data structures built by a single participant. Ultimately, the collected trees are merged into *complete rationale trees*, *i.e.*, data structures built by combining multiple *individual rationale trees*. Multiple *rationale mappings* and *individual rationale trees* are collected for each data point. On the other hand, only one *complete rationale tree* is provided for each. Figure 5.8 illustrates such a process.

Assumptions. While the considered formalization includes three types of mappings, our methodology focuses on detailing *external mappings* and *resolution mappings*, excluding *internal mappings* from the actual collection. This choice was made since careful inspection of well-known datasets revealed that such a level of detail is typically not covered by explanations. While such mappings might contribute to capturing

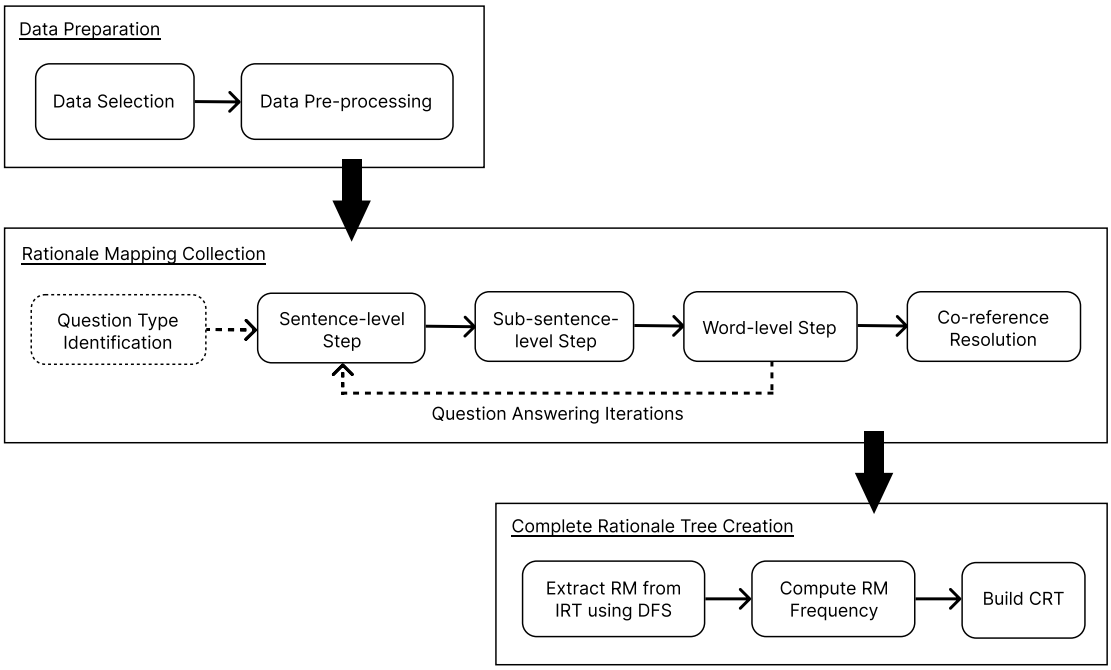


Figure 5.8: *A schematic representation of the process to generate Rationale Trees.*

human rationale at its fullest, *internal mappings* would introduce further complexity and a higher risk of human errors. Furthermore, we provide users with all the task elements required to create *rationale mappings* without having them perform the NLP task first.

Data Preparation. Before collecting *rationale mappings*, it is necessary to choose a set of data points for each task, *i.e.*, the texts associated with an NLP task. Such data must be complete enough to describe a task instance, *i.e.*, including input(s) and output, or these must be derivable from the data. Furthermore, the task generated from the data point must have an output (*e.g.*, questions not allowing an answer in the provided paragraph must be discarded as no mappings could be extracted). Question Answering data instances must satisfy an additional constraint, *i.e.*, multiple questions cannot be asked in the same text, as its formalization only allows for a single *abstractive mapping* for each *rationale tree*. Instances including multiple questions can be dropped or properly split into multiple valid instances, leading to multiple data points with one question each. Finally, additional pre-processing and text-cleaning operations may be needed based on the specific characteristics of the chosen dataset.

Rationale Mappings Collection. Before defining *rationale mappings* for a given data instance, participants are provided with a theoretical description of the applied formalization, followed by guided exercises to strengthen their understanding of the activity. Each guided exercise is partially pre-compiled to show how the task should be performed and includes a sample solution users can use to assess the validity of their mappings. After correctly completing these exercises, participants proceed with the actual data collection activity. An example of the *rationale mapping* collection process is displayed in Figure 5.9. The annotation process involves a sequence of four *ratio-*

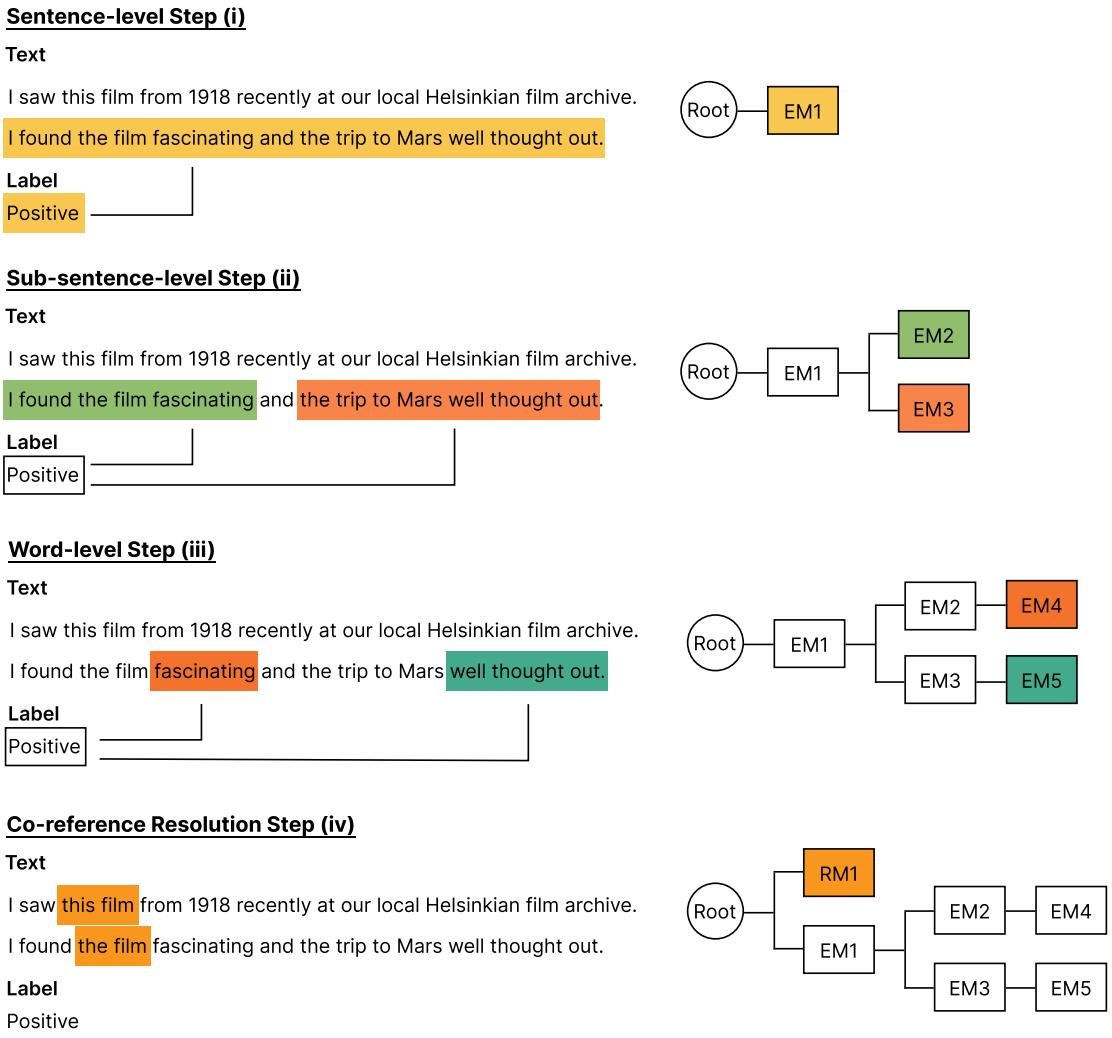


Figure 5.9: *Rationale Mapping Collection Process for Sentiment Analysis.*

nale mapping creation steps: a sentence-level step (i), a sub-sentence-level step (ii), a word-level step (iii), and a final coreference resolution step (iv). While the latter allows for collecting *resolution mappings*, the others guide the user in providing *external mappings* with different levels of detail. Such a process is the same for all tasks besides Question Answering. In particular, an additional initial step to define the *abstract mapping* and three iterations of the first steps (i-iii), one for each couple of input(s) and output (*i.e.*, question-paragraph, paragraph-answer, and question-answer) is performed. During these steps, participants are asked to select the texts to be included in the mappings. In particular, when performing the sentence-level creation step (i), the texts defined in the mapping are complete sentences from the content involved. The texts extracted in the sub-sentence-level creation step (ii) are slices from the content chosen in the previous step (i). In the word-level creation step (iii), the texts involved in the mappings are single words taken from the sub-sentences defined in the sentence-level step (i). Potential duplicates may arise across steps (ii) and (iii) as words can be seen as instances of simple sub-sentences. If so, only a single instance of a mapping is kept. Concerning the coreference resolution step (iv), participants are asked to pick

Algorithm 1 Rationale Tree Creation Algorithm

```
1: procedure ADDNODE(nodeToAdd, currentNode, siblings, parentNode)
2:     if isAncestor(currentNode, nodeToAdd) then
3:         if currentNode.children is empty then
4:             currentNode.children.push(nodeToAdd)
5:         else
6:             return ADDNODE(nodeToAdd,currentNode.firstChild,
                              currentNode.children.pop(),currentNode)
7:         end if
8:     else
9:         if siblings is empty then
10:            parentNode.children.push(nodeToAdd)
11:        else
12:            return ADDNODE(nodeToAdd,siblings.nextSibling,
                              siblings.pop(),parentNode)
13:        end if
14:    end if
15: end procedure
```

texts referring to the same entity, at least one of which must be included in one of the texts chosen in the first steps (i-iii). Some data points may result in a simplified activity as some steps may become trivial, *e.g.*, a data point with only one sentence in Sentiment Analysis makes the sentence-level creation step (i) very simple. Considering the content provided to the user has to be correct, such a step can be potentially skipped. Furthermore, one label associated with the task will be chosen for each user-defined mapping. These can be automatically inferred through task-specific strategies. For example, in Sentiment Analysis, the labels in External Mappings are directly associated with the sentence's sentiment since such a use-case is associated with the simplification defined for the task. Finally, *rationale mappings* are obtained by combining the texts provided by human actors and the automatically inferred labels. A potential step to collect free-text human rationale can be included to extend the content of the collected mappings.

The mappings provided by the same crowd-worker for each data point can be organized into *individual rationale trees* by applying Algorithm 1, which leverages the definition of a parent-child relationship in a *rationale tree* and relies on the assumption that the node to be inserted is a child of an existing node. In this regard, it is always possible to guarantee that a node is added before its children by sorting them according to the word indices associated with each text in a mapping, *i.e.*, by assessing that one of the child node's texts is contained in its corresponding parent's node text. Furthermore, the root node always exists, as previously described.

Complete Rationale Tree Creation. *Complete rationale trees* are obtained by merging all the *individual rationale trees* produced for each data point. In this process, *rationale mappings* are extracted by applying tree search algorithms to each *individual rationale tree*. *Complete rationale trees* are created using the same algorithm applied to obtain *individual rationale trees* while considering all *rationale mappings* together. As these may appear multiple times, *complete rationale trees* include a frequency score for each node. Such a score describes the ratio between the number of Individual Rationale Trees containing the node and the total number of *individual rationale trees* used to build the *complete rationale tree*. Such a score allows tuning the level of detail of the tree to cut branches or intermediate nodes based on a chosen threshold.

5.5 Preliminary Method Validation

The proposed data collection approach underwent preliminary experiments to assess and improve its effectiveness before its implementation.

Before implementing the application, a preliminary study was conducted to validate the proposed methodology. In particular, it focused on assessing whether the initial design of the approach was understandable and whether it would produce the expected outcome. This section describes the preliminary experiment setup and briefly discusses the results.

Experiment Setup. For each of the three tasks of interest, ten instances were randomly sampled from well-known datasets (*i.e.*, Large Movie Review dataset [264] for Sentiment Analysis, CNN/Daily Mail dataset [298] for Text Summarisation, and SQuAD 2.0 dataset [345] for Question Answering). Such instances were translated into Italian to ease the task for the participants. A tutorial[1], three guided exercises and their solutions, and two sets of four instances each to annotate were prepared for each task. The experiment involved 30 participants (13 females and 17 males) proficient in Italian with different backgrounds and ages (average age of 25.5 years and a standard deviation of 7.60). Participants were uniformly distributed to each task and set of tests, *i.e.*, five participants for each set for each task. The experiment was conducted on paper using coloured markers or a shared Google document at the participant's discretion. Participants read the tutorial and solved the guided exercises, checking the solutions if needed. Then, they performed the provided activities. Although they could not ask the researcher supervising their activity any questions to prevent potential bias, they could consult the tutorial whenever required.

Discussion. This experiment aimed to assess the process's understandability and the results' quality while gathering insights on the constraints and logic to embed in the final application. Participants had much more freedom regarding annotations in this setting than in a digital one. Consequently, it is essential to discern between errors that could be prevented by the application's logic and those that could not when assessing the results and the collected feedback.

All participants correctly completed the activity, and the collected results were manually inspected to assess their quality. Participants deemed the tutorial and the guided exercises effective in explaining the proposed methodology. The results of this experiment support such a statement as all participants provided complete and coherent results. Indeed, the mappings provided by the participants showed high agreement, especially at sentence (i) and word level (iii). At sub-sentence level (ii), mappings were centred around the same pieces of information, although slight differences in the words included in the mapping were identified. Such detail was particularly emphasized in Text Summarization, as one may expect given the vast amount of text compared to the other tasks.

On the other hand, a series of undesired behaviours were observed. For instance,

[1]The tutorial for the QA task can be found in the Additional Material section 7.1. Tutorials can be found at the following link https://tinyurl.com/2s3s9w4v, last accessed September 30, 2024

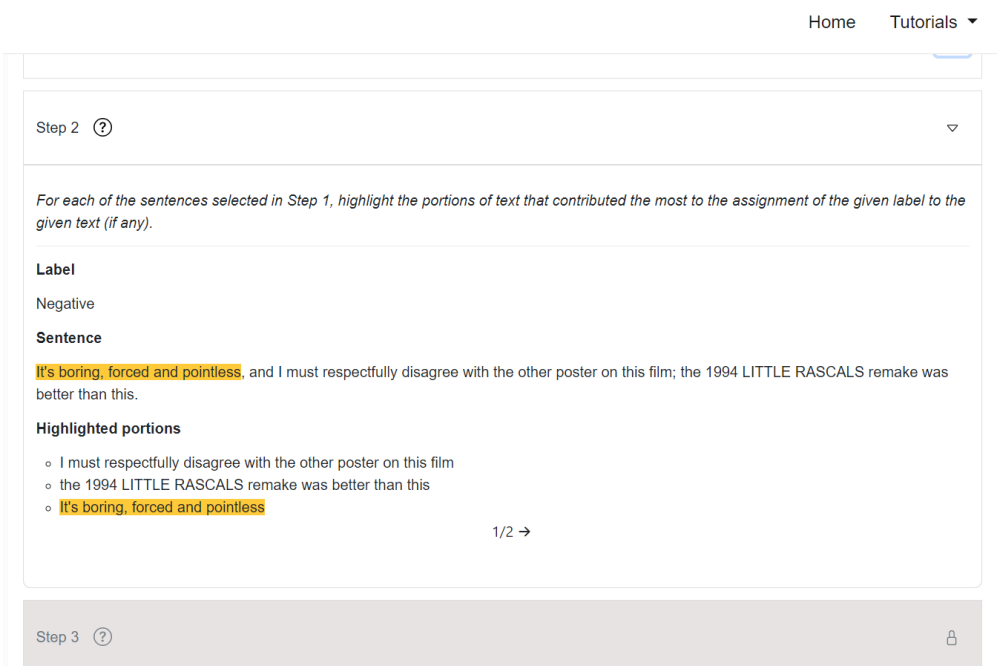


Figure 5.10: *A screenshot illustrating the sub-sentence step for Sentiment Analysis.*

some participants generated duplicated mappings, highlighted complete sentences when performing the sub-sentence-level (ii) or word-level (iii) step, or created mappings involving non-continuous portions of text. The first can be solved by ignoring multiple instances of the exact mapping. Preventing full-sentence selection would fix the second. The latter can be resolved by only allowing users to highlight continuous text portions. While these issues were observed across all tasks, other task-specific behaviours were identified. In particular, improper behaviour identified in Question Answering revealed that a few participants did not understand the difference between the required iterations. All these behaviours were considered when designing the final version of the web-based application so that the UI could precisely inform the user of the steps involved in a task while preventing undesired behaviours.

5.6 Implementation

A web-based application was designed and implemented to collect Rationale Mappings.

Data Structure. The final dataset to be collected describes *rationale trees* for a chosen set of data points from well-known datasets (as previously listed). *Rationale mappings* are stored as tuples, including the text(s) extracted from the original text, the indexes representing the position of the mapping's first and last word in the original text, the label, the mapping type, and a potential free-text rationale. Some mappings (*e.g.*, *abstractive mapping*) require additional data (*e.g.*, the question and its specialization). In *rationale trees*, each node additionally stores a reference to their child nodes, if any. The root node keeps the task's input(s) and output. In *complete rationale trees*,

each Rationale Mapping is associated with its frequency score.

Requirements and Design. The main requirement of the application is to enable users to provide the rationale they apply to a set of NLP tasks of interest organized as *rationale mappings*. This involves displaying the users a data point and allowing them to perform the steps prescribed for a given task. A login system is implemented to keep track of their activity and avoid showing users the same data instance twice. Instances are appointed to users to evenly spread the number of annotations on all the data points. The homepage allows users to select any of the three tasks. When a task is chosen for the first time, a tutorial and three guided examples are provided to teach the user about *rationale mappings* and the process implemented in the application. After completing these exercises, they are prompted to go through the actual data collection process. They are displayed the instance to annotate and the panels that allow them to perform the required annotation steps (i-iv). Each panel displays the text(s) and the components the annotator works on. In particular, the sentence-level step (i) is implemented to allow the user to pick the sentences from a list, while the sub-sentence (ii) and the word-level (iii) steps require users to select and highlight portions of the shown text (as shown in Figure 5.10). Whenever a mapping is created, it is added to a list visible to the user, allowing them to delete any undesired mapping. Finally, coreferences (iv) are identified by highlighting portions of the texts. Annotators can choose colours to highlight terms referring to the same entity. For each coreference group, *i.e.*, terms referring to the same entity, they select the entity such terms refer to. Users can consult the tutorial at any moment.

The interface is structured to guide users by unlocking a step only after completing the previous one. In Question Answering, the three iterations are performed separately, only showing the texts involved in each specific iteration and thus emphasizing their separation. In Question Answering, users pick the question type from a predefined list and highlight the portion of text used to define it. Finally, users are shown an explanatory warning when they make a known mistake (*e.g.*, when creating a duplicated mapping). Warnings allow users to improve their understanding of the methodology while preventing incomplete or redundant data submission.

System Architecture. The application is implemented as a client-server web application. The back-end is implemented in Node.js and connects to a MongoDB database with a collection for each task of interest. The front end is implemented using HTML and Javascript. The application's back-end implements the connection with the database and the endpoints serving instances to users and receiving (and validating) the submitted annotations. The database stores the data points the users annotate. The database consists of three collections, one per task, storing the data points the users annotate. Each data point is characterized by its input and output texts, according to its specific task, and by an array field where all Individual Rationale Trees are stored. The implemented application was deployed on a server with a Linux operative system, using Nginx as a web server to conduct the main experiment.

5.7 Experiment

The developed application was shared to perform the data collection and build Rationale Trees. Participants compiled questionnaires to assess its effectiveness.

Experiment Setup. The application was shared with 151 participants (130 males and 21 females) with an average age of 23.8 and a standard deviation of 1.36. The experiment considered 20 data points sampled randomly from the same datasets used in the Preliminary Validation. The input text(s) length in Sentiment Analysis and Question Answering is lower than 1000 characters, while in Text Summarization, it is lower than 2500. The experiment allowed the collection of 1'495 *individual rationale trees* (569 for Sentiment Analysis, 473 for Text Summarization, and 453 for Question Answering). These are organized into *complete rationale trees* using a Python script, finally leading to 20 *complete rationale trees* per task, one for each data point. The application additionally records the time a user takes to complete a mapping creation task. The average time to annotate an instance is 5 minutes for Sentiment Analysis, 14 minutes for Text Summarization, and almost 7 minutes for Question Answering. The length of the input texts probably causes the longer time taken for Text Summarization.

Participants were additionally asked to fill out a form to review their experience after using the application. The form gathered basic information about the user and general feedback about the application. Moreover, users were asked questions based on the System Usability Scale (SUS) [49] to evaluate usability, as well as questions inspired by the NASA-TLX method [171]. These have been slightly adjusted to improve clarity while maintaining their meaning.

Discussion. Inspecting the questionnaires and the participant feedback contributed towards determining potential areas of improvement while underlining the application's practical design. The SUS score computed from the submitted questionnaires is 65.7, demonstrating that the system's usability is sufficient and falls in the 40th percentile ranking [51]. The partial contribution of each question to the overall score was computed and analyzed alongside the collected feedback to investigate this result further. When interpreting each question's contribution to the SUS score, the closer their score to 10, the better. The question with the lowest score (*i.e.*, a value of 4.64) questions whether the users would use the system frequently. This outcome was expected since most of the efforts of this work have been directed towards making the process understandable and smooth for users rather than making it entertaining. Moreover, the questions addressing the system's complexity and cumbersomeness resulted in a pretty low effect on the total score (*i.e.*, a value of 6.21 and 5.76, respectively). Furthermore, many users consider Question Answering the most complex and repetitive task as it involves multiple texts and a more complex rationale mapping creation process. One may remove the question-answer iteration and derive those mappings from those provided by the users in previous iterations to ease the annotation process for this task. This may be possible by considering the texts of the question-paragraph and paragraph-answer mappings that share the exact paragraph text. Another question providing a low contribution towards the final score (*i.e.*, a value of 6.14) evaluates the participants' confidence in using the system. The comments provided by the users unveil that many

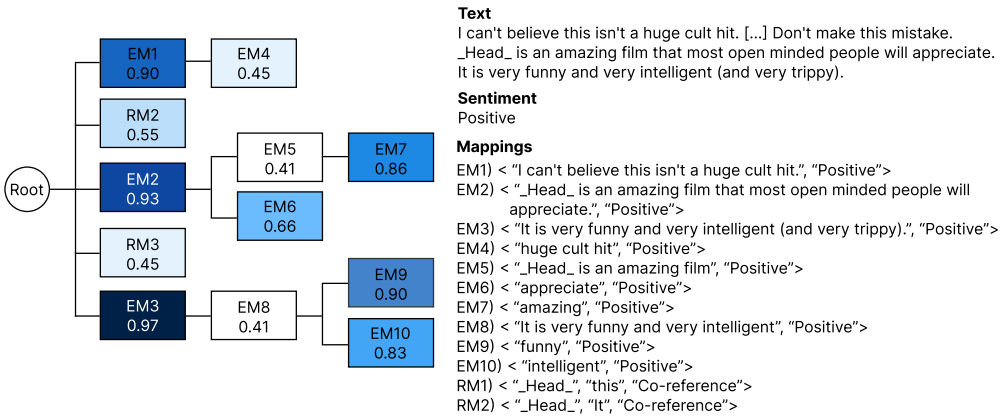


Figure 5.11: *A complete rationale tree for Sentiment Analysis. Nodes are coloured according to their frequency score. The higher the score, the darker the colour. Only rationale mappings with a frequency score greater than 0.4 were reported.*

doubts are related to why three iterations are needed in Question Answering. Other than removing the last iteration, an improvement in this direction could be adding optional sections to the tutorial, further detailing the reasons behind users' annotation steps. On the other hand, questions assessing whether users think that most people would learn to use this system very quickly, whether users deemed they needed the support of a technical person to use the system, and whether they needed to learn many things before using the system positively impacted the final score (*i.e.*, a value of 6.89, 8.61, and 7.77, respectively), confirming the effectiveness of the tutorial and the process.

An approximated NASA-TLX score of 56.9 was computed. Even though it is considered a high result [336], it was quite expected as performing the tasks requires reading and understanding a lot of text (*e.g.*, the tutorial, the data points involved in the activity, etc.). Striving to reduce the workload perceived by users, it would be possible to provide them with hints on the portions of text that are likely to be involved in mappings. Such hints could be displayed to users as light highlights in the text. Despite the potential reduction in the user's workload, adopting it requires evaluating the bias such an approach may introduce since users may follow such advice unthinkingly, resulting in Complete Rationale Trees lacking complexity. Another way to reduce the workload may be allowing users to perform and submit only some annotation steps (i-iv) for each data point, providing their outcome as a starting point for another user's task iteration. Similarly, task-specific changes could be applied. For instance, it would be possible to manually pre-process the data points of the Text Summarization task to reduce their length and complexity by splitting the content into multiple tasks that would still be merged into a single outcome. The latter solution may increase the wordiness of the generated summary, as users would have fewer sentences to combine when applying an abstractive approach. Despite such a consequence, iteratively performing the task to achieve a high-quality result could be possible.

Rationale Trees. The collected dataset consists of *individual* and *complete rationale trees* for the tasks of interest. *Individual rationale trees* are made up of *rationale mappings* provided by single users, whilst a *complete rationale tree* for a particular instance combines all the *rationale mappings* provided by all the users. In *complete ra-*

tionale trees, one may observe that all the *rationale mappings* defined in the sentence-level step (i), *i.e.*, the direct children of the root node, are assigned a higher (or at most equal) frequency than their child nodes. Such behaviour can be explained by inspecting the proposed method. In particular, sub-sentences can only be extracted from the sentences chosen at the sentence-level step (i). Hence, no mapping associated with any of its sub-sentences can exist if a sentence is not selected. On the other hand, one may pick no sub-sentences from a chosen sentence. It is possible to create *complete rationale trees* containing only nodes with a frequency score above a certain threshold, filtering out some nodes or completely pruning some branches. In this way, one can adjust the level of detail at which the tree captures the rationale while still considering the fundamental nodes to describe the reasoning leading from the input text(s) to the output. For instance, considering the tree depicted in Figure 2, removing the nodes with a frequency score lower than 0.45 would remove EM5 and EM8, EM7 would become a child of EM2, and EM9 and EM10 would become children of EM3. The dataset is publicly available on GitHub[2].

5.8 Final Remarks

This chapter describes a novel approach to structuring and collecting human knowledge for Natural Language Processing tasks. We reported on the literature about human knowledge and data structuring in NLP and XAI, as well as argumentation mining, which extensively inspired this work. We explained the concept of *rationale mapping*, its specializations, and how these can be structured into *rationale trees* to describe the reasoning process a human interpreter applies in language-based tasks. Task-specific mappings, potential simplifications, and extensions were detailed for each one. We propose and design an approach to collect such structures, validate them through a preliminary experiment, and implement them into a web application. Data is collected by engaging human interpreters in performing the implemented data flow, finally leading to a dataset tailored to improve model explainability while being intrinsically human-understandable. Experiments revealed the approach's effectiveness in collecting Rationale Mappings and Trees while proposing exciting improvements. We argue these representations contribute towards representing human knowledge to be applied to XAI tasks while also being a suitable way of shaping explanations provided by XAI methods or self-explaining models (*e.g.*, LLMs). Furthermore, a first round of observations on the outcome of the data collection process revealed that even though these structures are effective with all the considered NLP tasks, the task that may benefit the most from these representations is Question Answering. In particular, one may consider the level of detail achieved when applying *rationale mapping* to simple NLP tasks (*e.g.*, Sentiment Analysis) unnecessary. On the other hand, complex tasks like Question Answering require a high level of detail to define a complete human rationale. Hence, future works will investigate such structures' applicability and potential extension when applied to complex QA scenarios (*e.g.*, when multiple answers are to be answered or when complex reasoning is to be used).

Future work will also involve assessing the understandability of *rationale trees* for datasets of interest, improving and detailing the labels applied to some of the proposed

[2]https://github.com/valentinanaldi99/RationaleMappingsDataset.git

mappings, and exploring the applicability of *rationale trees* to other NLP tasks. Furthermore, the data collection approach can be improved to ease the process and extend the method to other NLP tasks, like Natural Language Inference (NLI) and Claim Verification.

CHAPTER 6

Explainable AI in Image Classification

This chapter discusses the research performed on explainable AI in the context of computer vision (CV). In particular, an approach generating local and class-wise explanations to unveil the networks' reasoning process was developed, finally improving the model's interpretability. Crowdsourcing and Gamification were fundamental in collecting human knowledge to generate the final explanations. This chapter is mainly built upon the article

1. Matteo Bianchi, Antonio De Santis, Andrea Tocchetti, and Marco Brambilla. Interpretable network visualizations: A human-in-the-loop approach for post-hoc explainability of cnn-based image classification. In Kate Larson, editor, Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pages 3715–3723. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track. DOI: 10.24963/ijcai.2024/411, URL: https://doi.org/10.24963/ijcai.2024/411

The work was carried out with PhD Matteo Bianchi and PhD Antonio de Santis. The PhD candidate proposed the initial research questions and the draft of the approach's design, further detailed and extended by their colleagues, and contributed to designing the data collection process and the experiments while supervising the development of the proposed methodology.

6.1 Interpretable Network Visualizations

A post-hoc explainability technique named Interpretable Network Visualization (INV) to generate comprehensive, class-wise explanations for image classification models using human knowledge is described.

Acknowledged the increasing complexity of computer vision models, XAI researchers developed techniques that produced explanations for their decision-making process. Regardless of their effectiveness, several critical limitations are yet to be addressed. Considering image classification models, most techniques focus on producing explanations as heatmaps by highlighting the pixels that contributed the most to the output. Such representations grant users insights about what the model is looking at for a specific data instance. In fact, recent XAI techniques mainly provide local explanations, *i.e.*, concerning a single data element, hence not providing a completely interpretable explanation of the model's decision-making process. Indeed, generalizing local explanations for explaining a model's internals is very difficult as the highlighted pixels are relevant only based on their context. Explanations encompassing the entire model's decision-making process are necessary to overcome such limitations.

A post-hoc explainability technique called Interpretable Network Visualizations (INV) using human knowledge to create comprehensive, class-wise explanations for image classifications is presented. The approach works on Convolutional Neural Networks (CNNs) without requiring modifications or performance trade-offs. The provided outcome details the features and patterns identified by the network at each layer, alongside their importance towards the output. These representations are generated through feature map clustering and the score computed through Grad-CAM [382]. Then, crowdsourcing and Gamification are applied to label the computed local visual explanations using human knowledge to improve model interpretability [430]. Finally, the proposed method combines visual and textual explanations and aggregates them to generate class-wise representations with different levels of detail. These are based on the number of layers considered for each group in the final aggregation. Experiments assessing INVs capabilities are carried out, comparing the proposed method against state-of-the-art approaches and revealing effectiveness at least comparable if not better under most XAI-related aspects (*e.g.*, interpretability, understandability, etc.).

6.1.1 Context-specific Related Works & Background

The background describing XAI approaches in Computer Vision is described, highlighting the most essential approaches towards our implementation.

Explainability for Computer Vision Models. Several approaches were designed to explain black-box computer vision models. Zeiler et al. [510] examined the internal workings of CNNs by reconstructing the activations of the intermediate layers and projecting them back to the input pixel space, finally providing a visual representation of the information extracted by the network from the input image up to a chosen layer. In a later study, Simonyan et al. [397] proposed a gradient-based visualization method to generate saliency maps to generalize the deconvolution approach. These

visualizations use the backpropagation method to compute the transfer function's gradients concerning the input image. The higher the gradient's value, the more significant the impact of a pixel on the final prediction. Similarly, SmoothGrad [403] and Guided Backpropagation [412] are widely adopted gradient-based techniques that enhance the quality of the generated visualizations even further. Despite their effectiveness and ease of implementation, gradient computation might sometimes result in an inaccurate assessment of feature importance mainly due to gradient saturation, which occurs when their score flattens in an input's vicinity (*i.e.*, when the gradients get close to zero). DeepLIFT [393] tackles this problem by computing each feature's contribution compared to a baseline input (*i.e.*, a completely black figure). Such an approach was combined with the gradient-based one, generating the so-called Integrated Gradients (IG) [428] methodology, overcoming gradient-based methods' limitations by integrating a prediction's gradients over a path from a baseline input to the actual input. Furthermore, such a novel approach addresses the implementation invariance issue caused by DeepLIFT. Despite overcoming such fundamental problems, IG enforces a trade-off between faithfulness and computational intensity, making it rarely worthwhile [269].

In the same context, activation maps were presented as an alternative to extract feature importance [529]. The proposed approach focuses on Global Average Pooling (GAP) layers rather than fully connected ones, reducing feature maps into a scalar value to represent each feature. Class Activation Maps (CAMs) are extracted through a weighted linear sum of the computed feature maps and then upscaled to the input image's size to generate the final heatmap. The latter represents the class-weighted average of the input image's sections observed by the CNN before the final prediction, providing class-specific visualizations by only highlighting regions relevant to each class. This approach applies to CNNs employing a GAP layer before the final prediction. Whenever a fully connected layer is used, architectural changes are required, such as substituting and re-training the last layer, potentially hindering model accuracy [257] and stressing the concept of post-hoc explainability.

Gradient-weighted Class Activation Mapping (Grad-CAM) [382] was later introduced as a generalization of the CAM approach. It applies GAP on the gradients for a given class concerning the feature maps. It generates the final class-specific map using a ReLU function while considering features that positively impact the prediction. Grad-CAM is often combined with other methods (*e.g.*, backpropagation [412]) to overcome its lack of fine-grained pixel-scale representations whenever needed. Grad-CAM provides computational efficiency and broad model applicability. Furthermore, it can be applied to any convolutional layer with lower efficacy, focusing on less semantically relevant local features. Grad-CAM++ [72] incorporates high-order derivatives when computing CAMs, consequently improving its outcome. Other variances were also proposed, *e.g.*, Eigen-CAM [295], Smooth Grad-CAM++ [319], and Score-CAM [467], resulting in a context-specific performance increase.

The discussed XAI methods generate heatmaps highlighting an image's areas with the highest influence on the prediction while lacking the motivation for which these are relevant. Despite the potential faithfulness of the explanations, their interpretation is deeply subjective. A combination of visual and textual explanations was identified as a possible solution to such a deficiency. Grad-CAM was combined with manually labelled images to automatically assign names to neurons at training time, finally gener-

ating textual descriptions for the heatmaps to improve their comprehensibility [37]. Another technique named Testing with Concept Activation Vectors (TCAV) [213] was also proposed. It determines a specific feature or human concept's importance towards predicting a particular class by inputting sample images representing a single feature and observing the network's prediction. Similarly, the method allows the detection of biases in networks. TCAV has also been improved, providing enhanced variations [372].

In conclusion, Grad-CAM (and its variances) and TCAV effectively explained AI decisions. Nonetheless, it has been shown they might introduce bias in explanations. In Grad-CAM, humans might misinterpret why a region is essential to the prediction due to biased assumptions. In TCAV, bias might be introduced due to the images chosen to represent a feature [446]. Hence, it is necessary to address and mitigate such potential biases when developing novel XAI techniques.

6.1.2 Interpretable Network Visualizations

An overview of INVs is provided, highlighting their features and providing a high-level description of the process to generate them.

While state-of-the-art computer vision XAI methods provide faithful explanations by highlighting the most critical region of the image towards the network's prediction, they do not give a complete understanding of the model's decision-making process. Image classification involves multiple feature extraction stages across various layers. Despite its complexity, providing a user with an overview of such a process potentially increases model transparency. This work proposes a post-hoc explainability framework to generate local and class-wise explanations as Interpretable Network Visualizations (INV), providing human-understandable and accurate visualizations of the features extracted by the CNN at each layer. These representations contribute to explaining a correct model's decision-making process or debugging an incorrect one. INVs are made of layers consisting of aggregated heatmaps, each highlighting areas of input images where important features were identified. Figure 6.1 provides a partial example of such visualizations (further detailed in later sections).

Each layer's aggregated heatmaps are generated by clustering feature maps based on similarity (*i.e.*, the area of the figure they deem essential for the prediction), resulting in a variable number of groups based on the generated explanations. Each aggregated heatmap is also assigned the relative importance of the feature maps towards the final prediction and a set of crowdsourced labels describing the represented concept. Multiple labels are assigned, accounting for heatmaps highlighting multiple entities and a score is assigned to them based on their relevance (*e.g.*, their frequency). Such crowdsourced labels are inherently human-understandable, hence enhancing INVs' interpretability. INVs might consider all CNN layers or just a subset of interest. In particular, deeper layers might be better suited for such representations as they learn more advanced semantic concepts and have wider receptive fields than shallow layers. Furthermore, potential biases might be identified by deconstructing the feature extraction process to analyze each part individually. Similarly, potential cross-layer feature correlations might be investigated, allowing for a more comprehensive understanding of their relationship. Crowdsourced labels allow for enhanced interpretability

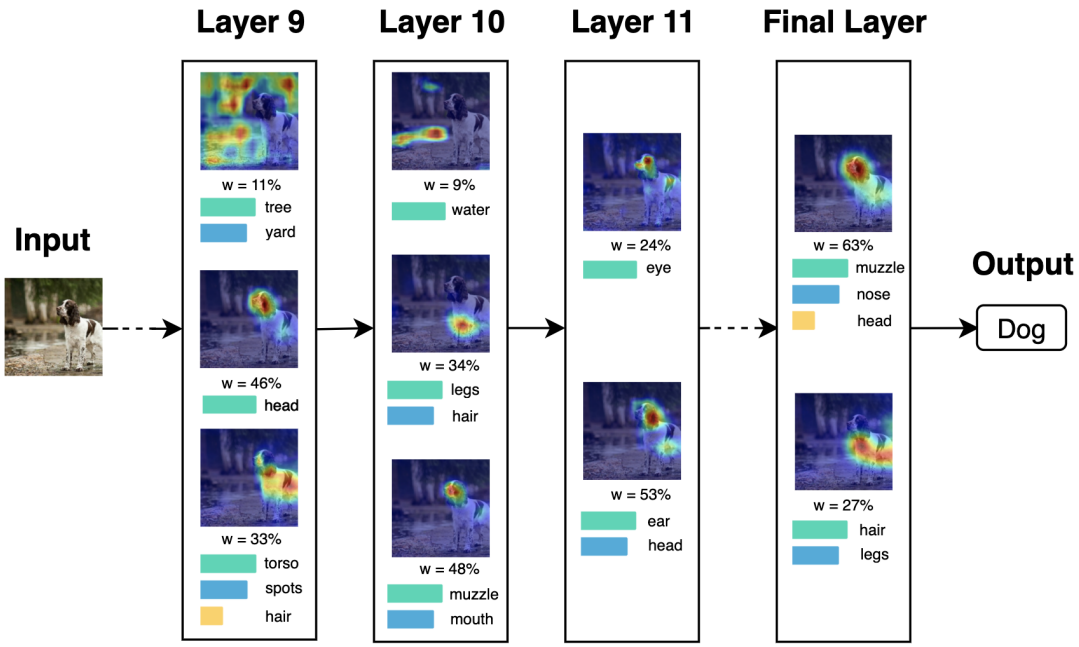


Figure 6.1: *An INV showing the layer-wise feature extraction process. Each layer includes a variable number of heatmaps representing the features identified by the network, each associated with labels describing human concepts defining these features and a weight representing the contribution of each feature towards the output. Features with meagre weight are omitted from the visualization.*

and contribute to aggregating layer-wise explanations across multiple images, providing a class-wise perspective of the network decision-making process.

6.1.3 Generating INVs

The process to generate INVs is described in detail. Three main steps are involved, *i.e.*, feature map analysis, human knowledge collection, and label analysis.

Given a CNN trained for image classification and an input image, INVs are built following these steps

- **Feature Maps Analysis**. Feature maps are extracted and clustered, and their weights are computed. Clusters are merged to generate representative heatmaps, referred to as cluster maps.
- **Human Knowledge Collection**. Labels are collected through crowdsourcing to enhance cluster maps' interpretability.
- **Label Analysis**. The collected labels are processed, filtered, and structured. Moreover, cluster maps with similar labels are also merged.

Feature Map Analysis. In the first step, the CNN is provided with an input image and feature maps and their importance is extracted. These are then clustered and

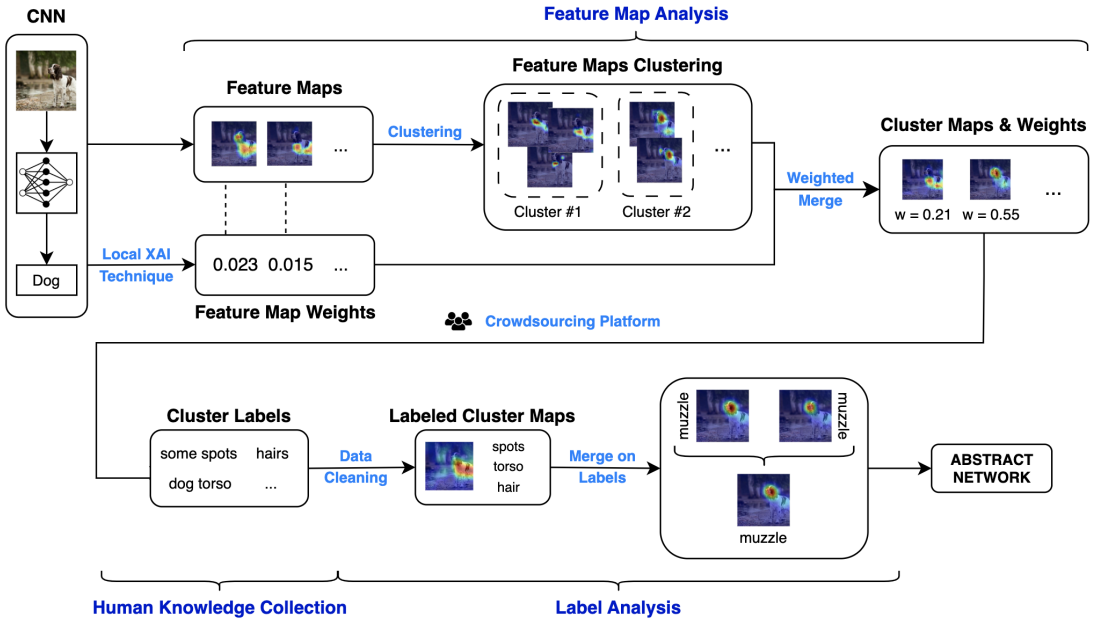


Figure 6.2: *The process for generating INVs. In the first step, feature maps and their weights are extracted from the network. These are then clustered to generate representative heatmaps (i.e., cluster maps). In the second step, human knowledge is collected to label clusters. In the final step, such labels are cleaned, and cluster maps with the same ones are merged together.*

merged into a single map to generate representative cluster maps for each layer.

Feature Maps and Weights Computation. Feature maps are extracted, and a ReLU activation function is applied to each convolutional layer to consider only positive activations, removing empty feature maps. Then, feature maps are associated with the corresponding class-specific weights towards the predicted class computed using Grad-CAM with Guided ReLU, considering positive gradients in positive activations' regions. These are further normalized to enhance weights' interpretability, guaranteeing a total weight per layer of 1.00 and allowing representations involving percentages. Feature maps with weights less than zero are filtered out. Furthermore, a weight threshold might be applied to exclude low-importance feature maps to improve the clustering process. This threshold is based on the number of each layer's feature maps and aims to reduce the number of feature maps while retaining the vast majority (*e.g.*, 70-90%) of the total weight in each layer. The number of feature maps per layer is often in the order of hundreds or more, making it overwhelming for humans to handle. Moreover, these can be clustered and merged into cluster maps as multiple filters in the same layer produce similar feature maps.

Preprocessing. Given the intrinsic complexity of clustering, a few pre-processing steps are performed.

- Feature maps are normalized through min-max normalization, making them comparable in region activation.
- A dimensionality reduction approach combining Principal Component Analysis

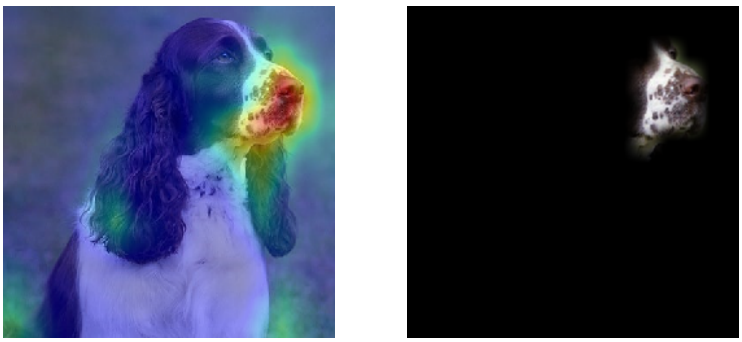
(PCA) [276] and t-distributed Stochastic Neighbor Embedding (t-SNE) [455] is applied since feature maps are highly susceptible to the curse of dimensionality. The first is applied for data visualization in low-dimensional spaces while preserving its local structure, making it efficient in clustering complex datasets. The second reduces the number of dimensions before applying t-SNE as it requires a high computational effort. PCA's parameters must be chosen based on the model and the layer of interest as feature maps' number and size might vary.

Clustering. Agglomerative clustering is then applied for each network layer as hierarchical clustering approaches were identified as the best-suited compared to density-based and centroid-based ones, which remove potentially interesting noisy points and perform well only when specific data distributions are applied, respectively. Furthermore, the chosen approach produces many clustering results for an incremental number of clusters, making it very flexible. A silhouette score metric (*i.e.*, a measure of cluster cohesion and separation) is applied to select the optimal number of clusters as it varies based on the input image and layer and depends on the number of features a convolution layer extracts. In this case, the average silhouette score is computed since visual inspection is unfeasible. Furthermore, many clusters are complex to manage from the human understandability and crowdsourcing perspectives. Hence, a reasonable range of clusters (*e.g.*, from three to eight) must be chosen based on model size, availability of resources for crowdsourcing, and human comprehensibility.

Merging Feature Maps. Finally, a cluster map representing the clustered feature maps is calculated using a weighted average approach for each cluster. The final weight is obtained by summing the weights of all feature maps belonging to a cluster, which are computed using Grad-CAM. Such a score indicates the significance of the cluster map to the predicted class. After computing the cluster maps and their scores for each layer, a threshold can be chosen to filter out low-weight maps based on the available crowdsourcing resources.

Human Knowledge Collection. Labels representing human concepts highlighted in the computed cluster maps are collected via crowdsourcing at this step. Such a process enables associating visual explanations with human-interpretable concepts and significantly reduces interpretation bias.

Cluster Map Masking. An interpretable visualization of a feature map is usually obtained by overlaying it on its input image after normalization, up-scaling, blurring, and colour mapping. Cluster maps underwent the same process, obtaining an equivalent representation, as in Figure 6.3a.



(a) *Cluster Map overlay.* **(b)** *Masked overlay.*

Figure 6.3: *On the left (a), an overlay of a cluster map identifying a dog's muzzle. On the right (b), its corresponding masked image showing only the highlighted area is represented.*

While this representation provides evidence on which parts of the image are recognized by the model, it might still drive humans to focus on other features, potentially generating labels that are not representative of the highlighted ones. Therefore, a masked image is generated by overlaying a binary mask hiding the non-highlighted portions to the input image, as represented in Figure 6.3b. Filtering the most critical pixels of a cluster map is achieved by choosing an estimate of the percentage of such pixels to show, computing its percentile, and finally considering only the values greater than such. Such a parameter can be adjusted to create different masks showing gradually larger portions (as depicted in Figure 6.4), producing a finite number of masks with a non-linearly increasing percentile since the information density decreases while moving away from the most highlighted pixels.

(a) *2.5%.* **(b)** *4.5%.* **(c)** *8%.* **(d)** *13%.* **(e)** *20%.* **(f)** *30%.*

Figure 6.4: *Six masks generated by gradually increasing the pixel percentage parameter.*

Gamification. A gamified activity named *Deep Reveal* was developed, deployed, and shared to collect the labels. Gamification increases engagement and enhances participants' effort, resulting in a higher amount and higher quality labels. Furthermore, Gamification can help design the activity's flow to drive participants' behaviour. In this use case, participants are driven to behave like the neural network, observing and analyzing features to guess the correct class. *Deep Reveal* is a web-based game in which participants are presented with masked images of cluster maps and must guess their class and explain the rationale behind their decision. Figure 6.5 shows the process implemented in the activity.

When guessing, users pick the class among the ones in a randomly generated subset from the model labels, including the correct one. In such a setting, a reasonable number of answers has to be chosen (*e.g.*, five options) as displaying all model classes might increase the task's complexity when an overwhelming amount of classes is available.

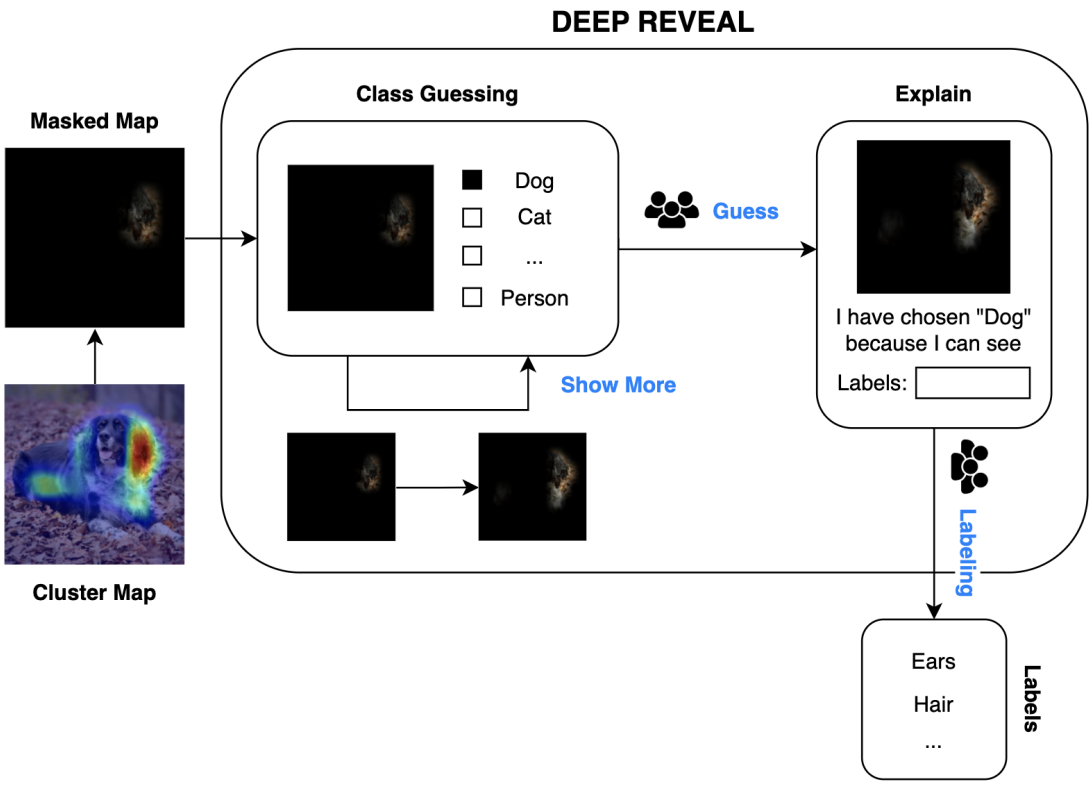


Figure 6.5: *A pipeline describing the process for crowdsourcing labels through Deep Reveal. A masked version of the cluster map is first shown. Users can try to guess the image right away or ask to see more of it. After submitting their guess, users are asked to provide labels to describe the represented elements that drove their decision.*

Additionally, randomness has to be used when generating each class subset, as deterministic approaches might introduce bias or determine recognizable paths for users. Whenever a masked map is too hard for a user to guess, they can gradually increase the displayed area up to five times to get more clues. If so, a masked map with an increased pixel percentage parameter is displayed (see Figure 6.4). After the displayed area is extended for the fifth time, the user is given the option to retire. Inspecting the number of lost games and resignations is important to determine whether the cluster map is discriminative in predicting the class. The guessing game design allows focusing user attention on the discriminative features, *i.e.*, the ones they deem important towards the prediction. Whenever users pick a class, they are asked to specify which recognized features led to their guess, prompting them to provide labels for the cluster maps. Such data is collected as free-text, avoiding introducing biases, *e.g.*, providing a list of options might bias the choice and hinder the data collection. A scoring system and a leaderboard were also introduced to enhance user engagement and drive competition. The first awards users for guessing the right class and providing labels while increasingly reducing the score whenever an extended masked map is requested. The second provides an overview of the total score associated with each player. The gamified activity also involves attention checks, *i.e.*, some masked maps are purposefully simple for the user to solve, allowing the detection of untrustworthy users whose data has to be excluded from the analysis.

Label Analysis. Completed the data collection step, the collected free-text labels must be analyzed and cleaned as these might contain errors (*e.g.*, misspelt words, etc.) or might require further processing (*e.g.*, long phrases, stop-words, synonyms, etc.). Labels corresponding to the class name and stopwords are filtered out as they do not provide additional information. Labels are then converted into word embeddings using *all-mpnet-base-v2*, a Sentence-BERT model [352]. Agglomerative Clustering using cosine similarity, complete linkage, and the average silhouette score to select the optimal number of clusters was applied to group similar labels. Cosine similarity is then applied to define a representative single-word label for each cluster. Finally, groups represented by labels with the same lemma are unified.

Each label is then assigned a score to attribute higher importance to each cluster's most frequently collected ones. Such a score considers label frequency and the percentage of images revealed (*i.e.*, based on the number of hints used) when labels were collected. The score *score(C, l)* for label *l* in cluster *C* is computed through an implementation-dependent heuristic that assigns a higher importance to the frequency as shown in Equation 6.1

$$score(c,l) = freq(c,l) \times \left(1 - \frac{avg\_ext\_map(c,l)}{2 \times (max\_ext\_map - 1)}\right) \tag{6.1}$$

where *freq(c, l)* is the frequency of label *l* in cluster *c*, *avg_ext_map* is the average number of times a user required an extended masked map, and *max_ext_map - 1* is the number of available hints (and extended masked maps) for the game. The second term ranges from 1 when no extended masked maps are requested to 0.5 when the largest available map is displayed. Furthermore, the term decreases linearly. Moreover, the score of labels from users who guessed incorrectly counts as one-fourth of the standard score as they have a higher probability of being imprecise. Such a score is assigned to each label to identify the ones best describing the cluster maps they were assigned to.

Moreover, various cluster maps might represent similar features with similar labels due to potential flaws in the clustering process or whenever a feature is represented multiple times in the image. Therefore, clusters with the most relevant labels are merged before the final INV is created. In particular, clusters sharing at least one label with the maximum score are merged. These might lead to merging multiple clusters (see Algorithm 2).

Algorithm 2 Cluster merge on maximum scoring labels

```
for all l ∈ layers do
    repeat
        for all c_i, c_j ∈ l.clusters, i ≠ j do
            if c_i.best_labels ∩ c_j.best_labels ≠ ∅ then
                create new cluster having best_labels = c_i.best_labels ∪ c_j.best_labels
                delete c_i and c_j from l.clusters
                add new cluster to l.clusters
                break
            end if
        end for
    until l.clusters changes
end for
```

Cluster maps are merged using a weighted average with a final weight equal to the sum of the merged map's weights. Their labels are also combined by applying a weighted average to their scores. Finally, an image's INV is generated by organizing its cluster maps, their weights, and labels for each layer.

6.1.4 Implementation

For the model, a VGG-16 network is augmented using Imagenette. Feature maps are extracted using Grad-CAM, processed using dimensionality reduction, PCA, and t-SNE, and finally clustered using Agglomerative Clustering. Masked cluster maps are also produced to be employed in the data collection activity.

Model and Dataset. The reference model for implementing INVs in this use case is the VGG-16 model pre-trained on ImageNet, available through the Keras python library. VGG-16 is a CNN architecture consisting of 13 convolutional, 5-max pooling and three dense, fully connected layers. The initial convolutional and the pooling layers were kept, and their weights were initialized. The input shape, dense layers, and output classes were updated to fit the Imagenette[1] dataset. It includes ten classes and 13'394 images. The model was trained using image augmentation, early stopping, dropout, transfer learning, and fine-tuning to prevent potential overfitting. Transfer learning iterations were initially performed to train the fully connected layers, after which the last three convolutional layers were fined-tuned, finally achieving a training and validation accuracy of 96.65% and 97.45%, respectively.

Feature Map Extraction. Extracting feature maps requires building new Keras models for each convolutional layer. These accept the same input and the corresponding convolutional layer output from the original model (*i.e.*, the layer's feature map). Each feature map's weight was computed using Grad-CAM's computation for gradients. Weights and feature maps are normalized through unit and min-max normalization, respectively. Feature maps are filtered considering a threshold varying according to the number of feature maps composing a layer and the model architecture (see Table 6.1)

Number of Feature Maps	64	128	256	512
Weight Threshold	1.4	0.7	0.35	0.175

Table 6.1: *A table associating the number of feature maps in a layer and the weight threshold to filter out feature maps.*

Clustering. After computing the feature maps, a clustering step was performed. For each layer, feature maps were selected by applying dimensionality reduction, transforming the normalized future maps into one-dimensional vectors, applying PCA with a minimum number of components equal to 50 or the library default value (*i.e.*, the minimum between the number of feature maps and the number of pixels), and applying t-SNE using default parameters. Following this pre-processing step, Agglomerative clustering with Euclidean distance and Ward linkage (*i.e.*, an approach linking clusters

[1] Imagenette - https://github.com/fastai/imagenette (Last Accessed, 11 November 2024)

that minimize the increase in the sum of squares distances between clusters) was applied. Clustering was tested with a number of clusters ranging between three and eight, and the value that maximized the average silhouette score was finally chosen. Each cluster's weight and final map were computed, additionally thresholding the results to exclude clusters with low weights to focus on the most relevant cluster during the labelling phase. This threshold was calculated for each layer using a heuristic, *i.e.*, as the maximum value between one-third of the weight of the most significant cluster and half of the average weight of all clusters in that layer.

Algorithm 3 Apply a mask to an image given a feature map and a percentile

function MASKIMAGE($image, heatmap, p$)
  $heatmap\_up \leftarrow$ UPSCALEHEATMAP($heatmap, image.shape$) ▷ up-scales the heatmap
  $percentile \leftarrow$ COMPUTEPERCENTILE($heatmap\_up, p$) ▷ p-th percentile of the heatmap
  for $pixel$ **in** $heatmap\_up$ **do** ▷ map computation
    $pixel \leftarrow 1$ **if** $pixel \geq percentile$ **else** 0
  end for
  $size\_ratio \leftarrow image.shape / heatmap.shape$
  $blur\_sigma \leftarrow$ COMPUTEBLURSIGMA($size\_ratio$) ▷ the higher ratio, the higher sigma
  $mask \leftarrow$ APPLYGAUSSIANFILTER($heatmap\_up, blur\_sigma, truncate = 1.5$)
  $masked\_image \leftarrow$ APPLYMASK($image, mask$) ▷ overlay the mask on the image
  return $masked\_image$
end function

Masking Cluster Maps. Following the clustering procedure, the resulting cluster maps were masked (see Algorithm 3). First, the cluster map is up-scaled to the original image size. Then, various masked cluster maps are computed based on a set of percentiles of choice (see Table 6.2), considering only the pixels where the value of the heatmap is greater or equal to the chosen percentile. The number of percentiles is determined based on the number of hints available in the Deep Reveal gamified activity. At the same time, their corresponding parameters were chosen to show the most essential part of the image increasingly.

Hints used	0	1	2	3	4	5
p parameter	2%	4.5%	8%	13%	20%	30%
Δp	-	2.5%	3.5%	5%	7%	10%

Table 6.2: *The list of the values for the parameter of the masking algorithm used to produce the masked maps for Deep Reveal. The delta for each step is also shown.*

Gaussian blurring was applied before masking the images for a smoother visualization, using the ratio between the shape of the input image and the shape of the heatmap (see Table 6.3).

Input image shape	512x512				
Heatmap shape	256x256	128x128	64x64	32x32	16x16
size_ratio	2	4	8	16	32
$sigma$ parameter	(3,3)	(6,6)	(12,12)	(16,16)	(24,24)
$truncate$ parameter	1.5				

Table 6.3: *A table representing the values chosen for the Gaussian filter implementation based on the shape of the input image and the cluster heatmap.*

Finally, the computed blurred masks for the considered percentiles are overlayed on the input image to achieve the final masked cluster map.

6.1.5 Deep Reveal

The design and implementation of a gamified activity named *Deep Reveal* to collect human-generated labels for masked heatmaps is described, reporting its main features and elements.

As one of the objectives of this research is to enhance visual explanations with textual labels, a gamified application named *Deep Reveal* was implemented and deployed. Users are engaged in a gamified image-guessing activity through which labels describing the most relevant part of the image are collected. The designed activity involves a variety of elements.

Login. A login system is included to keep track of the played matches, preventing users from labelling the same image more than once. Users provide a username, an e-mail address, and a password. Users can also play as guests, ensuring they are not given the same picture twice, only within each session. Furthermore, identifying users is fundamental to filtering out data from untrustworthy users.

Leaderboard. Users are organized in a leaderboard based on their score, the number of played games, win rate and win streak. The leaderboard shows the top 50 players based on a sorting chosen among the available statistics.

Gameplay. In each match, players are asked to guess the class of the image and describe the most important details that led to their guess. They are shown a masked heatmap revealing a small portion only and six classes to choose from (including the right one). Players start each game with 300 points, and they can increase the visible region of the image up to five times by incurring a penalty of 50 points. Players are awarded extra points whenever they guess multiple images right in a row (so-called win streak bonus). As players pick one of the available choices, they are prompted to provide the features that drove their guess, awarding them with 25 points whenever they provide any. Following this last step, the content of the image, the correct class, and the points earned are displayed. The application balances the number of games per cluster map, potentially resulting in each cluster map having a similar number of labels. The web interface of the application in the labelling phase is shown in Figure 6.6

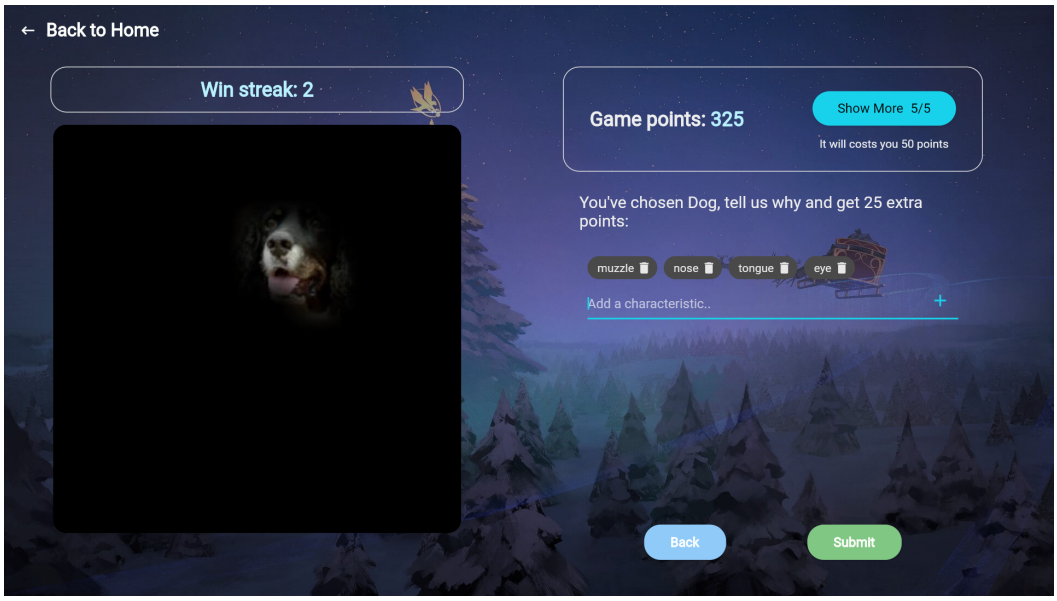


Figure 6.6: *A screenshot showcasing the labelling phase of the game. The user describes the characteristics that led to their guess as labels.*

Attention Checks. Attention checks are added to the game to ensure proper data collection and filter out content from untrustworthy players. In particular, each player's third match always presents the same image (the one represented in Figure 6.6), which is very simple to guess. A player incorrectly classifying this entity will be marked as untrustworthy, and their data will not be considered for the final analyses.

Implementation. *Deep Reveal* was implemented as a web application. Its back end was implemented in Python, using the Django REST framework to generate the application server APIs and a PostgreSQL database to store the data. Python was chosen as it allows for the generation of masked images on the fly, reducing the efforts for data storage. The client-side was implemented using Flutter Web, a portable platform allowing mobile and desktop deployments. Nginx was employed to deploy the web application.

6.1.6 Experiments & Results

Deep Reveal is shared with participants, and their knowledge is collected, analyzed, and cleaned to generate INVs, for which an example is reported. Statistics on the collected data are also reported.

Number of Images and Layers. Assuming the CNN of choice, five images for each of the ten classified classes were considered, leading to 50 images. Furthermore, only the last nine convolutional layers were analyzed, as initial layers focus on less semantically meaningful local features (*e.g.*, edges, outlines, etc.).

Feature Map Analysis. After selecting the images for the experiment, they underwent the cluster map generation process. Hence, feature maps and their weights were

computed, normalized, and thresholded to remove those with weights lower than a threshold of choice. Then, feature maps were processed using PCA and t-SNE, clustered choosing the number of clusters that maximizes the silhouette score, and thresholding was used to keep only the most relevant ones. As presented in Table 6.4, the deeper the layer, the higher the percentage of removed feature maps. On the other hand, the total leftover weight showed a slight upward trend. Such an observation suggests the weights are concentrated in deeper layers, representing more relevant information through fewer clusters. Finally, the computed feature maps for each cluster are merged, resulting in about 1'950 cluster maps to be labelled.

Layer Name	Feature Maps Threshold	Leftover Clusters	Leftover Feature Maps	Leftover Weight
block3_conv1	<0.35%	4.3	52.7%	70.2%
block3_conv2	<0.35%	4.4	51.4%	68.6%
block3_conv3	<0.35%	4.8	46.2%	65.2%
block4_conv1	<0.175%	4.2	46.5%	71.9%
block4_conv2	<0.175%	4.8	44.9%	72.1%
block4_conv3	<0.175%	5.2	35.9%	72.6%
block5_conv1	<0.175%	4.6	30.4%	73.8%
block5_conv2	<0.175%	3.6	25.4%	72.5%
block5_conv3	<0.175%	3.3	14.6%	78.4%

Table 6.4: *A table presenting the results of thresholding and clustering. The first was assigned a lower percentage in deeper layers as twice as many feature maps were computed. The values for leftover clusters, feature maps, and weight are averaged across each layer's images.*

Human Knowledge Collection. Deep Reveal was deployed as a web application and shared with 210 participants to collect labels. Participants were asked to fill out a usability and workload questionnaire (see Additional Material 7.2) after completing the experiment to validate the design and implementation of the tool, identify potential issues, and collect feedback for improvements. The usability questionnaire was based on the System Usability Scale (SUS) [49], and the workload one was adapted from the NASA-TLX [171] method. Participants could insert labels in Italian and English, allowing for broader data collection at the cost of an additional translation step. In the end, about 10'000 labels were collected. Only the data provided by one participant was discarded as they did not pass the attention check. A comprehensive overview of the outcomes of the crowdsourcing activity is provided in Table 6.5.

Games	Wins	Losses	Resigns	Total raw labels	Raw labels per map	Hints per game
7948	7150	688	110	9968	5.096	1.6

Table 6.5: *A table presenting the outcomes of the crowdsourcing activity.*

Label Analysis. The collected labels are translated from Italian to English and manually validated to avoid losing semantics in the image's context for a few cases (*e.g.*, "Esso" was translated to "it", although Esso is the name of a well-known American oil company). Labels were then filtered (as previously discussed), mapped to word embeddings and clustered, finally identifying a representative single-word label for each cluster. Each label was assigned a score and used to merge cluster maps according to

the ones with the highest score. Table 6.6 reports some statistics about the labels and the cluster maps before and after the analyses.

Total labels	Unique labels	Labels per Map (pre-merging)	Cluster Maps (post-merging)	Weight per Map (post-merging)
12082	430	6.18	1192	27.07%

Table 6.6: *A table showing the outcome of the label analysis step, also representing the number of labels obtained and the cluster maps resulting from the merging using labels. The average number of non-unique labels per map before merging and the number of cluster maps after merging are also reported.*

Interpretable Network Visualizations. After analyzing the collected content, INVs are built for each class of interest by organizing cluster maps for each layer, displaying their weight and the top three highest-scoring labels. Figure 6.7 provides an example of such visualization for the *chainsaw* class.

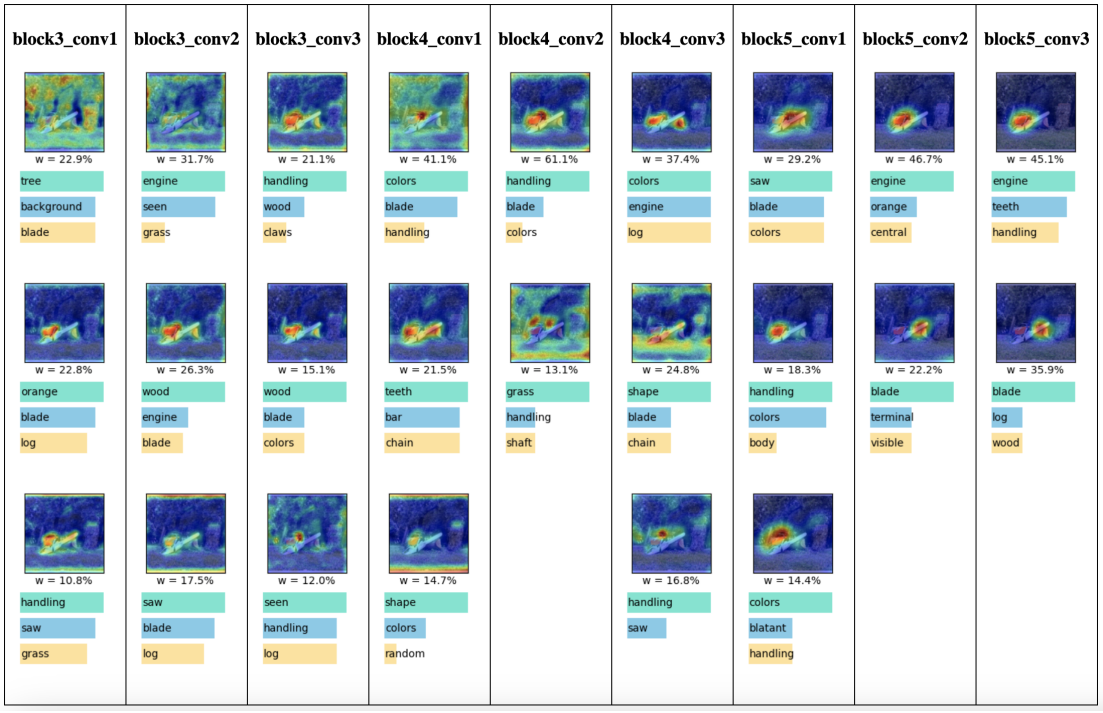


Figure 6.7: *An INV for an image of a chainsaw showcases the labelled features identified by the network in the last convolutional layers. Only the most important maps (i.e., the ones that were not filtered out) are shown in the INV, reporting their weights. In this use case, the engine and the blade (i.e. the saw) were the most important features identified by the network. However, other features (i.e., the handling and the colour) were important too.*

Each feature map was also assigned a detailed visualization, including a plot of the collected labels with their corresponding scores, the number of feature maps involved in the cluster, and additional information about the data collection process (*i.e.*, the masked image and the number of wins, losses, and resigns.)

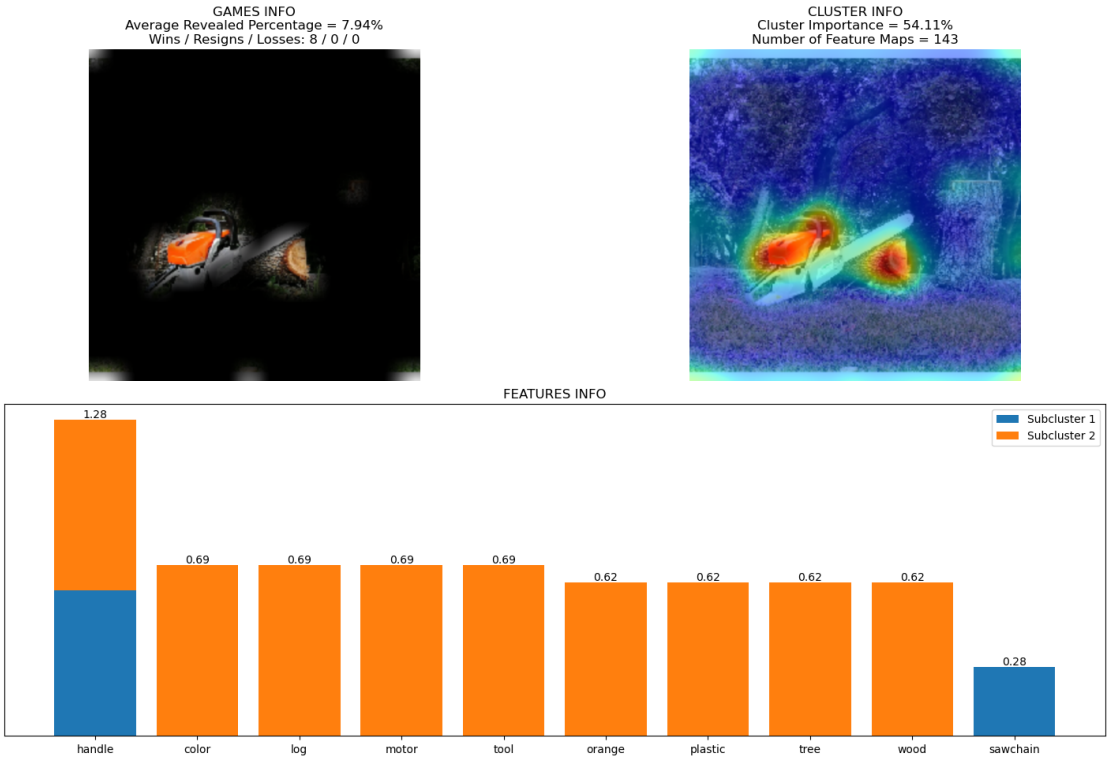


Figure 6.8: *A detailed visualization of a cluster map of an image of a chainsaw. A masked image representing the average image portion revealed by users when guessing the class and labelling the cluster map is shown on the top left. Information about the cluster is also presented, including the cluster map overlay, its importance, and the number of feature maps it involves. The bar plot displays the labels describing the cluster map ordered by score, specifying each one's contribution towards the final score and highlighting the previously performed merge. Information about the total wins, losses, and resigns is also provided.*

6.1.7 INVs Discussion

A subset of INVs of interest is described to highlight their features and potential gaps.

Three use cases are presented and discussed, while all the INVs for the 50 images considered in the experiments are available online[2].

Chainsaw INV (Figure 6.7). In this INV, the prediction is mainly determined by two class-discriminative elements, *i.e.*, the *engine* and the *blade* (*i.e.*, the saw), as the model increasingly focuses on them as the layers get deeper. Non-discriminative features (*e.g.*, *tree*, *wood*, and *grass*) are slightly relevant in shallow layers and irrelevant in the deeper ones. Furthermore, it was observed that the network potentially learned to identify the motor feature based on its colour (*i.e.*, orange), similar to some users in the labelling step. Further investigation revealed that the predicted class was a chainsaw when inputting a completely orange image, showing a coherent behaviour between

[2]INVs - https://github.com/Antonio-Dee/interpretable-network-visualizations (Last accessed 15 November 2024)

models and humans.

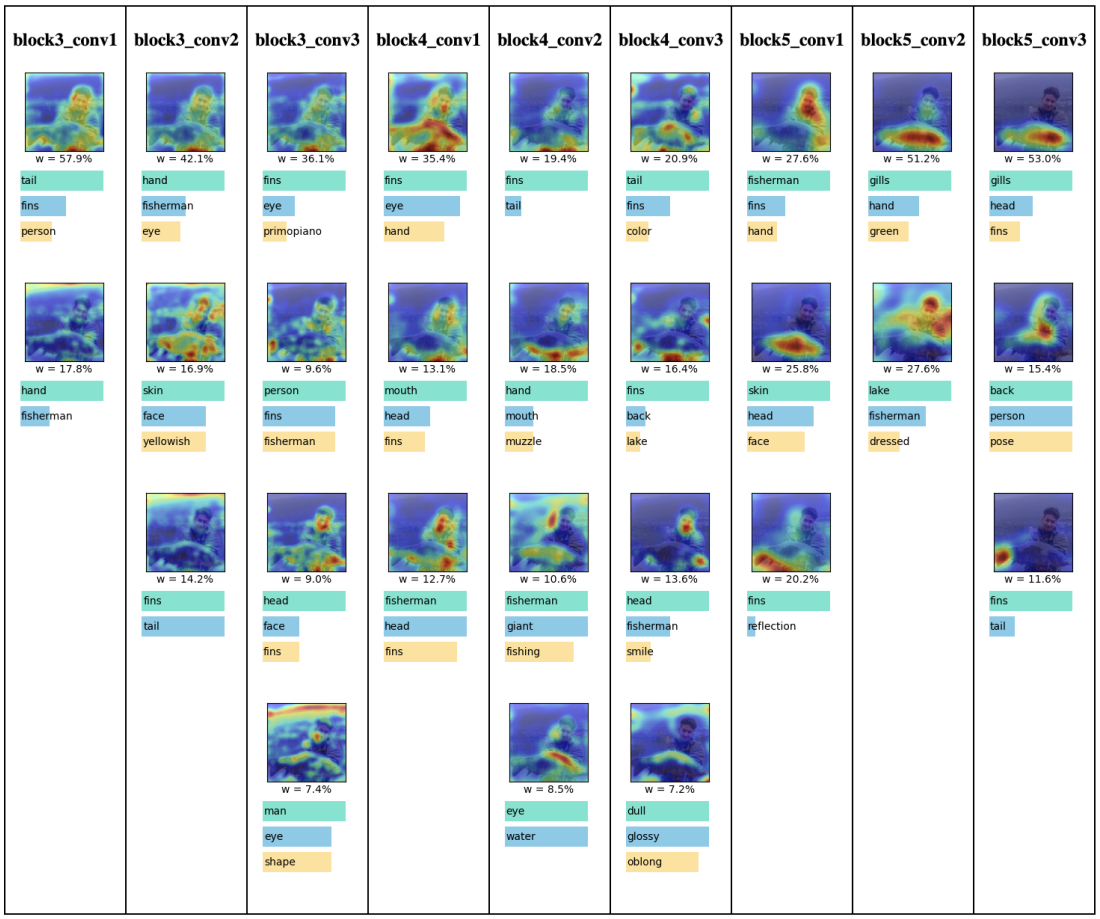


Figure 6.9: *An INV for an image of a trench showcases the labelled features identified by the network in the last convolutional layers. In this case, the most significant features for the classification were its gills and fin, as well as the lake and the fisherman. Other features were also important (, eye, mouth, and scales).*

Tench INV (Figure 6.9). In this INV, an image of a Tench is provided to the network, which identifies the *gills*, *back fin*, and the *person* in the background to classify the entity. The latter contributes towards the prediction across all the inspected layers, sometimes even achieving a higher relevance than other class-discriminating features, revealing a slight overfitting. Other features (*e.g.*, the *lake* and the *background*) also achieve a similar effect, although they are not considered relevant towards the final layer. On the other hand, relevant features like the fish's *scales*, *eyes*, and *mouth* were not considered as such by the network, especially in the deepest layers. However, further inspection revealed that inputting an image with a pattern of the fish's scales drove the network to predict the *Tench* class with high confidence, demonstrating it properly learned this feature and associated it with the right class.

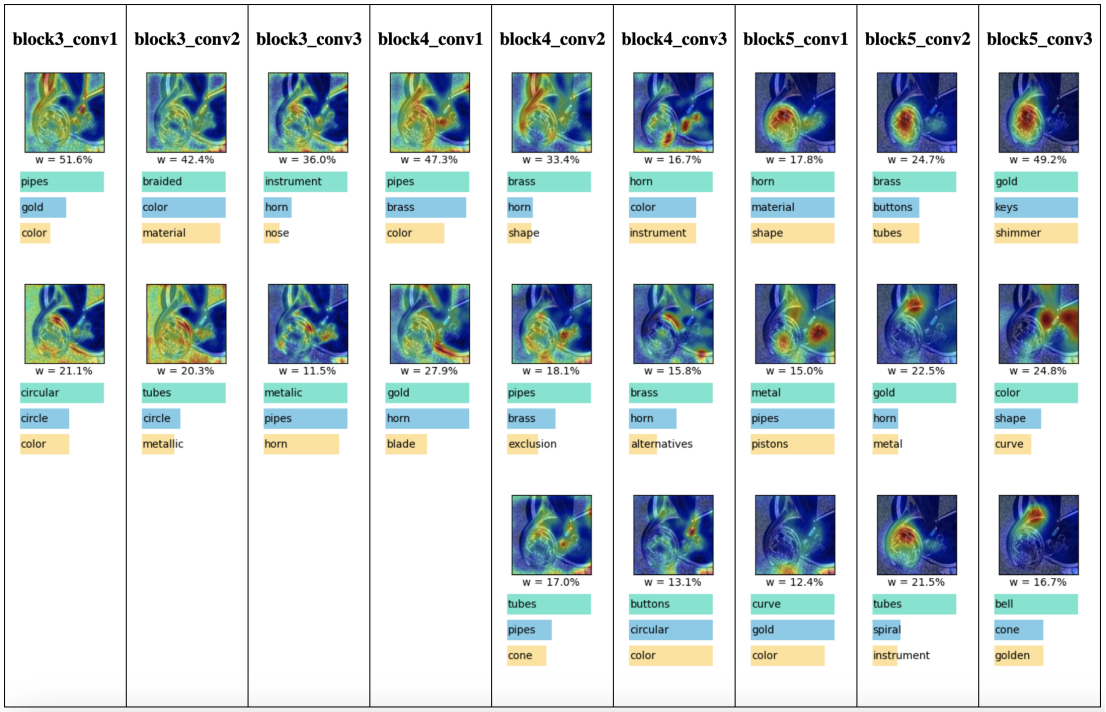


Figure 6.10: *An INV for an image of a French horn showcases the labelled features identified by the network in the last convolutional layers. In this case, the most significant features were the brass, the pipes, and the horn.*

French Horn INV (Figure 6.10). In this last INV, the main features the network applied to classify an image of a French horn are displayed. In this case, the proposed method performed "semantically" poorly compared to other images. In particular, most of the identified features are strongly tied to the material and the colour (*e.g.*, *gold*, *brass*, etc.). On the other hand, other context-specific features (*e.g.*, *pipes*, *bell*, etc.) were not considered as important. Such poor performance might be due to the inherent complexity of the class, which requires context-specific knowledge about musical instruments, leading most participants to focus on high-level or easily recognizable features only. On the other hand, a few participants provided detailed descriptions involving classes like *pipes* or *buttons*, although these were not considered relevant by the applied methodology and were filtered out in the process. This last example was reported to show that sometimes context-specific knowledge is essential towards achieving a proper INV.

The presented case studies revealed INVs' capability to provide a comprehensive overview of the most essential features for correct prediction, their importance, and the network's layer-wise decision-making process. Similarly, the approach provides insights on class unique features (*e.g.*, the colour of the chain saw's motor). On this last point, additional validation approaches should be applied as the fact that humans classified an entity using such class-unique features suggests the network might do the same. Furthermore, the lack of domain-specific knowledge might oversimplify the visualization (as discussed for the *French Horn* INV), highlighting the need for domain experts whenever complex entities are involved. However, while involving

	Understandability	Usefulness	Trustworthiness	Informativeness	Satisfaction
Grad-CAM	5.99 ±1.14	5.91 ±1.03	5.62 ±1.13	5.02 ±1.71	5.57 ±1.43
LIME	4.80 ±1.64	4.96 ±1.56	4.56 ±1.57	4.00 ±1.93	4.39 ±1.93
SHAP	4.95 ±1.51	4.87 ±1.53	4.66 ±1.56	4.14 ±1.90	4.18 ±2.06
Simpl INV	**6.05** ±1.17	5.90 ±1.06	**5.76** ±1.23	5.41 ±1.52	5.74 ±1.31
INV	5.92 ±1.40	**5.92** ±1.32	5.74 ±1.38	**5.82** ±1.50	**5.83** ±1.35

Table 6.7: *A table outlining the outcomes of the comparative analysis between INVs and other state-of-the-art methods. Although statistical significance was only observed in informativeness, INVs were slightly superior in most aspects.*

experts might result in detailed visualizations, they might overestimate the network's specificity as various classes with very different features might be involved. On the other hand, involving people with and without such expertise might provide different detail levels and broader views, finally leading to a better INV.

6.1.8 Comparative Analysis with Human Subjects

Two different types of INVs are compared against state-of-the-art XAI approaches under several XAI-related aspects. Its effectiveness is proven comparable (or better) to such methods.

A survey to compare INV with other state-of-the-art local XAI methods (*i.e.*, Grad-CAM, LIME, and SHAP) is reported to assess its capability to explain a model's decision-making process. SHAP's Gradient Explainer implementation was used to estimate Shapley values. A simplified version of INV was also considered, including only the final layer and a single label per cluster map. Participants were provided with a brief description clarifying each XAI approach and four instances of explanations from different classes. A set of randomly selected images among the ones employed in the experiments were selected and applied for all methods. Regarding the survey, the approach proposed by Aechtner [4] was applied, assessing understandability, usefulness, trustworthiness, informativeness, and satisfaction on a 7-point Likert scale. Information on whether participants had a background in AI was also collected. The questionnaire was completed by 165 participants, of which 71% had a background in AI. Results are shown in Table 6.7. At first sight, it can be noticed that SHAP and LIME consistently underperformed (one point on average) compared to INV, Simplified INV, and Grad-CAM. Although INV methods performed the best in all aspects, a t-test revealed a meaningful difference (p-value < 0.05) only in terms of informativeness. No meaningful difference between INV, Simplified INV, and Grad-CAM was identified concerning the other aspects. In conclusion, INV is superior in informativeness compared to state-of-the-art methods while at least equal in different aspects.

6.1.9 From INVs to Class-wise Representations

Finally, INVs are aggregated to generate Class-wise INVs to provide comprehensive representations for explaining CV models.

The generated INVs describe a model's local, layer-wise decision-making process, *i.e.*, a visualization for each input involving feature maps with weights and labels with scores. This representation is common to all INVs, making it easy to merge them into descriptive representations for each class. These might be performed by either aggregating multiple consecutive layers (*i.e.*, representing the different stages of the model's decision-making pipeline) or all layers (*i.e.*, representing the complete model's decision-making process). Such a representation is called c-INVs (class-wise INVs). While the proposed approach was mainly targeted at generating local explanations, a potential approach for generating such complete visualizations is presented. It is important to note that the quality of such an outcome depends on the number of images considered and labels collected and the quality of the generated INVs. Hence, this section describes some preliminary results that might not represent the classes considered in great detail.

c-INVs Generation. c-INVs are generated through a hierarchical aggregation of INVs, *e.g.*, layers are organized in groups of three, and the cluster maps from the considered INVs are merged. Layers can also be grouped together to generate a complete class-wise representation for which a TCAV score computed using 20-30 images and 50 runs is also provided. Considering the INVs computed for a class of choice, only the labels with the maximum score for each cluster map are kept as their score is a proxy for their trustworthiness and a label's capacity to describe a cluster map. Such labels drive the subsequent cluster map aggregation, generating new groups for the final class-wise representation. These aggregations are sorted based on their global score, *i.e.*, a score computed as the sum of the individual scores of each cluster map's label achieving the highest score[3]. Global scores measure label quality across multiple cluster maps, generalizing their individual scores. Such scores could have also been normalized to describe the distribution of labels across the layer groups. The final visualization also includes a subset of the cluster maps for which the label scored the highest. Such maps' weights are also averaged and assigned to each group. These visualizations provide an overview of the features learned by the network for a specific class and their importance.

Chainsaw c-INV. As a first example of a c-INV, the *chainsaw* class was considered, as represented in Figure 6.11. The proposed visualization aggregates layers in groups of three and considers only the top six labels as the other groups' computed global scores drop significantly. It can be observed that the network focused on the *sawchain* and the *handle* while also considering the *wood* and *tree* entities somewhat relevant.

[3]For instance, a global score of 4.2 indicates that approximately five users added that label across the three layers and all images of that class

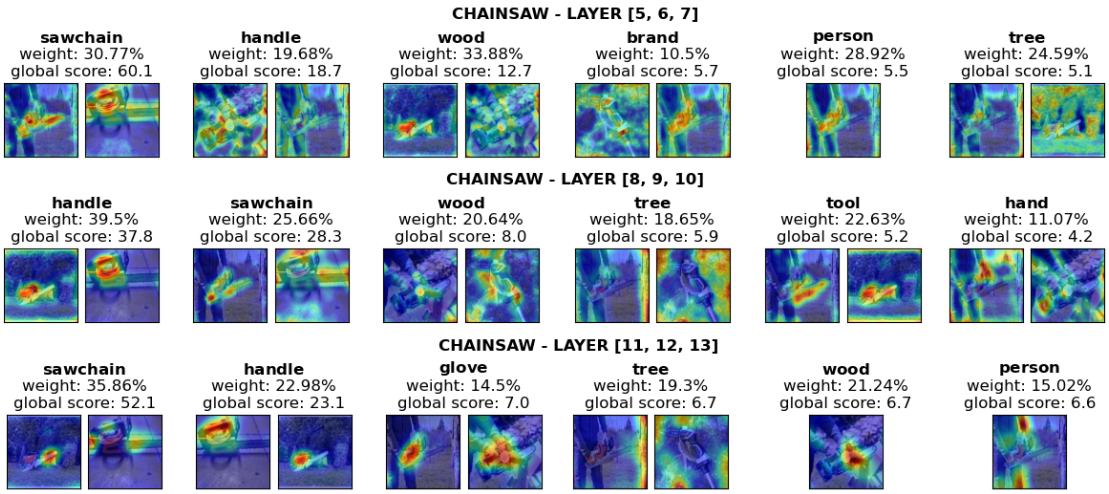


Figure 6.11: *A c-INV for the chainsaw class. For each row, the main features extracted from each group of layers are represented and detailed with a weight representing the feature's importance towards the prediction and a global score measuring the label's trustworthiness. In this case, the sawchain and the handle are the most important extracted features.*

Tench c-INV. A c-INV for the *tench* class is represented in Figure 6.12. In the proposed representation, the *fin* and the *eye* features are considered the most important. The *person* and *hand* features are still relevant towards the prediction, achieving a higher score in shallow layers and their corresponding groups. Such behaviour highlights the importance of considering a broad and heterogeneous dataset. In particular, when a broad dataset is considered, such a bias might not be reported as very few images might include a person. On the other hand, a heterogeneous dataset might allow an understanding of whether the network learned to classify images based on unwanted or biased elements.

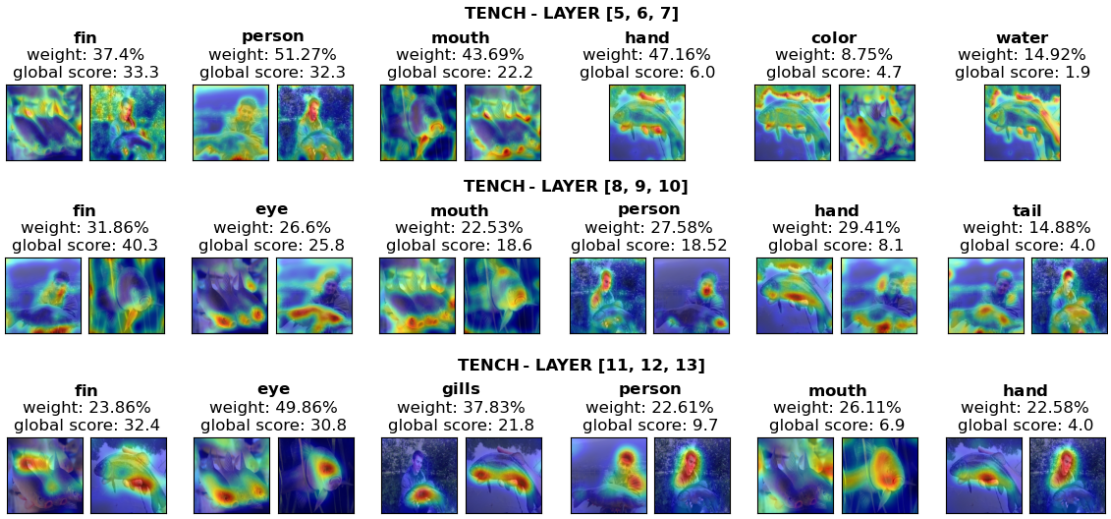


Figure 6.12: *A c-INV for the tench class. Such a visualization focuses on some tench's features, like fins, eyes, and gills, as well as external ones, like the human and their hand.*

As previously described, c-INVs might be generated with a varying number of

layer groups and a varying number of presented features (*e.g.*, based on a threshold of choice). In particular, all the inspected layers might be grouped, generating simple class-wise explanations representing the most critical features in the model's decision-making process. Figure 6.13 represents a complete c-INV for the *tench* class, highlighting that the tench's features (*e.g.*, fins, eyes, etc.) mainly drove the model's classification. However, the *person* and their features are relevant across most layers, too. Although these explanations might be intuitive and straightforward, they lack detailed insights (*e.g.*, they do not represent the network's decision-making process across different layers). Furthermore, weights are not defined as averaging weights from shallow and deep layers, which might cause a drop in relevance. Regarding TCAV scores, it was proven statistically significant for most features (p-value < 0.05). The only one with a slightly lower score was the *eye* feature, showing a slight discrepancy mainly due to the difference in computation between INV and TCAV (*i.e.*, the first uses the gradients of the pre-softmax score while the latter uses the gradients of the loss). Combining these methods achieves better explainability as the first generates the explanations and their labels while the second verifies their reliability. The results for every class are available on GitHub[4].

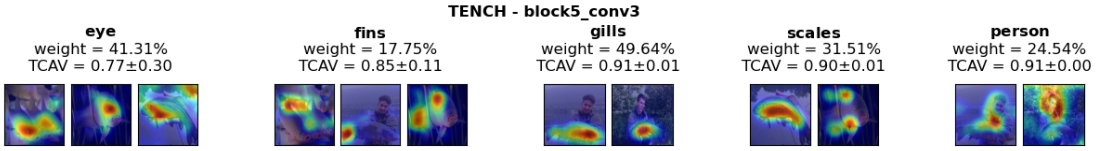


Figure 6.13: *A simple c-INV for the tench class, generated by grouping the features extracted from the last layer and summarizing the most important features.*

6.1.10 Deep Reveal Assessment

A questionnaire on *Deep Reveal* was answered by the participants, providing insights into its usability and workload while collecting useful feedback to improve its implementation.

As previously mentioned, a questionnaire was set in place to validate Deep Reveal's usability and workload (see Additional Material 7.2 for more details). Feedback on the activity and potential improvements was also collected and discussed here. The questionnaire was compiled by 152 out of 210 participants, resulting in a SUS score of 80.9/100, a NASA-TLX score of 38.1/100, and an overall user experience rating of 3.95/5. The final SUS score of 80.9 corresponds to a percentile ranking of 90, consequently representing a system with good or excellent usability [51]. The final NASA-TLX score[5] of 38.1 is considered somewhat high [336], suggesting the proposed approach still has room for improvement in terms of workload. On this point, guessing and labelling masked images might require effort as they might not be simple to interpret for humans. Indeed, the choice of the pictures to guess was one of

[4]Additional c-INVs - https://github.com/Antonio-Dee/interpretable-network-visualizations/tree/main/Additional%20Results (Last Accessed 13 November 2024)

[5]For practical reasons, a simplified version of the NASA-TLX that excludes the weighing step was preferred. Therefore, the resulting score should be considered more of an estimate than an accurate evaluation.

the most criticized aspects, mainly due to their complexity. Moreover, a significant amount of feedback reported that some matches were disproportionately more difficult than others, causing frustration and confusion. This might be due to some cluster maps not highlighting relevant information to discriminate between classes. Future improvements will balance match difficulty, generating masked cluster maps with higher percentiles to reveal a broader portion of the image whenever complex-to-understand images are involved. Similarly, difficulty levels might be introduced based on cluster maps' weight, as users used slightly fewer hints when playing with more important cluster maps (see Figure 6.14). Further feedback also revealed challenges in assigning labels to features users were unfamiliar with, emphasizing the importance of engaging experts with context-specific knowledge. An improved and varied gameplay, graphics, and score systems could also be implemented.

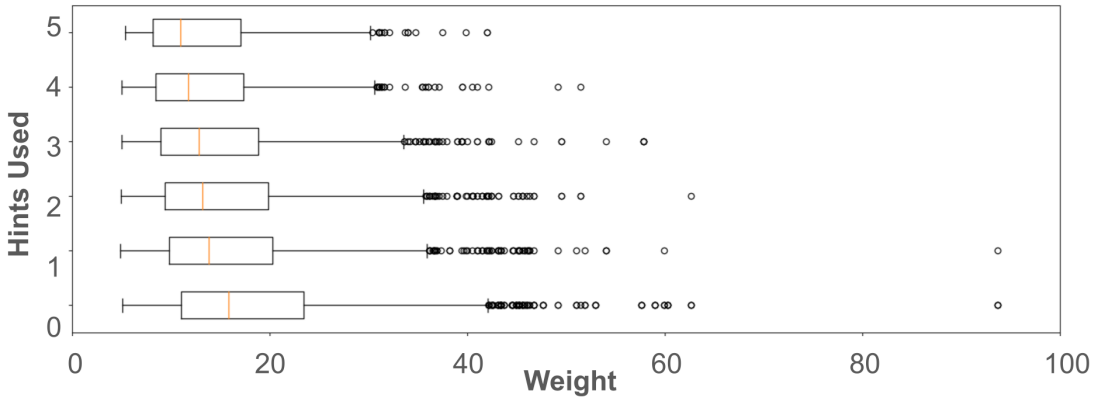


Figure 6.14: *A box plot depicting the number of hints used compared to the weight of the cluster maps. A slight trend reports users requesting more hints when feature maps with lower weights are presented. Furthermore, higher weights might correspond to easier-to-guess cluster maps since the number of hints requested can be seen as a proxy of the difficulty.*

6.1.11 Final Remarks

This chapter described the process, implementation, and validation of Interpretable Network Visualizations (INV), a novel approach for generating post-hoc local explanations for CNN-based image classification models. The proposed approach generates a detailed visualization of the features extracted from the input image at each layer through feature map clustering and merging. A weight is computed for each resulting cluster map to represent their contribution towards the final prediction. Human-understandable labels were collected and assigned to such maps by engaging humans in an online image-guessing game named *Deep Reveal*. Such labels were cleaned and processed, dropping further groupings based on the most relevant ones collected for each map at each layer. Finally, INVs were generated and further aggregated to create class-wise representations named c-INV, explaining a model's decision-making process with various levels of detail. INVs and a simplified version were assessed, revealing an effectiveness at least comparable with the most employed state-of-the-art techniques, if not slightly better.

Future works will involve improvements in the labelling step, including but not limited to extending *Deep Reveal* to collect labels to incorrectly predicted images, imple-

menting approaches to methodically validate label's correctness, and assessing the potential application of LLMs and multimodal models (*e.g.*, Llama [447]) in the proposed process, potentially easing the need for extensive crowdsourcing. Furthermore, extending the approach to labelling the network's neurons is another future research topic of interest as it might result in a single labelling iteration and on-the-fly explainability.

CHAPTER 7

Conclusion

This PhD dissertation describes the research performed by the candidate in the context of explainable AI. Several topics and contexts were studied, and the results obtained for each were reported.

The initial background describes the literature about XAI, gamification, crowdsourcing and a combination of such topics. The field of XAI is described, providing details on why explaining AI models is essential and the main definitions utilized in the literature. Context-specific background about each topic is provided within each chapter. Crowdsourcing is then explored, describing its main features. Gamification is theoretically described, focusing on the main principles that drive its usage. Furthermore, its primary design patterns and principles are described in the context of XAI.

The subsequent chapter focuses on the research on crowdsourcing and gamification. The candidate developed multiple software applications by employing principles to drive long-lasting engagement, finally providing the design of an XAI-centred platform bridging the gap between practitioners and novices.

As one of the main topics involved in the research is the employment of human knowledge in XAI, a chapter organizing their role in this context is proposed. It describes how humans have been involved in the different steps of the ML and XAI pipeline, reporting and discussing examples from the literature. Similarly, their contribution towards improving the robustness of AI models is presented.

XAI in the context of Natural Language Processing and Computer Vision is then described, reporting on the developed approaches. In the first context, a formalization to organize human knowledge was created. A novel XAI method aimed at enhancing model interpretability through human-interpretable labels is described in the latter. Humans were actively employed in these methods, using their knowledge as the method's core.

In conclusion, this dissertation provides evidence of humans' role in such a fundamentally performance-driven context (*i.e.*, mainly focused on improving model performance rather than their understandability), as explainability has risen in importance only over the last decade. Indeed, humans must be involved in most AI applications, as the main objective of XAI is to make such systems trustworthy, interpretable, and usable in real-world scenarios.

Following, some takeaways and perspectives learned from the research performed are outlined. Moreover, a brief discussion on the applicability of such techniques to late-breaking Large Language Models (LLMs) systems and Generative AI (GenAI) is provided. Finally, perspectives on future works are also provided.

Perspectives on Explaining Models. Explaining models in today's society is paramount for various reasons, *e.g.*, abide by regulations, improve users' trustworthiness, understand and enhance models' behaviour, and many more. While some researchers deemed explanations unnecessary in some cases, understanding a model's decision-making process is essential, regardless of the context. From a model's perspective, explanations must faithfully represent its behaviour. Indeed, faithfulness is one of the most important properties explanations must have to be considered useful, as non-faithful proofs of a model's behaviour will most likely mislead its users and (potentially) its developers.

As extensively discussed throughout this dissertation, explanations are essential to provide proof of a model's behaviour to a broad range of actors with different roles in the XAI cycle (as discussed in Chapter 4), *e.g.*, developers use this evidence to debug and fix or improve their models, users need explanations to trust these systems, etc. Explanations must be crafted to be as complete and understandable as possible, considering the actor's role and expertise as the main discerning factors. For example, example-based explanations and decision trees might be suited for final users as they are easy to understand and provide simple evidence of a system's behaviour. On the other hand, these might be too simple for experts or developers to completely understand a model's behaviour to ensure its proper functioning, hence requiring more complex representations or advanced and comprehensive tools. Moreover, recent studies demonstrated that background and education strongly impact humans' perception of explanations, making it essential to consider such aspects when designing XAI approaches [113].

Choosing the most suitable representation also depends on the context. This does not only depend on whether the system is built for Computer Vision, Natural Language Processing, etc. It also involves the actual context in which the model is applied (*e.g.*, banking, e-commerce, etc.) and the users' understanding of the topic. For example, providing an overview of the relevance of multiple complex banking parameters towards a model's decision might confuse the final user if not properly supported by context experts when inspecting them. Furthermore, whenever complex (*i.e.*, difficult to understand) explanations are generated, these might be extended and/or combined with other simpler and more understandable ones to achieve an even more interpretable outcome (as shown in Chapter 6). Context complexity is also relevant whenever data collection processes are set in place, as users might require specific knowledge to provide useful data properly.

Despite such additional complexity, mechanisms like gamification, games with a purpose, and properly engineered user engagement might contribute towards designing effective data collection activities by designing simple and enjoyable tasks for crowd workers that mask the inherent data or context complexity (as discussed in Chapter 3). Additionally, as AI is permeating everybody's daily life, it is more than ever relevant to educate users of AI-driven applications on such topics, providing people with tools and knowledge to deem a system trustworthy. In particular, bridging the gap between users and research communities might not only contribute towards this objective, but it might also be proven helpful in providing researchers with a trusted user base contributing to their research in various ways (*e.g.*, by providing data, thoughts, and perceptions). In conclusion, properly explaining AI models and systems is essential both from the user and model perspective, not only for their correct implementation but also for their rightful deployment and usage in today's society. Furthermore, users must be involved in the AI development loop as they have been proven to be extremely valuable assets for building, improving, and validating such systems.

A Methodological Approach for Designing Gamified Processes. Throughout this dissertation, most of the proposed approaches involved gamification or were implemented as games with a purpose whose design aimed at simplifying complex data collection tasks. Aside from the lessons learned for engaging users in such activities (as reported in Chapter 3), the design process starting from the problem is another relevant aspect to discuss. The essential elements to consider when designing a gamified activity are (1) the outcome to achieve and its features (*i.e.*, the data to collect, its structure, etc.), (2) the activity's input, (3) the process to be gamified, (4) the application's context, and finally (5) the users' expertise (about the context). When it comes to the expected outcome of the gamified process, this thesis mainly involved (structured) data to be collected from a given text or image (*e.g.*, labels). The process is usually designed based on the input and the outcome to achieve, as it represents how the latter is produced from the first, and it must involve all the necessary steps to (incrementally) produce the desired output. Gamification can be applied to the whole process or some of its steps, affecting the complete design of the application or parts of it, respectively. For example, one might consider designing applications based on well-known games when they fit the input and the desired outcome and when they ease and/or make the data collection process more enjoyable for the users involved.

For example, *Codenames* can be used as a complete gamified process to collect words that link groups of terms, which can then be used to build graphs representing such connections for explainability purposes. Furthermore, customizing the game to involve context-specific words would also allow for categorizing groups of terms. Small and targeted gamified elements can be employed whenever a comprehensive gamification approach can not be set in place. One of the most straightforward approaches involves awarding players with points, which can be used to build leaderboards. Similarly, achievements can be awarded when certain conditions are met. It is essential to define score- and achievement-awarding criteria based on the activity and its steps and test the effect of introducing such gamified elements. Furthermore, these components are more effective whenever simple and short-term data collection activities are set in place as they represent extrinsic motivation factors. Indeed, intrinsic motivation must

be leveraged whenever a long-lasting engagement is required.

The context and the users' expertise are strictly related, and the latter is essential to achieve a good task outcome, *i.e.*, whenever inexpert users are involved in an activity, the quality of the outcome might suffer from their lack of experience. For instance, it is essential to involve expert doctors when labelling clinical images in the medical context (regardless of their usage, *e.g.*, training or validating models or assessing XAI approaches). Such a need might be (sometimes partially) tackled by training users before the activity or by properly organizing its steps to simplify it. Sometimes, even splitting the process to assign users with different knowledge to different stages might be helpful whenever various areas of expertise are needed. Of course, their applicability still depends on the context, *e.g.*, it is pretty unthinkable to allow ordinary people to label complex medical images of tumours.

After the application's initial design is built, a preliminary test with users is highly advised to ensure its proper functioning and to discover potential gaps and flaws. If users are already engaged, this initial test might be performed on paper to avoid implementation overheads and costs. In this case, an intrinsic gap between the digital and the physical version must be considered when implementing the final application, *i.e.*, some problems or constraints might get fixed just by digitalizing the process (as experienced in Chapter 5). Testing the application by involving external actors, rather than its developers or designers, is strongly advised as it will provide novel insights and opinions. Such an approach can be applied iteratively to refine the tool before deploying it. Questionnaires to assess the systems' usability, complexity, and user effort might be helpful at this stage in understanding the developed application. User profiling might also be beneficial in getting even more insights into specific behaviours and/or biases. After a preliminary validation, the process can be implemented and set in place. Additional tests with small batches of users are advised to detect potential flaws in the final implementation. As the process is successfully validated, the data collection can start by involving users through the researcher's network or well-known crowdsourcing platforms.

Furthermore, it is advised for the final application to include attention checks to detect misbehaving users whose data is not to be considered towards the outcome. Regarding the data collection processes presented in this dissertation in the context of XAI, the desired process and its outcome influenced most of the application's design. In the gamified approach presented in Chapter 6, the activity was aimed at collecting labels describing the most important parts of the explanations of a classification model to enhance the final local explanation and generate class-wise ones. The objective was to drive the users to behave like a classification model, *i.e.*, classifying an image by focusing on the most relevant parts towards the model's prediction while describing these elements. Acknowledged the features of the use case (as previously listed in point (1)) and the desired behaviour, the crowdsourcing task was completely gamified through a *peek-a-boom*-like game design and further enhanced with simple gamified elements (*i.e.* points and leaderboards) since a short-term data collection process was developed. As most of the activities involved in this thesis employed (simple) images, the user's expertise did not affect the initial design. However, it was discovered that engaging knowledgeable users would have enhanced the outcome in some cases (*e.g.*, figures representing French horns). In the approach proposed in Chapter 5 to generate

Rationale Mappings, the level of detail of the explanation also affected the final design. Indeed, it involved an iterative approach aimed at refining the links between the users' thought process applied when performing the NLP task and its data, as the final explanation involves a multi-level tree in which the higher the depth of the tree, the higher the detail of the mapping. In conclusion, several elements must be considered when implementing gamified or data collection processes. It is fundamental to align the final design with their features (as initially listed in point (1)) to achieve a complete and enjoyable gamified application.

XAI in GenAI and LLMs. Even though this thesis does not directly address explainability applied to late-breaking advances in AI (*i.e.*, Generative AI and LLMs), some of the outcomes achieved and side projects [150] performed might still be relevant towards this topic. Generative AI is a computational technique capable of generating seemingly new, meaningful content from training data [127]. LLMs are just an example of such technologies for text generation. Regarding such systems' explainability, they are considered self-explainable models [190, 268] as they provide explanations for their reasoning, driving researchers to validate and explore the faithfulness [268] and features [190] of their explanations in various contexts and applications. In particular, their outcome can still be compared with the one collected from humans to validate other aspects, *e.g.*, their human-likeliness. For example, LLMs might be asked to produce Rationale Mappings (presented in Chapter 5), which will be compared with the ones extracted from humans to validate the human-likeliness of the reasoning applied by the system. Achieving high compatibility might result in improved trust from its users, especially when context-specific systems or models working in limited contexts are applied (*e.g.*, digital assistants for banking). Although this will not allow for deeper exploration of the model's internals, it will still provide an initial exploration for comparing human and model behaviour.

Additionally, human perception of LLMs' outcomes and communication is an essential aspect to investigate [215], as it shapes the system's perceived trustworthiness and consequently its applicability to real scenarios. Towards this objective, we deem analyzing people's reaction to LLMs answering questions represents their (emotional) trust towards AI models [150], especially when controversial questions are involved. In the considered use case, volunteer streamers performed live sessions asking thoughtfully crafted questions derived from the fundamental notion of trust in a GPT model. The crowd's reactions toward the model's answers were collected and analyzed, reporting similarities in cognitive and emotional reactions' distribution and intensity, and sentiment distribution (primarily neutral). Additionally, the use case presented in Chapter 3 involving a data collection activity for labelling images might help check a GenAI model's knowledge of specific entities. In particular, workers labelling images usually focus on features represented in the provided pictures. At the same time, GenAI models (*e.g.*, LLMs) acquire knowledge beyond an entity's representation, hence making representation-agnostic data collection processes interesting and more helpful compared to standard labelling.

Moreover, recent research has focused on assessing LLMs' capabilities in crowdsourcing (so-called AI Sourcing [84]), exploring their proficiency in providing human-like and human-aligned outcomes [45, 308]. Preliminary results demonstrated such

models' effectiveness by either supporting humans or behaving as such [45, 308], providing scalability with lower time constraints and cost. In this case, additional attention has to be paid to the quality of the collected data [217] since fundamental conclusions on their effectiveness are yet to be drawn. Furthermore, while this approach might provide benefits, it is still necessary to ensure that *human data remains human* [459] as it is more unique, diverse, and complete than synthetic data.

Open Challenges. This thesis addresses several topics regarding collecting, structuring, and representing human knowledge to enhance explanations through crowdsourcing and gamification. Nonetheless, some challenges are yet to be addressed, including but not limited to knowledge assessment, accounting for potential biases, and scalability.

Regarding knowledge assessment, it is necessary to consider which properties are essential, *e.g.*, completeness, quality, and more. In particular, some properties can be evaluated after the data has been processed. For example, the quality of the collected labels in INVs might be assessed after their collection by inspecting them (even through automated procedures), performing additional crowdsourcing on specific data instances whenever needed. Regarding other properties (*e.g.*, data coverage), one might design applications capable of covering all data instances of interest. Furthermore, data completeness is another fundamental aspect that depends on the context and subjectivity, *e.g.*, a few correct labels might be enough in simple cases. In contrast, many might be needed in complex ones.

Biases can be accounted for in the crowd's selection process and the design of the gamified application. In the first case, one might consider a heterogeneous crowd by design when crowdsourcing platforms are employed. Whenever crowds involving people with common features are considered, one might want to ensure the experiment is not affected by inherent biases. Furthermore, gamified applications might be designed to include (live) analysis to detect biases, perform additional iterations to collect broader knowledge whenever needed, and finally lead to an unbiased data collection.

Scalability can be achieved by involving a larger crowd in data collection activities, finally attaining larger datasets. Whenever such a solution is not applicable due to limited time or resources, open data sources can potentially be considered whenever the content is coherent with the research, and their structure can be tailored to extend the collected data. For example, one might consider labeled images from existing research to expand the one collected in INVs, extending the proposed pipeline to include an additional "integration step". Such inclusion should also consider the process employed to collect the data. Indeed, only labels strictly correlated with the most relevant areas at different layers must be considered in the mentioned case. Furthermore, these might also be weighted accordingly to account for their source. Additionally, one might develop gamified approaches to cover multiple data instances, limiting the impact on the complexity and effort through carefully designed applications. As a last option, one might employ GenAI models to generate human-like knowledge [84] or mimic human behaviour [45, 308] as previously described.

Future Work. This dissertation strictly focuses on human factors and perspectives in XAI, researching methods and approaches to use human knowledge and generate

human-understandable explanations. Future work will involve refining some of the techniques presented (*e.g.* the ones presented in Chapters 5 and 6). In particular, rationale mappings will be applied to specific use cases, like complex question-answering, further customizing and improving the mappings design (*e.g.*, their features, types, etc.) and their data collection process. INVs will be enhanced to improve the final representations. Additionally, potential improvements to intermediate steps (*e.g.*, clustering, label cleaning, and score computation) will be researched. Besides enhancing previously developed approaches, these will be adapted to be applied to modern Generative AI approaches (as previously discussed).

Additional Material

7.1 Additional Material Chapter 05

This last section provides the tutorial text for the QA task. The tutorials for the other tasks can be found at the following link `https://tinyurl.com/2s3s9w4v`

You will be shown a question, a paragraph, and an answer.
Here's an example we will consider in this tutorial:

Question
Where was the University of Paris located?

Paragraph
By the end of the 12th century, Paris had become the political, economic, religious, and cultural capital of France. Important areas were The Île de la Cité - the site of the royal palace, the easter extremity - for Notre Dame, and the Left Bank. The latter was the site of the University.

Answer
The Left Bank.

Step 1
In Step 1, we ask you to identify the Question Type. Usually, when we want to understand what a question is asking, we look for some keywords in it. For example, if the question contains the word "*when*", it is immediately clear that it is time-related.
Here's the list of question types you can pick from.

- **Yes/No Question**, i.e., questions that require looking for confirmation in the paragraph.
- **Wh-Question**, i.e., questions that require looking for the answer based on the type of wh-question (e.g., Who, What, etc.).
- **Choice Question**, i.e., questions that require picking the answer among the ones proposed in the question based on the paragraph.
- **Disjunctive or Tag Question**s, i.e., questions that require looking for confirmation in the paragraph

Questions belonging to the Wh-Question type are further specialized into the following categories:

Specialization	**Keywords**
Person	Who, Whose, Whom
Information	What, How
Location	Where
Time	When
Reason	Why, What for, How come, Why don't
Quantity	How many, How much, How far, How long, etc.
Choice	Which, Whom

The list of keywords indicates the words most frequently associated with the corresponding specialization, but some others may not be reported in the table.

Once you have identified the question type and possibly the specialization, we ask you to highlight the words in the question that allowed you to do so.

Question
Where was the University of Paris located?

Question Type: Wh-Question
Specialization: Location

Question - Paragraph (QP)
In this step, you will work on the question and paragraph. This step is further divided into three sub-steps.

QP - Step 1
In Step 1, you will find a list containing all the sentences in the paragraph. Amongst them, we ask you to tick those you consider useful for answering the question.

Question
Where was the University of Paris located?

Paragraph sentences

- ☐ By the end of the 12th century, Paris had become the political, economic, religious, and cultural capital of France.
- ☑ Important areas were The Île de la Cité - the site of the royal palace, the easter extremity - for Notre Dame, and the Left Bank.
- ☑ The latter was the site of the University.

From now on, you will work only on the sentences you selected, ignoring all of the others.

QP - Step 2
In Step 2, we ask you to identify the continuous portions of text in the paragraph that are useful for answering the question. While doing this, we also ask you to identify the portions of text in the question that allowed you to understand that a portion of the paragraph is important.

Question
Where was the University of Paris located?

Relevant paragraph sentences

- ☐ Important areas were The Île de la Cité - the site of the royal palace, the easter extremity - for Notre Dame, and the Left Bank.

- [ ] The latter was the site of the University.

QP - Step 3
In Step 3, we ask you to identify the single words in the paragraph that are useful for answering the question. While doing this, we also ask you to identify the single words in the question that allowed you to understand that a portion of the paragraph is important.

Question
Where was the University of Paris located?

Relevant paragraph sentences

- [ ] Important areas were The Île de la Cité - the site of the royal palace, the easter extremity - for Notre Dame, and the Left Bank.
- [ ] The latter was the site of the University.

Paragraph - Answer (PA)
In this step, you will work on the paragraph and the answer. This step is further divided into three sub-steps.

PA - Step 1
In Step 1, you will find a list containing all the sentences in the paragraph. Amongst them, we ask you to tick those you think have been used for providing the answer.

Answer
The Left Bank.

Paragraph sentences

- [ ] By the end of the 12th century, Paris had become the political, economic, religious, and cultural capital of France.
- [x] Important areas were The Île de la Cité - the site of the royal palace, the easter extremity - for Notre Dame, and the Left Bank.
- [ ] The latter was the site of the University.

From now on, you will work only on the sentences you selected, ignoring all of the others.

PA - Step 2
In Step 2, we ask you to identify the continuous portions of text in the paragraph that have been used for providing the answer. While doing this, we also ask you to identify the portions of text in the answer that allowed you to understand that a portion of the paragraph is important.

Answer
The Left Bank.

Relevant paragraph sentences

- ☑ Important areas were The Île de la Cité - the site of the royal palace, the easter extremity - for Notre Dame, and the Left Bank.

PA - Step 3
In Step 3, we ask you to identify the single words in the paragraph that have been used for providing the answer. While doing this, we also ask you to identify the single words in the answer that allowed you to understand that a portion of the paragraph is important.

In our example, it is not necessary to perform this step as the portions of text found in PA - Step 2 are identical.

Question - Answer (QA)
In this step, you will work on the question and the answer. This step is further divided into three sub-steps.

QA - Step 1
In Step 1, you will find a list containing all the sentences in the answer. Amongst them, we ask you to tick those that directly provide an answer to the question.

Question
Where was the University of Paris located?

Answer sentences

- ☑ The Left Bank.

From now on, you will work only on the sentences you selected, ignoring all of the others.

QA - Step 2
In Step 2, we ask you to map continuous portions of text in the question to the associated portions of text in the answer.

Question
Where was the University of Paris located?

Relevant answer sentences

- ☑ The Left Bank.

QA - Step 3
In Step 3, we ask you to map single words in the question to the associated single words in the answer.

In our example, it is not necessary to perform this step as there are no single words to map between question and answer.

Coreferences

As a last step, we ask you to identify the coreferences in the question, paragraph, and answer separately.
In a text, there is a coreference when two or more words, possibly different from one another, refer to the same entity.
We ask you to pick a different color for each group of words referring to the same entity.
Also, for each group, you should identify its “main entity”, i.e. the entity itself, to which the other words refer.
You should highlight only the coreferences that you deemed useful for solving the task.

Question
Where was the University of Paris located?

No coreferences in the question.

Paragraph
By the end of the 12th century, Paris had become the political, economic, religious, and cultural capital of France. Important areas were The Île de la Cité - the site of the royal palace, the easter extremity - for Notre Dame, and the Left Bank. The latter was the site of the University.

Main entity: the Left Bank

Answer
The Left Bank.

No coreferences in the answer.

7.2 Additional Material Chapter 06

Participant information

1. Please provide the email address that you used to register on Deep Reveal.

   _____

Usability

	Strongly Disagree				Strongly Agree
1. I think that I would like to play Deep Reveal often.	1	2	3	4	5
2. I found Deep Reveal to be unnecessarily complex.	1	2	3	4	5
3. I found Deep Reveal to be easy to use.	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use Deep Reveal.	1	2	3	4	5
5. I found the various functions of Deep Reveal were well integrated.	1	2	3	4	5
6. I thought there was too much inconsistency in Deep Reveal.	1	2	3	4	5
7. I would imagine that most people would learn to use Deep Reveal very quickly.	1	2	3	4	5
8. I found Deep Reveal very cumbersome to use.	1	2	3	4	5
9. I felt very confident using Deep Reveal.	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with Deep Reveal.	1	2	3	4	5

Workload

How mentally demanding was Deep Reveal?

0	1	2	3	4	5	6	7	8	9	10

Very Low Very High

How strenuous was it to use Deep Reveal?

0	1	2	3	4	5	6	7	8	9	10

Very Low Very High

How hurried or rushed was the pace of the game?

0	1	2	3	4	5	6	7	8	9	10

Very Low Very High

How successful were you in playing Deep Reveal?

0	1	2	3	4	5	6	7	8	9	10

Perfect Failure

How hard did you have to work to accomplish your level of performance?

0	1	2	3	4	5	6	7	8	9	10

Very Low Very High

How insecure, discouraged, irritated, stressed, and annoyed were you?

0	1	2	3	4	5	6	7	8	9	10

Very Low Very High

Feedback

1. Which of the following would you consider to be the weak points of the game? (Multiple choices allowed)
   A. Gameplay B. Graphics C. In-game images D. Score system
   E. Adding characteristics F. None of the above G. Other: _____

2. How did you find the overall experience?

1	2	3	4	5

3. Additional feedback (Optional)
   _____

Bibliography

[1] Maged Abdelaty, Sandra Scott-Hayward, Roberto Doriguzzi-Corin, and Domenico Siracusa. Gadot: Gan-based adversarial training for robust ddos attack detection. In *CNS*, pages 119–127. IEEE, 2021.

[2] Ahmed Abusnaina, Mohammed Abuhamad, Hisham Alasmary, Afsah Anwar, Rhongho Jang, Saeed Salem, DaeHun Nyang, and David Mohaisen. Dl-fhmc: Deep learning-based fine-grained hierarchical learning approach for robust malware classification. *Trans. on Dependable and Secure Computing*, 19(5):3432–3447, 2022.

[3] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[4] Jonathan Aechtner, Lena Cabrera, Dennis Katwal, Pierre Onghena, Diego Penroz Valenzuela, and Anna Wilbik. Comparing user perception of explanations developed with xai methods. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, 2022.

[5] Shazia Afzal, Arunima Chaudhary, Nitin Gupta, Hima Patel, Carolina Spina, and Dakuo Wang. *Data-Debugging Through Interactive Visual Explanations*, pages 133–142. Springer International Publishing, 05 2021.

[6] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *UAI*, pages 2114–2124. PMLR, 2021.

[7] Amira Ahmed and Frances Johnson. Gamification as a way of facilitating emotions during information-seeking behaviour: A systematic review of previous research. In Katharina Toeppe, Hui Yan, and Samuel Kai Wah Chu, editors, *Diversity, Divergence, Dialogue*, pages 85–98, Cham, 2021. Springer International Publishing.

[8] Sheikh Waqas Akhtar, Saad Rehman, Mahmood Akhtar, Muazzam A. Khan, Farhan Riaz, Qaiser Chaudry, and Rupert Young. Improving the robustness of neural networks using k-support norm based adversarial training. *IEEE Access*, 4:9501–9511, 2016.

[9] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023.

[10] Ahmed Allam, Zlatina Kostova, Kent Nakamoto, Peter Johannes Schulz, et al. The effect of social support features and gamification on a web-based intervention for rheumatoid arthritis patients: randomized controlled trial. *Journal of medical Internet research*, 17(1):e3510, 2015.

[11] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods, 2018.

[12] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks, 2018.

[13] David Alvarez-Melis, Harmanpreet Kaur, Hal Daumé III, Hanna M. Wallach, and Jennifer Wortman Vaughan. A human-centered interpretability framework based on weight of evidence. *CoRR*, abs/2104.13299, 2021.

[14] Kasun Amarasinghe, Kevin Kenney, and Milos Manic. Toward explainable deep neural network based anomaly detection. In *HSI*, pages 311–317, 2018.

[15] Isaac Ampomah, James Burton, Amir Enshaei, and Noura Al Moubayed. Generating textual explanations for machine learning models performance: A table-to-text task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3542–3551, Marseille, France, June 2022. European Language Resources Association.

[16] Emily Amspoker and Miriam R L Petruck. A gamified approach to frame semantic role labeling. In Eduard Dragut, Yunyao Li, Lucian Popa, Slobodan Vucetic, and Shashank Srivastava, editors, *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, pages 37–42, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[17] Avishek Anand, Kilian Bizer, Alexander Erlei, Ujwal Gadiraju, Christian Heinze, Lukas Meub, Wolfgang Nejdl, and Bjoern Steinroetter. Effects of algorithmic decision-making and interpretability on human behavior: Experiments using crowdsourcing. In *AAAI Conference on Human Computation & Crowdsourcing*, 12 2018.

[18] Ariful Islam Anik and Andrea Bunt. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *CHI*, pages 1–13, 2021.

[19] ML Anupama, P Vinod, Corrado Aaron Visaggio, MA Arya, Josna Philomina, Rincy Raphael, Anson Pinhero, KS Ajith, and P Mathiyalagan. Detection and robustness evaluation of android malware classifiers. *Journal of Computer Virology and Hacking Techniques*, pages 1–24, 2021.

[20] Paolo Arcaini, Andrea Bombarda, Silvia Bonfanti, and Angelo Gargantini. Dealing with robustness of convolutional neural networks for image classification. In *AITest*, pages 7–14, 2020.

[21] Stephanie Armbruster and Valentin Klotzbücher. Lost in lockdown? covid-19, social distancing, and mental health in germany. Diskussionsbeiträge 2020-04, University of Freiburg, Wilfried Guth Endowed Chair for Constitutional Political Economy and Competition Policy, Freiburg i. Br., 2020.

[22] Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. Marta: Leveraging human rationales for explainable text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5868–5876, May 2021.

[23] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online, November 2020. Association for Computational Linguistics.

[24] Shriya Atmakuri, Tejas Chheda, Dinesh Kandula, Nishant Yadav, Taesung Lee, and Hessel Tuinhof. Robustness of explanation methods for nlp models, 2022.

[25] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing, 2021.

[26] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In *NeurIPS*, volume 34, pages 5644–5655. Curran Associates, Inc., 2021.

[27] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Ready player one! eliciting diverse knowledge using a configurable game. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1709–1719, New York, NY, USA, 2022. Association for Computing Machinery.

[28] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. How can explainability methods be used to support bug identification in computer vision models? In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[29] Agathe Balayn, Natasa Rikalo, Jie Yang, and Alessandro Bozzon. Faulty or ready? handling failures in deep-learning computer vision models until deployment: A study of practices, challenges, and needs. In *CHI'23*, 2023.

[30] Agathe Balayn, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. What do you mean? interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the Web Conference 2021*, WWW '21, page 1937–1948, New York, NY, USA, 2021. Association for Computing Machinery.

[31] Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. Certifying geometric robustness of neural networks. In *NeurIPS*, volume 32. Curran Associates, Inc., 2019.

[32] Aaron Bangor, Phil Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *J. Usability Stud.*, 4:114–123, 04 2009.

[33] Oshrat Bar, Amnon Drory, and Raja Giryes. A spectral perspective of dnn robustness to label noise. In *AIStats*, volume 151 of *PMLR*, pages 3732–3752. PMLR, 28–30 Mar 2022.

[34] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *Knowledge Eng. Review*, 26:365–410, 12 2011.

[35] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[36] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya V. Nori, and Antonio Criminisi. Measuring neural net robustness with constraints. In *NeurIPS*, page 2621–2629. Curran Associates, 2016.

[37] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *CoRR*, abs/1704.05796, 2017.

[38] Kevin Bauer, Moritz von Zahn, and Oliver Hinz. Expl(ai)ned: The impact of explainable artificial intelligence on cognitive processes. *Leibniz Institute for Financial Research SAFE Working Paper Series*, 2021.

[39] Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel Valstar, Björn Schuller, and Elisabeth Andre. explainable cooperative machine learning with nova. *KI - Künstliche Intelligenz*, 34, 01 2020.

[40] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS*, pages 325–342. PMLR, 2021.

[41] Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. Enhancing robustness of machine learning systems via data transformations. In *CISS*, pages 1–5. IEEE, 2018.

[42] Devis Bianchini, Daniela Fogli, and Davide Ragazzi. Promoting citizen participation through gamification. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, New York, NY, USA, 2016. Association for Computing Machinery.

[43] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. Winogavil: Gamified association benchmark to challenge vision-and-language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26549–26564. Curran Associates, Inc., 2022.

[44] Aleksandar Bojchevski and S. Gunnemann. Certifiable robustness to graph perturbations. *NeurIPS*, 32, 2019.

[45] Francesco Bombassei De Bona, Gabriele Dominici, Tim Miller, Marc Langheinrich, and Martin Gjoreski. Evaluating explanations through llms: Beyond traditional user studies, 2024.

[46] Johan Bos and Malvina Nissim. Uncovering noun-noun compound relations by gamification. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 251–255, Vilnius, Lithuania, May 2015. Linköping University Electronic Press, Sweden.

[47] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Andrea Mauri, and Riccardo Volonterio. Pattern-based specification of crowdsourcing applications. In Sven Casteleyn, Gustavo Rossi, and Marco Winckler, editors, *Web Engineering*, pages 218–235, Cham, 2014. Springer International Publishing.

[48] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[49] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[50] John Brooke. *SUS – a quick and dirty usability scale*, pages 189–194. 01 1996.

[51] John Brooke. Sus: a retrospective. *Journal of Usability Studies*, 8:29–40, 01 2013.

[52] Abhigna B.S., Nitasha Soni, and Shilpa Dixit. Crowdsourcing – a step towards advanced machine learning. *Procedia Computer Science*, 132:632–642, 2018. International Conference on Computational Intelligence and Data Science.

[53] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541. ACM, 2006.

[54] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[55] Collin Burns and Jacob Steinhardt. Limitations of post-hoc feature alignment for robustness. In *CVPR*, pages 2525–2533, 2021.

[56] Garrick Cabour, Andrés Morales, Elise Ledoux, and Samuel Bassetto. Towards an explanation space to align humans and explainable-ai teamwork, 2021.

[57] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. Discovering and validating ai errors with crowdsourced failure reports. *ACM on Human-Computer Interaction*, 5(CSCW2):1–22, 2021.

[58] Marco Antonio Calijorne Soares and Fernando Silva Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646, 2020.

[59] Marco Antonio Calijorne Soares and Fernando Silva Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646, 2020.

[60] Stefano Calzavara, Lorenzo Cazzaro, Claudio Lucchese, Federico Marcuzzi, and Salvatore Orlando. Beyond robustness: Resilience verification of tree-based classifiers. *Computers & Security*, 121:102843, 2022.

[61] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations, 2018.

[62] Judy Cameron. Negative effects of reward on intrinsic motivation—a limited phenomenon: Comment on deci, koestner, and ryan (2001). *Review of Educational Research*, 71:29–42, 2001.

[63] Ginevra Carbone, Matthew Wicker, Luca Laurenti, A. Patane, L. Bortolussi, and Guido Sanguinetti. Robustness of bayesian neural networks to gradient-based attacks. In *NeurIPS*, volume 33, pages 15602–15613. Curran Associates, 2020.

[64] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[65] Lucas Carstens and Francesca Toni. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO, June 2015. Association for Computational Linguistics.

[66] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[67] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[68] Irene Celino. Who is this explanation for? human intelligence and knowledge graphs for explainable AI. *CoRR*, abs/2005.13275, 2020.

[69] Vanessa Cesário. Guidelines for combining storytelling and gamification: Which features would teenagers desire to have a more enjoyable museum experience? In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–6, New York, NY, USA, 2019. Association for Computing Machinery.

[70] Alvin Chan, Yi Tay, Yew Soon Ong, and Jie Fu. Jacobian adversarially regularized networks for robustness. *arXiv preprint arXiv:1912.10185*, 2019.

[71] C. Chang, G. Adam, and A. Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *2021 CVPR*, pages 15207–15216, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.

[72] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *CoRR*, abs/1710.11063, 2017.

[73] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. Machine explanations and human understanding, 2022.

[74] Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. Can rationalization improve robustness?, 2022.

[75] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples, 2017.

[76] Shang-Tse Chen, C. Cornelius, J. Martin, and D. Horng Chau. ShapeShifter: Robust physical adversarial attack on faster r-CNN object detector. In *Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2019.

[77] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *ECML/KDD*, pages 52–68. Springer, 2018.

[78] Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. Robust and accurate object detection via adversarial learning. In *CVPR*, pages 16622–16631, 2021.

[79] Yu Chen, Lingfei Wu, and Mohammed Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *NeurIPS*, volume 33, pages 19314–19326. Curran Associates, Inc., 2020.

[80] Minhao Cheng, Pin-Yu Chen, S. Liu, S. Chang, C.-J. Hsieh, and P. Das. Self-progressing robust training, 2020.

[81] Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. Cat: Customized adversarial training for improved robustness, 2020.

[82] Xiang Cheng, Yunzhe Hao, Jiaming Xu, and Bo Xu. Lisnn: Improving spiking neural networks with lateral interactions for robust object recognition. In *IJCAI*, pages 1519–1525, 7 2020.

[83] Seok-Hwan Choi, Jin-Myeong Shin, Peng Liu, and Yoon-Ho Choi. Argan: Adversarially robust generative adversarial networks for deep neural networks against adversarial examples. *IEEE Access*, 10:33602–33615, 2022.

[84] Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. Crowdsourcing or ai sourcing? *Commun. ACM*, 68(4):24–27, March 2025.

[85] Eric Chu, Nabeel Gillani, and Sneha Priscilla Makini. Games for fairness and interpretability. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 520–524, New York, NY, USA, 2020. Association for Computing Machinery.

[86] Eric Chu, Deb Roy, and Jacob Andreas. Are visual explanations useful? A case study in model-in-the-loop prediction. *CoRR*, abs/2007.12248, 2020.

[87] Eric Clarke, Tia DeNora, and Jonna Vuoskoski. Music, empathy and cultural understanding. *Physics of Life Reviews*, 15:61–88, 2015.

[88] Dennis Collaris and Jack J. van Wijk. Explainexplore: Visual exploration of machine learning explanations. In Fabian Beck, Jinwook Seo, and Chaoli Wang, editors, *2020 IEEE Pacific Visualization Symposium, PacificVis 2020 - Proceedings*, pages 26–35, United States, jun 2020. Institute of Electrical and Electronics Engineers. 13th IEEE Pacific Visualization Symposium, PacificVis 2020 ; Conference date: 14-04-2020 Through 17-04-2020.

[89] Bernat Coma-Puig and Josep Carmona. An iterative approach based on explainability to improve the learning of fraud detection models. *CoRR*, abs/2009.13437, 2020.

[90] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296:103471, 2021.

[91] Luca Console, Daniele Theseider Dupre, and Pietro Torasso. A theory of diagnosis for incomplete causal models. In *IJCAI*, pages 1311–1317, 1989.

[92] Alvaro H. C. Correia and Freddy Lecue. Human-in-the-loop feature selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2438–2445, Jul. 2019.

[93] Francesco Croce, M. Andriushchenko, V. Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *NeurIPS*, volume 1, 2021.

[94] Benjamin M.P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153, 2016.

[95] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China, December 2020. Association for Computational Linguistics.

[96] Kim de Bie, Ana Lucic, and Hinda Haned. To trust or not to trust a regressor: Estimating and explaining trustworthiness of regression predictions. *CoRR*, abs/2104.06982, 2021.

[97] Sam DeJohn. Beyond protest: Examining the decide madrid platform for public engagement, May 2018.

[98] Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving adversarial robustness via unlabeled out-of-domain data, 2020.

[99] Sebastian Deterding, Rilla Khaled, Lennart Nacke, and Dan Dixon. Gamification: Toward a definition. pages 6–9, 01 2011.

[100] Chaitanya Devaguptapu, Devansh Agarwal, Gaurav Mittal, Pulkit Gopalani, and Vineeth N Balasubramanian. On adversarial robustness: A neural architecture search perspective. In *ICCV*, pages 152–161, 2021.

[101] Jay DeYoung, Sarthak Jain, Nazneen Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron Wallace. Eraser: A benchmark to evaluate rationalized nlp models. pages 4443–4458, 01 2020.

[102] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models, 2019.

[103] Jonathan Dinu, Jeffrey P. Bigham, and J. Zico Kolter. Challenging common interpretability assumptions in feature attribution explanations. *ArXiv*, abs/2012.02748, 2020.

[104] Andrea Dittadi, Samuele Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. Generalization and robustness implications in object-centric learning, 2021.

[105] Ann-Kathrin Dombrowski, Christopher J Anders, Klaus-Robert Müller, and Pan Kessel. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121:108194, 2022.

[106] Minjing Dong, Yanxi Li, Yunhe Wang, and Chang Xu. Adversarially robust neural architectures, 2020.

[107] Yinpeng Dong, Qi-An Fu, X. Yang, T. Pang, H. Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness, 2019.

[108] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap?, 2022.

[109] Tianyu Du, Shouling Ji, Lujia Shen, Yao Zhang, Jinfeng Li, Jie Shi, Chengfang Fang, Jianwei Yin, Raheem Beyah, and Ting Wang. Cert-rnn: Towards certifying the robustness of recurrent neural networks. In *CCS*, pages 516–534, 2021.

[110] Xiaohu Du, Jie Yu, Shasha Li, Zibo Yi, Hai Liu, and Jun Ma. Combating word-level adversarial text with robust adversarial training. In *2021 Intl.Joint Conf. on Neural Networks (IJCNN)*, pages 1–8, 2021.

[111] Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *ICLR*, 2020.

[112] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), January 2023.

[113] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. The who in xai: How ai background shapes perceptions of ai explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[114] Upol Ehsan and Mark O. Riedl. Human-centered explainable AI: towards a reflective sociotechnical approach. *CoRR*, abs/2002.01092, 2020.

[115] Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. Fool me twice: Entailment from Wikipedia gamification. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online, June 2021. Association for Computational Linguistics.

[116] Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen, Daniel Keim, and Christopher Collins. Visual analytics for topic model optimization based on user-steerable speculative execution. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):374–384, 2019.

[117] Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021.

[118] Muhammed Elhadi, Ahmed Alsoufi, Ahmed Msherghi, Entisar Alshareea, Aimen Ashini, Taha Nagib, Nada Abuzid, Sanabel Abodabos, Hind Alrifai, Eman Gresea, Wisal Yahya, Duha Ashour, Salma Abomengal, Noura Qarqab, Amel Albibas, Mohamed Anaiba, Hanadi Idheiraj, Hudi Abraheem, Mohammed Fayyad, Yosra Alkilani, Suhir Alsuwiyah, Abdelwahap Elghezewi, and Ahmed Zaid. Psychological health, sleep quality, behavior, and internet use among people during the covid-19 pandemic: A cross-sectional study. *Frontiers in Psychiatry*, 12, 2021.

[119] Enrique Estellés-Arolas and Fernando González-Ladrón de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2):189–200, 2012.

[120] Enrique Estellés-Arolas and Fernando González-Ladrón de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2):189–200, 2012.

[121] Vladimir Estivill-Castro, Eugene Gilmore, and Rene Hexel. Human-in-the-loop construction of decision tree classifiers with parallel coordinates. pages 3852–3859, 10 2020.

[122] Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. *CoRR*, abs/2112.04417, 2021.

[123] Nils Feldhus, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Moller. Constructing natural language explanations via saliency map verbalization. *ArXiv*, abs/2210.07222, 2022.

[124] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *IJCAI-20*, pages 2206–2212, 7 2020. Main track.

[125] Shi Feng and Jordan L. Boyd-Graber. What can AI do for me: Evaluating machine learning interpretations in cooperative play. *CoRR*, abs/1810.09648, 2018.

[126] William Ferguson, Dhruv Batra, Raymond Mooney, Devi Parikh, Antonio Torralba, David Bau, David Diller, Josh Fasching, Jaden Fiotto-Kaufman, Yash Goyal, Jeff Miller, Kerry Moffitt, Alex Montes de Oca, Ramprasaath R. Selvaraju, Ayush Shrivastava, Jialin Wu, and Stefan Lee. Reframing explanation as an interactive medium: The equas (explainable question answering system) project. *Applied AI Letters*, 2(4):e60, 2021.

[127] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126, Feb 2024.

[128] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. Keep it simple: Reward and task design in crowdsourcing. In *Italian Chapter of SIGCHI*, pages 1–4, 2013.

[129] Andrea Fiorillo, Gaia Sampogna, Vincenzo Giallonardo, Valeria Del Vecchio, Mario Luciano, Umberto Albert, Claudia Carmassi, Giuseppe Carrà, Francesca Cirulli, Bernardo Dell'Osso, and et al. Effects of the lockdown on the mental health of the general population during the covid-19 pandemic in italy: Results from the comet collaborative network. *European Psychiatry*, 63(1):e87, 2020.

[130] Luciano Floridi. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1(6):261–262, 2019.

[131] James Fox and Sivasankaran Rajamanickam. How robust are graph neural networks to structural noise?, 2019.

[132] Scott Freitas, Shang-Tse Chen, Zijie J Wang, and Duen Horng Chau. Unmask: Adversarial detection and defense through robust feature alignment. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1081–1088. IEEE, 2020.

[133] Sorelle A. Friedler, Chitradeep Dutta Roy, Carlos Scheidegger, and Dylan Slack. Assessing the local interpretability of machine learning models. *CoRR*, abs/1902.03501, 2019.

[134] Yonty Friesem. Chapter 2 - empathy for the digital age: Using video production to enhance social, emotional, and cognitive skills. In Sharon Y. Tettegah and Dorothy L. Espelage, editors, *Emotions, Technology, and Behaviors*, Emotions and Technology, pages 21–45. Academic Press, San Diego, 2016.

[135] Laura Beth Fulton, Ja Young Lee, Qian Wang, Zhendong Yuan, Jessica Hammer, and Adam Perer. Getting playful with explainable ai: Games with a purpose to improve human understanding of ai. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery.

[136] Ujwal Gadiraju and Jie Yang. What can crowd computing do for the next generation of ai systems? In *CSW@NeurIPS*, 2020.

[137] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, Jan 2017.

[138] Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, and Yanjun Qi. Deepcloak: Masking deep neural network models for robustness against adversarial samples, 2017.

[139] Xiang Gao, Ripon K. Saha, Mukul R. Prasad, and Abhik Roychoudhury. Fuzz testing based data augmentation to improve robustness of deep neural networks. In *ICSE*, ICSE '20, page 1147–1158, New York, NY, USA, 2020. ACM.

[140] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[141] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2018.

[142] Simon Geisler, Tobias Schmidt, Hakan Şirin, Daniel Zügner, Aleksandar Bojchevski, and Stephan Günnemann. Robustness of graph neural networks at scale. In *NeurIPS*, volume 34, pages 7637–7649. Curran Associates, Inc., 2021.

[143] Margarita Geleta, Jiacen Xu, Manikanta Loya, Junlin Wang, Sameer Singh, Zhou Li, and Sergio Gago-Masague. Maestro: A gamified platform for teaching ai robustness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15816–15824, Jul. 2024.

[144] Anthony Gerber and Briann Fischetti. The impact of escape room gamification using a teleconferencing platform on pharmacy student learning. *Medical Science Educator*, 32, 2022.

[145] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of theAssoc. for Comp. Linguistics*, 9:346–361, 2021.

[146] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Klaus Mueller. Explainable active learning (XAL): an empirical study of how local explanations impact annotator experience. *CoRR*, abs/2001.09219, 2020.

[147] A Ghorbani and al. Towards automatic concept-based explanations. In *NeurIPS*, 2019.

[148] Sanjukta Ghosh, Rohan Shet, Peter Amon, Andreas Hutter, and André Kaup. Robustness of deep convolutional neural networks for image degradations. In *ICASSP*, pages 2916–2920, 2018.

[149] Marios M Giakalaras. Gamification and storytelling. *Univ. Aegean*, 8:1–7, 2016.

[150] Mathyas Giudici, Federica Liguori, Andrea Tocchetti, and Marco Brambilla. Unveiling human-ai interaction and subjective perceptions about artificial intelligent agents. In Kostas Stefanidis, Kari Systä, Maristella Matera, Sebastian Heil, Haridimos Kondylakis, and Elisa Quintarelli, editors, *Web Engineering*, pages 414–418, Cham, 2024. Springer Nature Switzerland.

[151] Claire Glanois, Paul Weng, Matthieu Zimmer, Dong Li, Tianpei Yang, Jianye Hao, and Wulong Liu. A survey on interpretable reinforcement learning. *Machine Learning*, 113(8):5847–5890, Aug 2024.

[152] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable ai: current status and future directions, 2021.

[153] Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. *Generalized but not Robust?* comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2705–2718, Dublin, Ireland, May 2022.

[154] PETER GOLDIE. Empathy with one's past. *The Southern Journal of Philosophy*, 49(s1):193–207, 2011.

[155] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Vice: Visual counterfactual explanations for machine learning models. *CoRR*, abs/2003.02428, 2020.

[156] Dou Goodman, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, and Zhang Huan. Advbox: a toolbox to generate adversarial examples that fool neural networks. *arXiv preprint arXiv:2001.05574*, 2020.

[157] Divya Gopinath, G. Katz, C S. Păsăreanu, and Clark Barrett. Deepsafe: A data-driven approach for assessing robustness of neural networks. In *Intl. symp. on automated technology for verification and analysis*, pages 3–19. Springer, 2018.

[158] Fabian Groh. Gamification: State of the art definition and utilization. *Proceedings of the 4th Seminar on Research Trends in Media Informatics*, pages 39–46, 01 2012.

[159] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, Apr 2022.

[160] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models, 2018.

[161] Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. FastIF: Scalable influence functions for efficient model interpretation and debugging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10333–10350, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[162] Lijie Guo, Elizabeth M. Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. Building trust in interactive machine learning via user contributed interpretable rules. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 537–548, New York, NY, USA, 2022. Association for Computing Machinery.

[163] Sidharth Gupta, P. Dube, and Ashish Verma. Improving the affordability of robustness training for dnns. In *CVPR*, June 2020.

[164] Jose L. Gómez-Urquiza, Juan Gómez-Salgado, Luis Albendín-García, María Correa-Rodríguez, Emilio González-Jiménez, and Guillermo A. Cañadas-De la Fuente. The impact on nursing students' opinions and motivation of using a "nursing escape room" as a teaching game: A descriptive study. *Nurse Education Today*, 72:73–76, 2019.

[165] Anthony Ha. Mindmixer raises $17m to help governments connect with their communities – techcrunch, Sep 2014.

[166] Christian Haase-Schütz, Rainer Stal, Heinz Hertlein, and Bernhard Sick. Iterative label improvement: Robust training by confidence based filtering and dataset partitioning, 2020.

[167] Tim Hahn, Ulrich Ebner-Priemer, and Andreas Meyer-Lindenberg. Transparent artificial intelligence – a conceptual framework for evaluating ai-based clinical decision support systems, Jan 2019.

[168] Juho Hamari and Janne Tuunanen. Player types: A meta-synthesis. *Transactions of the Digital Games Research Association*, 1:29–53, 03 2014.

[169] Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions, 2020.

[170] James Hardee. An overview of empathy. *The Permanente Journal*, 01 2003.

[171] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[172] Casper Harteveld, Sam Snodgrass, Omid Mohaddesi, Jack Hart, Tyler Corwin, and Guillermo Romera Rodriguez. The development of a methodology for gamifying surveys. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '18 Extended Abstracts, page 461–467, New York, NY, USA, 2018. Association for Computing Machinery.

[173] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, jul 2020. Association for Computational Linguistics.

[174] Alexander Heimerl, Katharina Weitz, Tobias Baur, and Elisabeth Andre. Unraveling ml models of emotion with nova: Multi-level explainable ai for non-experts. *IEEE Transactions on Affective Computing*, pages 1–1, 2020.

[175] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*, 2021.

[176] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

[177] P Henriksen, K Hammernik, D Rueckert, and A Lomuscio. Bias field robustness verification of large neural image classifiers. 2021.

[178] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[179] Julia Hirschberg and Christopher D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[180] P Hitzler and MK Sarker. Human-centered concept explanations for neural networks. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342(337):2, 2022.

[181] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[182] Fred Hohman, Arjun Srinivasan, and Steven M. Drucker. Telegam: Combining visualization and verbalization for interpretable machine learning. *IEEE Visualization Conference (VIS)*, 2019.

[183] Elizabeth A. Holm. In defense of the black box. *Science*, 364(6435):26–27, 2019.

[184] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[185] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020.

[186] Ramtin Hosseini, Xingyi Yang, and Pengtao Xie. Dsrna: Differentiable search of robust neural architectures, 2020.

[187] Jeff Howe. The rise of crowdsourcing, wired. *http://www. wired. com/wired/archive/14.06/crowds. html*, 2006.

[188] X Hu, H Wang, A Vegesana, and al. Crowdsourcing detection of sampling biases in image datasets. In *Proc. of WWW*, pages 2955–2961, 2020.

[189] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *2019 EMNLP-IJCNLP*, pages 2391–2401, 2019.

[190] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations, 2023.

[191] Miroslav Hudec, Erika Mináriková, Radko Mesiar, Anna Saranti, and Andreas Holzinger. Classification by ordinal sums of conjunctive and disjunctive functions for explainable ai and interpretable machine learning solutions. *Knowledge-Based Systems*, 220:106916, 2021.

[192] Christina Humer, Andreas Hinterreiter, Benedikt Leichtmann, Martina Mara, and Marc Streit. Reassuring, misleading, debunking: Comparing effects of xai methods on human decisions. *ACM Trans. Interact. Intell. Syst.*, may 2024. Just Accepted.

[193] Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 279–287, 2019.

[194] W.A. IJsselsteijn, Y.A.W. de Kort, and K. Poels. *The Game Experience Questionnaire*. Technische Universiteit Eindhoven, 2013.

[195] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *ISWC*, pages 486–504. Springer, 2014.

[196] Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proc. of the 58th Annual Meeting of theAssoc. for Comp. Linguistics*, pages 6740–6750, Online, July 2020. Assoc. for Comp. Linguistics.

[197] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 19–35. IEEE, 2018.

[198] Peter A. Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T. Morrison. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference, 2018.

[199] Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In *NeurIPS*, volume 33, pages 10558–10570. Curran Associates, Inc., 2020.

[200] Malhar Jere, Maghav Kumar, and Farinaz Koushanfar. A singular value perspective on model robustness. *arXiv preprint arXiv:2012.03516*, 2020.

[201] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4211–4222. Curran Associates, Inc., 2020.

[202] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*, volume 34, pages 8018–8025, 2020.

[203] Jonghoon Jin, Aysegul Dundar, and Eugenio Culurciello. Robust convolutional neural networks under adversarial noise, 2015.

[204] Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *SIGKDD*, page 66–74, New York, NY, USA, 2020. ACM.

[205] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. EUCA: A practical prototyping framework towards end-user-centered explainable artificial intelligence. *CoRR*, abs/2102.02437, 2021.

[206] Jose Nunes Da Junior, Antonio Leite, Jean-Yves Winum, Andrea Basso, Ulisses Sousa, David Nascimento, and Samuel Alves. Hsg400 – design, implementation, and evaluation of a hybrid board game for aiding chemistry and chemical engineering students in the review of stereochemistry during and after the covid-19 pandemic. *Education for Chemical Engineers*, 36, 04 2021.

[207] Takuhiro Kaneko, Yoshitaka Ushiku, and Tatsuya Harada. Label-noise robust generative adversarial networks, 2018.

[208] Rea Karachiwalla and Felix Pinkow. Understanding crowdsourcing projects: A review on the key design elements of a crowdsourcing initiative. *Creativity and Innovation Management*, 30(3):563–584, 2021.

[209] Pragya Katyayan and Nisheeth Joshi. Design and development of rule-based open-domain question-answering system on squad v2.0 dataset, 2022.

[210] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition, 2020.

[211] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744, Jan 2023.

[212] B Kim, M Wattenberg, J Gilmer, and al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In *ICML*, 2018.

[213] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). 2017.

[214] Hyun Kim, Minsoo Cho, and Seung-Hoon Na. ExplainMeetSum: A dataset for explainable meeting summarization aligned with human intent. In *Proc. of the 61st Annual Meeting of theAssoc. for Comp. Linguistics (Volume 1: Long Papers)*, pages 13079–13098, Toronto, Canada, July 2023. Assoc. for Comp. Linguistics.

[215] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA, 2024. Association for Computing Machinery.

[216] Marvin Klingner, Andreas Bar, and Tim Fingscheidt. Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation. In *CVPR Workshops*, June 2020.

[217] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. Chatgpt: Jack of all trades, master of none. *Inf. Fusion*, 99(C), November 2023.

[218] Pang Wei Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, Weihua Hu, Michihiro Yasunaga, R. L. Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, pages 5637–5664. PMLR, 2021.

[219] Konstantinos Kontoangelos, Marina Economou, and Charalambos Papageorgiou. Mental health effects of covid-19 pandemia: A review of clinical and psychological traits. *Psychiatry Investigation*, 17:491–505, 06 2020.

[220] Martijn J.L. Kors, Gabriele Ferri, Erik D. van der Spek, Cas Ketel, and Ben A.M. Schouten. A breathtaking journey. on the design of an empathy-arousing mixed-reality game. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '16, page 91–104, New York, NY, USA, 2016. Association for Computing Machinery.

[221] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *I. Journal of Computer Vision*, 129(3):736–760, 2021.

[222] Piotr Kotlinski, Xi-Jing Chang, Yang Chih-Yun, Wei-Chen Chiu, and Yung-Ju Chang. Using gamification to create and label photos that are challenging for computer vision and people. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '20 Adjunct, page 59–62, New York, NY, USA, 2020. Association for Computing Machinery.

[223] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online, November 2020. Association for Computational Linguistics.

[224] Maria Kouvela, Ilias Dimitriadis, and Athena Vakali. Bot-detective: An explainable twitter bot detection service with crowdsourcing functionalities. In *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, MEDES '20, page 55–63, New York, NY, USA, 2020. Association for Computing Machinery.

[225] Sean Kross and Philip Guo. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *ACM on Human-Computer Interaction*, 5(CSCW2):1–28, 2021.

[226] Ashwin Kumar, Stylianos Loukas Vasileiou, Melanie Bancilhon, Alvitta Ottley, and William Yeoh. Vizxp: A visualization framework for conveying explanations to users in model reconciliation problems. *Proceedings of the International Conference on Automated Planning and Scheduling*, 32(1):701–709, Jun. 2022.

[227] Emanuele La Malfa and Marta Kwiatkowska. The king is naked: on the notion of robustness for natural language processing. In *AAAI*, volume 36, pages 11047–11057, 2022.

[228] Emanuele La Malfa, Min Wu, L. Laurenti, B. Wang, A. Hartshorn, and Marta Kwiatkowska. Assessing robustness of text classification through maximal safe radius computation. In *EMNLP*, pages 2949–2968. ACL, November 2020.

[229] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):59–67, Oct. 2019.

[230] Isaac Lage and Finale Doshi-Velez. Learning interpretable concept-based models with human feedback. *CoRR*, abs/2012.02898, 2020.

[231] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior, 2018.

[232] Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *37th ICML*, ICML'20. JMLR.org, 2020.

[233] Paul Lam and Alan Tse. Gamification in everyday classrooms: Observations from schools in hong kong. *Frontiers in Education*, 6, 2022.

[234] Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. Qed: A framework and dataset for explanations in question answering, 2020.

[235] Fenareti Lampathaki, Carlos Agostinho, Yuri Glikman, and Michele Sesana. Moving from 'black box' to 'glass box' artificial intelligence in manufacturing with xmanai. In *2021 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, pages 1–6, 2021.

[236] Alfred Laugros, Alice Caplier, and Matthieu Ospici. Addressing neural network robustness with mixup and targeted labeling adversarial training, 2020.

[237] John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, December 2019.

[238] Hyungyu Lee, Ho Bae, and Sungroh Yoon. Gradient masking of label smoothing in adversarial robustness. *IEEE Access*, 9:6453–6464, 2021.

[239] Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *CHI*, pages 1–13, 2021.

[240] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions, 2016.

[241] Klas Leino, Z. Wang, and M. Fredrikson. Globally-robust neural networks. In *ICML*, volume 139, pages 6212–6222. PMLR, 18–24 Jul 2021.

[242] Regina Lenart-Gansiniec, Wojciech Czakon, Łukasz Sułkowski, and Jasna Pocek. Understanding crowdsourcing in science. *Review of Managerial Science*, 17(8):2797–2830, Nov 2023.

[243] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. FIND: human-in-the-loop debugging deep text classifiers. *CoRR*, abs/2010.04987, 2020.

[244] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350 – 1371, 2015.

[245] Alexander Levine and Soheil Feizi. Improved, deterministic smoothing for l_1 certified robustness, 2021.

[246] Dongfang Li, Baotian Hu, Qingcai Chen, Tujie Xu, Jingcong Tao, and Yunan Zhang. Unifying model explainability and robustness for joint text classification and rationale extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10947–10955, 2022.

[247] Dongyue Li and Hongyang Zhang. Improved regularization and robustness for fine-tuning in neural networks. In *NeurIPS*, volume 34, pages 27249–27262. Curran Associates, Inc., 2021.

[248] Jinfeng Li, Tianyu Du, Shouling Ji, Rong Zhang, Quan Lu, Min Yang, and Ting Wang. {TextShield}: Robust text classification based on multimodal embedding and neural machine translation. In *USENIX*, pages 1381–1398, 2020.

[249] Linyi Li, Zexuan Zhong, Bo Li, and Tao Xie. Robustra: Training provable robust neural networks over reference adversarial space. In *IJCAI*, pages 4711–4717, 2019.

[250] Xin Li, Xiangrui Li, Deng Pan, and Dongxiao Zhu. Improving adversarial robustness via probabilistically compact loss with logit constraints, 2020.

[251] Yanxi Li, Zhaohui Yang, Yunhe Wang, and Chang Xu. Neural architecture dilation for adversarial robustness. In *NeurIPS*, volume 34, pages 29578–29589. Curran Associates, Inc., 2021.

[252] Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie C. K. Cheung, and Siva Reddy. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment, 2022.

[253] Q. Vera Liao and Kush R. Varshney. Human-centered explainable ai (xai): From algorithms to user experiences, 2021.

[254] Rhema Linder, Sina Mohseni, Fan Yang, Shiva K. Pentyala, Eric D. Ragan, and Xia Ben Hu. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters*, 2(4):e49, 2021.

[255] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[256] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. Towards hybrid human-ai workflows for unknown unknown detection. In *Web Conf.*, pages 2432–2442, 2020.

[257] Tianyuan Liu, Hangbin Zheng, Jinsong Bao, Pai Zheng, Junliang Wang, Changqi Yang, and Jun Gu. An explainable laser welding defect recognition method based on multi-scale class activation mapping. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022.

[258] Zhe Liu, Yufan Guo, and Jalal Mahmud. When and why does a model fail? A human-in-the-loop error detection framework for sentiment analysis. *CoRR*, abs/2106.00954, 2021.

[259] Lissette López-Faican and Javier Jaen. Designing gamified interactive systems for empathy development. In *Companion Publication of the 2021 ACM Designing Interactive Systems Conference*, DIS '21 Companion, page 27–29, New York, NY, USA, 2021. Association for Computing Machinery.

[260] Yang Lou, Ruizi Wu, Junli Li, Lin Wang, Xiang Li, and Guanrong Chen. A learning convolutional neural network approach for network robustness prediction. *IEEE Transactions on Cybernetics*, 53(7):4531–4544, 2023.

[261] Xiaotian Lu, Arseny Tolmachev, Tatsuya Yamamoto, Koh Takeuchi, Seiji Okajima, Tomoyoshi Takebayashi, Koji Maruhashi, and Hisashi Kashima. Crowdsourcing evaluation of saliency-based xai methods. In Yuxiao Dong, Nicolas Kourtellis, Barbara Hammer, and Jose A. Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, pages 431–446, Cham, 2021. Springer International Publishing.

[262] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[263] Hongnan Ma, Kevin McAreavey, Ryan McConville, and Weiru Liu. Explainable ai for non-experts: Energy tariff forecasting. In *2022 27th International Conference on Automation and Computing (ICAC)*, pages 1–6, 2022.

[264] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[265] Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997.

[266] Divyam Madaan, Jinwoo Shin, and Sung Ju Hwang. Learning to generate noise for multi-attack robustness, 2020.

[267] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*. OpenReview.net, 2018.

[268] Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful?, 2024.

[269] Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. *CoRR*, abs/2110.08412, 2021.

[270] Lucie Charlotte Magister, Dmitry Kazhdan, Vikash Singh, and Pietro Liò. Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks. *CoRR*, abs/2107.11889, 2021.

[271] Nika Mahnic. Gamification of politics:start a new game! *Teorija in Praksa*, 51:143–161, 01 2014.

[272] Social Decision making Lab. Go viral!, January 2022.

[273] Ravi Mangal, Aditya V. Nori, and Alessandro Orso. Robustness of neural networks: A probabilistic and practical approach, 2019.

[274] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness, 2019.

[275] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *WACV*, pages 1859–1868, January 2021.

[276] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[277] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[278] Brad Miller, Alex Kantchelian, Sadia Afroz, Rekha Bachwani, E. Dauber, L. Huang, M. C. Tschantz, A. D. Joseph, and J Doug Tygar. Adversarial active learning. In *Workshop on Artificial Intelligent and Security*, pages 3–14, 2014.

[279] David J Miller, Xinyi Hu, Zhicong Qiu, and George Kesidis. Adversarial learning: a critical review and active learning study. In *Intl. Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

[280] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, pages 3243–3249, 2018.

[281] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, pages 7721–7735. PMLR, 2021.

[282] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *35th ICML*, volume 80, pages 3578–3586. PMLR, 10–15 Jul 2018.

[283] Swati Mishra and Jeffrey M. Rzeszotarski. Crowdsourcing and evaluating concept-driven explanations of machine learning models. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.

[284] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *FAccT*, pages 220–229, 2019.

[285] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Embedding human knowledge in deep neural network via attention map. *CoRR*, abs/1905.03540, 2019.

[286] Raquel Mochales and Marie-Francine Moens. Argumentation mining. *Artificial Intelligence and Law*, 19:1–22, 03 2011.

[287] Sina Mohseni, Jeremy E Block, and Eric Ragan. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 22–31, New York, NY, USA, 2021. Association for Computing Machinery.

[288] Jisoo Mok, Byunggook Na, Hyeokjun Choe, and Sungroh Yoon. Advrush: Searching for adversarially robust neural architectures, 2021.

[289] Andreea Molnar. The effect of interactive digital storytelling gamification on microbiology classroom interactions. In *2018 IEEE Integrated STEM Education Conference (ISEC)*, pages 243–246, 2018.

[290] Mohammad Momeny, Ali Mohammad Latif, Mehdi Agha Sarram, Razieh Sheikhpour, and Yu Dong Zhang. A noise robust convolutional neural network for image classification. *Results in Engineering*, 10:100225, 2021.

[291] Edwin Carlos Montiel-Vázquez, Jorge Adolfo Ramírez Uresti, and Octavio Loyola-González. An explainable artificial intelligence approach for detecting empathy in textual communication. *Applied Sciences*, 12(19), 2022.

[292] Seyed-Mohsen Moosavi-Dezfooli, Ashish Shrivastava, and Oncel Tuzel. Divide, denoise, and defend against adversarial attacks, 2018.

[293] Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations, 2021.

[294] Katelyn Morrison, Mayank Jain, Jessica Hammer, and Adam Perer. Eye into ai: Evaluating the interpretability of explainable ai techniques through a game with a purpose. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), oct 2023.

[295] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *CoRR*, abs/2008.00299, 2020.

[296] Aamir Mustafa, S. H. Khan, M. Hayat, R. Goecke, Jianbing Shen, and Ling Shao. Deeply supervised discriminative learning for adversarial defense. *Trans. on Pattern Analysis and Machine Intelligence*, 43(9):3154–3166, 2021.

[297] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In *Proceedings of the 44th International Conference on Software Engineering*, pages 413–425, 2022.

[298] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics.

[299] Vedant Nanda, Till Speicher, John P Dickerson, Krishna P Gummadi, and Muhammad Bilal Zafar. Unifying model explainability and robustness for joint text classification and rationale extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10947–10955, 2022.

[300] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. *CoRR*, abs/2004.14546, 2020.

[301] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *CoRR*, abs/1802.00682, 2018.

[302] Shweta Narkar, Yunfeng Zhang, Q Vera Liao, Dakuo Wang, and Justin D Weisz. Model lineupper: Supporting interactive model comparison at multiple levels for automl. In *26th Intl. Conf. on Intelligent User Interfaces*, pages 170–174, 2021.

[303] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness, 2020.

[304] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *CoRR*, abs/2201.08164, 2022.

[305] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana, jun 2018. Association for Computational Linguistics.

[306] Francisco Antonio Nieto-Escamez and María Dolores Roldán-Tapia. Gamification as online teaching strategy during covid-19: A mini-review. *Frontiers in Psychology*, 12, 2021.

[307] Kun-Peng Ning, Lue Tao, Songcan Chen, and Sheng-Jun Huang. Improving model robustness by adaptively correcting perturbation levels with active queries. In *EAAI*, pages 9161–9169. AAAI Press, 2021.

[308] Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, and Fabrice Lefèvre. Open-source large language models as multilingual crowdworkers: Synthesizing open-domain dialogues in several languages with no examples in targets and no machine translation, 2025.

[309] Ardavan Salehi Nobandegani, Kevin da Silva Castanheira, Timothy O'Donnell, and Thomas R Shultz. On robustness: An undervalued dimension of human rationality. In *CogSci*, page 3327, 2019.

[310] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS*, December 2021.

[311] Mahsan Nourani, Joanie T. King, and Eric D. Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. *CoRR*, abs/2008.09100, 2020.

[312] Mahsan Nourani, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan, Nicholas Ruozzi, and Vibhav Gogate. Don't explain without verifying veracity: An evaluation of explainable AI with video activity recognition. *CoRR*, abs/2005.02335, 2020.

[313] Mehdi Nourelahi, Lars Kotthoff, Peijie Chen, and Anh Nguyen. How explainable are adversarially-robust cnns? *arXiv preprint arXiv:2205.13042*, 2022.

[314] B Nushi, E Kamar, and al. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *AAAI*, 2017.

[315] Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards accountable AI: hybrid human-machine analyses for characterizing system failure. *CoRR*, abs/1809.07424, 2018.

[316] Haruna Ogawa, Hitoshi Nishikawa, Takenobu Tokunaga, and Hikaru Yokono. Gamification platform for collecting task-oriented dialogue data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7084–7093, Marseille, France, May 2020. European Language Resources Association.

[317] Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In Alexandra Balahur, Saif M. Mohammad, Veronique Hoste, and Roman Klinger, editors, *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[318] Christopher Olah, Ludwig Schubert, and Alexander Mordvintsev. Feature visualization. *Distill*, 2017.

[319] Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldemariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *CoRR*, abs/1908.01224, 2019.

[320] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA, 2016.

[321] Jonas Oppenlaender, Tahir Abbas, and Ujwal Gadiraju. The state of pilot study reporting in crowdsourcing: A reflection on best practices and guidelines. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), apr 2024.

[322] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1):89, 2021.

[323] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA, 2009. Association for Computing Machinery.

[324] Tianyu Pang, Huishuai Zhang, Di He, Yinpeng Dong, Hang Su, Wei Chen, Jun Zhu, and Tie-Yan Liu. Two coupled rejection metrics can tell adversarial examples apart. In *CVPR*, pages 15223–15233, 2022.

[325] Mervi Pantti and Minttu Tikka. *Cosmopolitan empathy and user-generated disaster appeal videos on YouTube*. Routledge, International, November 2013.

[326] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE SP*, pages 582–597, 2016.

[327] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *AISTATS*, volume 151, pages 4574–4594. PMLR, 28–30 Mar 2022.

[328] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, jan 2013.

[329] Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust, 2019.

[330] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. Investigating robustness and interpretability of link prediction via adversarial modifications. In *ACL*, pages 3336–3347, Minneapolis, Minnesota, June 2019. ACL.

[331] Mikki Phan, Joseph Keebler, and Barbara Chaparro. The development and validation of the game user experience satisfaction scale (guess). *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58, 09 2016.

[332] Maura Pintor, Daniele Angioni, Angelo Sotgiu, Luca Demetrio, Ambra Demontis, Battista Biggio, and Fabio Roli. Imagenet-patch: A dataset for benchmarking machine learning robustness against adversarial patches, 2022.

[333] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *ACM on Human-Computer Interaction*, 5(CSCW1):1–25, 2021.

[334] ROBERT PLUTCHIK. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press, 1980.

[335] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. Manipulating and measuring model interpretability. *CoRR*, abs/1802.07810, 2018.

[336] Atyanti Dyah Prabaswari, Chancard Basumerda, and Bagus Wahyu Utomo. The mental workload analysis of staff in study program of private educational organization. *IOP Conference Series: Materials Science and Engineering*, 528(1):012018, may 2019.

[337] Nicolas Pröllochs, Stefan Feuerriegel, and Dirk Neumann. Learning interpretable negation rules via weak supervision at document level: A reinforcement learning approach. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 407–413, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[338] Yada Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and Kai-Wei Chang. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. In *ACL-IJCNLP*, pages 3320–3331, August 2021.

[339] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 5582–5591, July 2019.

[340] Yanmin Qian, Hu Hu, and Tian Tan. Data augmentation using generative adversarial networks for robust speech recognition. *Speech Communication*, 114:1–9, 2019.

[341] Hamon R, Junklewitz H, and Sanchez Martin JI. Robustness and explainability of artificial intelligence. (KJ-NA-30040-EN-N (online)), 2020.

[342] Mashfiqui Rabbi, Meredith Philyaw-Kotov, Jinseok Lee, Anthony Mansour, Laura Dent, Xiaolei Wang, Rebecca Cunningham, Erin Bonar, Inbal Nahum-Shani, Predrag Klasnja, Maureen Walton, and Susan Murphy. Sara: A mobile app to engage users in health data collection. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, page 781–789, New York, NY, USA, 2017. Association for Computing Machinery.

[343] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy, 2020.

[344] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proc. of the 57th Annual Meeting of theAssoc. for Comp. Linguistics*, pages 4932–4942, Florence, Italy, July 2019. Assoc. for Comp. Linguistics.

[345] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018.

[346] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

[347] Qing Rao and Jelena Frtunikj. Deep learning for self-driving cars: Chances and challenges. In *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, pages 35–38, 2018.

[348] Arijit Ray, Giedrius Burachas, Yi Yao, and Ajay Divakaran. Lucid explanations help: Using a human-ai image-guessing game to evaluate machine explanation helpfulness. *CoRR*, abs/1904.03285, 2019.

[349] V C Raykar, S Yu, and al. Learning from crowds. *JMLR*, 11(Apr), 2010.

[350] Juan Carlo Rebanal, Yuqi Tang, Jordan Combitsis, Kai-Wei Chang, and Xiang 'Anthony' Chen. Xalgo: Explaining the internal states of algorithms via question answering. *CoRR*, abs/2007.07407, 2020.

[351] Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of theAssoc. for Comp. Linguistics*, 7:249–266, 2019.

[352] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

[353] Carl O. Retzlaff, Alessa Angerschmid, Anna Saranti, David Schneeberger, Richard Röttger, Heimo Müller, and Andreas Holzinger. Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research*, 86:101243, 2024.

[354] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. *AAAI*, 35(11):9419–9427, May 2021.

[355] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[356] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[357] Mireia Ribera Turró and Agata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. 03 2019.

[358] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. In *CHI*, pages 1–13, 2021.

[359] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1):107–136, 2006.

[360] Ellen Riloff and Michael Thelen. A rule-based question answering system for reading comprehension tests. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Sytems - Volume 6*, ANLP/NAACL-ReadingComp '00, page 13–19, USA, 2000. Association for Computational Linguistics.

[361] Christos Rodosthenous and Loizos Michael. A hybrid approach to commonsense knowledge acquisition. In *STAIRS 2016*, pages 111–122. IOS Press, 2016.

[362] Sudipta Singha Roy, Sk. Imran Hossain, M. A. H. Akhand, and Kazuyuki Murase. A robust system for noisy image classification combining denoising autoencoder and convolutional neural network. *Intl.Journal of Advanced Computer Science and Applications*, 9(1), 2018.

[363] Andras Rozsa, Manuel Gunther, and Terrance E. Boult. Towards robust deep neural networks with bang, 2016.

[364] Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska. Global robustness evaluation of deep neural networks with provable guarantees for the l_0 norm, 2018.

[365] Richard Ryan and Edward Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American psychologist*, 55:68–78, 01 2000.

[366] Marta Sabou, Kalina Bontcheva, and Arno Scharl. Crowdsourcing research opportunities: Lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '12, New York, NY, USA, 2012. Association for Computing Machinery.

[367] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.

[368] Michael Sailer, Jan Ulrich Hense, Sarah Katharina Mayr, and Heinz Mandl. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, 69:371–380, 2017.

[369] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[370] Wojciech Samek and Klaus-Robert Müller. *Towards Explainable Artificial Intelligence*, pages 5–22. Springer International Publishing, Cham, 2019.

[371] Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *ACL*, pages 1975–1985, New Orleans, Louisiana, June 2018. ACL.

[372] Antonio De Santis, Riccardo Campi, Matteo Bianchi, and Marco Brambilla. Visual-tcav: Concept-based attribution and saliency maps for post-hoc explainability in image classification, 2024.

[373] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. In *Conf. on Empirical Methods in Natural Language Processing*, 2019.

[374] Jeff Sauro. Measuring usability with the system usability scale (sus), February 2011.

[375] Philipp Schmidt and Felix Biessmann. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 431–449, Cham, 2020. Springer International Publishing.

[376] Martin Schuessler, Philipp Weiß, and Leon Sixt. Two4two: Evaluating interpretable machine learning - A synthetic dataset for controlled experiments. *CoRR*, abs/2105.02825, 2021.

[377] Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. F1 is not enough! models and evaluation towards user-centered explainable question answering. *CoRR*, abs/2010.06283, 2020.

[378] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. Human interpretation of saliency-based explanation over text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 611–636, New York, NY, USA, 2022. Association for Computing Machinery.

[379] Hendrik Schuff, Hsiu-Yu Yang, Heike Adel, and Ngoc Thang Vu. Does external knowledge help explainable natural language inference? automatic evaluation vs. human ratings. *CoRR*, abs/2109.07833, 2021.

[380] Katie Seaborn and Deborah I. Fels. Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74:14–31, 2015.

[381] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness?, 2021.

[382] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[383] Manisha Senadeera, Thommen Karimpanal George, Sunil Gupta, Stephan Jacobs, and Santu Rana. Emote: An explainable architecture for modelling the other through empathy, 2023.

[384] Rita Sevastjanova, Wolfgang Jentner, Fabian Sperrle, Rebecca Kehlbeck, Jürgen Bernard, and Mennatallah El-assady. Questioncomb: A gamification approach for the visual explanation of linguistic phenomena through interactive labeling. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), sep 2021.

[385] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time?, 2019.

[386] Rulin Shao, Z. Shi, J. Yi, P-Y. Chen, and C-J. Hsieh. On the adversarial robustness of vision transformers, 2021.

[387] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 882–892, New York, NY, USA, 2022. Association for Computing Machinery.

[388] Saima Sharmin, Nitin Rathi, Priyadarshini Panda, and Kaushik Roy. Inherent adversarial robustness of deep spiking neural networks: Effects of discrete input encoding and non-linear activations, 2020.

[389] Hua Shen and Ting-Hao Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):168–172, Oct. 2020.

[390] Max W Shen. Trust in ai: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient. *arXiv preprint arXiv:2202.05302*, 2022.

[391] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[392] Krist Shingjergji, Deniz Iren, Felix Böttger, Corrie Urlings, and Roland Klemke. Interpretable explainability in facial emotion recognition and gamification for data collection. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2022.

[393] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017.

[394] Dule Shu, Nandi O Leslie, Charles A Kamhoua, and Conrad S Tucker. Generative adversarial attacks against intrusion detection systems using active learning. In *Workshop on Wireless Security and Machine Learning*, pages 1–6, 2020.

[395] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017.

[396] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014.

[397] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[398] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin T. Vechev. Boosting robustness certification of neural networks. In *ICLR*, 2019.

[399] Vinay Singh, Iuliia Konovalova, and Arpan Kumar Kar. When to choose ranked area integrals versus integrated gradient for explainable artificial intelligence – a comparison of algorithms. *Benchmarking: An International Journal*, 30(9):3067–3089, Jan 2023.

[400] Sahil Singla, Surbhi Singla, and Soheil Feizi. Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100. In *ICLR*, 2022.

[401] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, pages 5972–5981, 2019.

[402] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods, 2019.

[403] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[404] Alison Smith and Jim Nolan. The problem of explanations without user feedback. *CEUR Workshop Proceedings*, 2068, mar 2018.

[405] Carol J Smith. Designing trustworthy ai: A human-machine teaming framework to guide development. *arXiv preprint arXiv:1910.03515*, 2019.

[406] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[407] Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *2020 FAccT*, pages 56–67, 2020.

[408] Kacper Sokol and Peter A. Flach. Explainability is in the mind of the beholder: Establishing the foundations of explainable artificial intelligence. *CoRR*, abs/2112.14466, 2021.

[409] Severine Soltani, Robert Kaufman, and Michael Pazzani. User-centric enhancements to explainable ai algorithms for image classification. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, number 44, 2022.

[410] Chang Song, Elias Fallon, and Hai Li. Improving adversarial robustness in weight-quantized neural networks, 2020.

[411] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *CoRR*, abs/1908.00087, 2019.

[412] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

[413] Digital Public Square. The coronavirus quiz, March 2022.

[414] Ramya Srinivasan and Beatriz San Miguel González. The role of empathy for artificial intelligence accountability. *Journal of Responsible Technology*, 9:100021, 2022.

[415] Joe Stacey, Yonatan Belinkov, and Marek Rei. Supervising model attention with human explanations for robust natural language inference. In *AAAI*, volume 36, pages 11349–11357, 2022.

[416] Matthew Staib. Distributionally robust deep learning as a generalization of adversarial training. 2017.

[417] Cor Steging, Silja Renooij, and Bart Verheij. Discovering the rationale of decisions: Experiments on aligning learning and reasoning. *CoRR*, abs/2105.06758, 2021.

[418] Alexander Steinmaurer, Martin Sackl, and Christian Gütl. Engagement in in-game questionnaires - perspectives from users and experts. In *2021 7th International Conference of the Immersive Learning Research Network (iLRN)*, pages 1–7, 2021.

[419] Nathalie Stembert and Ingrid Mulder. Love your city! an interactive platform empowering citizens to turn the public domain into a participatory domain. 05 2013.

[420] Julia Strout, Ye Zhang, and Raymond J. Mooney. Do human rationales improve machine explanations? *CoRR*, abs/1905.13714, 2019.

[421] E Štrumbelj and I Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 2014.

[422] David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *ICML*, pages 9155–9166. PMLR, 2020.

[423] Dong Su, H. Zhang, H. Chen, J. Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? a comprehensive study on the robustness of 18 deep image classification models. In *ECCV*, pages 644–661, Cham, 2018. Springer.

[424] Sabah Suhail, Mubashar Iqbal, Rasheed Hussain, and Raja Jurdak. Enigma: An explainable digital twin security solution for cyber–physical systems. *Computers in Industry*, 151:103961, 2023.

[425] Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020.

[426] Ke Sun, Zhanxing Zhu, and Zhouchen Lin. Enhancing the robustness of deep neural networks by boundary conditional gan. *arXiv preprint arXiv:1902.11029*, 2019.

[427] Weidi Sun, Yuteng Lu, Xiyue Zhang, Zhanxing Zhu, and Meng Sun. Global robustness verification networks, 2020.

[428] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017.

[429] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.

[430] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, page 109–119, New York, NY, USA, 2021. Association for Computing Machinery.

[431] Jesús Sánchez-Martín, Mario Corrales-Serrano, Amalia Luque-Sendra, and Francisco Zamora-Polo. Exit for success. gamifying science and technology for university students using escape-room. a preliminary approach. *Heliyon*, 6(7):e04340, 2020.

[432] Maite Taboada. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347, 2016.

[433] Jihoon Tack, Sihyun Yu, Jongheon Jeong, Minseon Kim, Sung Ju Hwang, and Jinwoo Shin. Consistency regularization for adversarial robustness. *AAAI*, 36(8):8414–8422, Jun. 2022.

[434] Savaş Takan, Duygu Ergün, and Gökmen Katipoğlu. Gamified text testing for sustainable fairness. *Sustainability*, 15(3), 2023.

[435] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. Commonsenseqa 2.0: Exposing the limits of ai through gamification, 2022.

[436] Wee-Kheng Tan and Chun Yu Hsu. The application of emotions, sharing motivations, and psychological distance in examining the intention to share covid-19-related fake news. *Online Information Review*, 47(1):59–80, Jan 2023.

[437] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, Philip H. S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques, 2021.

[438] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, volume 33, pages 18583–18599. Curran Associates, 2020.

[439] Matteo Terzi, Gian Antonio Susto, and Pratik Chaudhari. Directional adversarial training for cost sensitive deep learning classification applications. *Engineering Applications of Artificial Intelligence*, 91:103550, 2020.

[440] Dang Duy Thang and Toshihiro Matsui. Image transformation can make neural networks more robust against adversarial examples, 2019.

[441] Sarah-Kristin Thiel. Reward-based vs. social gamification: Exploring effectiveness of gamefulness in public participation. pages 1–6, 10 2016.

[442] Sarah-Kristin Thiel, T.P. Ertiö, and Matthias Baldauf. Why so serious? the role of gamification on motivation and engagement in e-participation. *Interaction Design and Architecture(s)*, pages 158–181, 12 2017.

[443] Sarah-Kristin Thiel and Peter Fröhlich. *Gamification as Motivation to Engage in Location-Based Public Participation?*, pages 399–421. 10 2017.

[444] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[445] V. Tjeng, K. Xiao, and R. Tedrake. Evaluating robustness of neural networks with mixed integer programming, 2017.

[446] Schrasing Tong and Lalana Kagal. Investigating bias in image classification using model explanations, 2020.

[447] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[448] Tamilla Triantoro, Ram Gopal, Raquel Benbunan-Fich, and Guido Lang. Personality and games: enhancing online surveys through gamification. *Information Technology and Management*, 21, Sept 2020.

[449] Yu-Lin Tsai, Chia-Yi Hsu, Chia-Mu Yu, and Pin-Yu Chen. Formalizing generalization and robustness of neural networks to weight perturbations, 2021.

[450] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR*, 2018.

[451] Hikaru Uchida, Masaki Matsubara, Kei Wakabayashi, and Atsuyuki Morishima. Human-in-the-loop approach towards dual process ai decisions. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3096–3098, 2020.

[452] Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. *Are Labels Required for Improving Adversarial Robustness?* Curran Associates Inc., Red Hook, NY, USA, 2019.

[453] Meet P. Vadera, Satya Narayan Shukla, Brian Jalaian, and Benjamin M. Marlin. Assessing the adversarial robustness of monte carlo and distillation methods for deep bayesian neural network classification, 2020.

[454] Pratik Vaishnavi, Tianji Cong, Kevin Eykholt, Atul Prakash, and Amir Rahmati. Can attention masks improve adversarial robustness?, 2019.

[455] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[456] Frans van Eemeren and Rob Grootendorst. A systematic theory of argumentation: The pragma-dialectical approach. *Argumentation*, 11 2003.

[457] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI conference*, pages 1–14, 2018.

[458] Noortje J. Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. Gamification for word sense labeling. In Alexander Koller and Katrin Erk, editors, *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 397–403, Potsdam, Germany, March 2013. Association for Computational Linguistics.

[459] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks, 2023.

[460] Maja Videnovik, Tone Vold, Georgina Dimova, Linda Vibeke Kiønig, and Vladimir Trajkovik. Migration of an escape room–style educational game to an online environment: Design thinking methodology. *JMIR Serious Games*, 10(3):e32095, Sep 2022.

[461] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *ArXiv*, abs/2006.00093, 2020.

[462] Marco Virgolin and Saverio Fracaros. On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence*, 316:103840, 2023.

[463] Luis Von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In *SIGCHI*, pages 75–78, 2006.

[464] Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of graph neural networks against adversarial structural perturbation, 2020.

[465] Danding Wang, Wencan Zhang, and Brian Y. Lim. Show or suppress? managing input uncertainty in machine learning model explanations, 2021.

[466] H. Wang and C-N. Yu. A direct approach to robust deep learning using adversarial networks. 2019.

[467] Haofan Wang, Mengnan Du, Fan Yang, and Zijian Zhang. Score-cam: Improved visual explanations via score-weighted class activation mapping. *CoRR*, abs/1910.01279, 2019.

[468] Jiakai Wang, Zixin Yin, Pengfei Hu, Aishan Liu, Renshuai Tao, Haotong Qin, Xianglong Liu, and Dacheng Tao. Defensive patches for robust recognition in the physical world. In *CVPR*, pages 2456–2465, 2022.

[469] Jun Wang, Changsheng Zhao, Junfu Xiang, and Kanji Uchino. Interactive topic model with enhanced interpretability. In *IUI Workshops*, 2019.

[470] Lijie Wang, Hao Liu, Shuyuan Peng, Hongxuan Tang, Xinyan Xiao, Ying Chen, Hua Wu, and Haifeng Wang. Dutrust: A sentiment analysis dataset for trustworthiness evaluation, 2021.

[471] Lijie Wang, Hao Liu, Shu ping Peng, Hongxuan Tang, Xinyan Xiao, Ying Chen, and Hua Wu. A sentiment analysis dataset for trustworthiness evaluation. *ArXiv*, abs/2108.13140, 2021.

[472] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I. Jordan. Robust optimization for fairness with noisy protected groups, 2020.

[473] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020.

[474] Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *EAAI*, pages 14024–14031. AAAI Press, 2021.

[475] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, Oct 2022.

[476] Stefan Webb, Tom Rainforth, Yee Whye Teh, and M. Pawan Kumar. A statistical approach to assessing neural network robustness, 2018.

[477] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach, 2018.

[478] Becky K. White, Annegret Martin, and James White. Gamification and older adults: Opportunities for gamification to support health promotion initiatives for older adults in the context of covid-19. *The Lancet Regional Health - Western Pacific*, page 100528, 2022.

[479] Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable natural language processing, 2021.

[480] Arie Wahyu Wijayanto, Jun Jin Choong, Kaushalya Madhawa, and Tsuyoshi Murata. Towards robust compressed convolutional neural networks. In *BigComp*, pages 1–8, 2019.

[481] Walt Woods, Jack Chen, and Christof Teuscher. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence*, 1(11):508–516, 2019.

[482] Peter Wright and John McCarthy. Empathy and experience in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 637–646, New York, NY, USA, 2008. Association for Computing Machinery.

[483] Yiting Wu and Min Zhang. Tightening robustness verification of convolutional neural networks with fine-grained linear approximation. In *AAAI*, volume 35, pages 11674–11681, 2021.

[484] Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France, May 2020. European Language Resources Association.

[485] Pulei Xiong, Scott Buffett, Shahrear Iqbal, Philippe Lamontagne, Mohammad Mamun, and Heather Molyneaux. Towards a robust and trustworthy machine learning system development: An engineering perspective. *Journal of Information Security and Applications*, 65:103121, 2022.

[486] Cong Xu, Xiang Li, and Min Yang. An orthogonal classifier for improving the adversarial robustness of neural networks. *Information Sciences*, 591:251–262, 2022.

[487] Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. How do we answer complex questions: Discourse structure of long-form answers, 2022.

[488] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *ICML*, volume 139, pages 11492–11501. PMLR, 18–24 Jul 2021.

[489] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, volume 32. Curran Associates, Inc., 2019.

[490] Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. Automatic text summarization methods: A comprehensive review, 2022.

[491] Y Yan, G M Fung, and al. Active learning from crowds. In *ICML*, pages 1161–1168, 2011.

[492] Ziang Yan, Yiwen Guo, and C. Zhang. Deep defense: Training dnns with improved adversarial robustness, 2018.

[493] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? 4 2020.

[494] J Yang, T Drake, A Damianou, and Y Maarek. Leveraging crowdsourcing data for deep active learning. an application: learning intents in alexa. In *WWW*, 2018.

[495] J Yang, A Smirnova, and al. Scalpel-cd: leveraging crowdsourcing and deep probabilistic modeling for debugging noisy training data. In *WWW*, pages 2158–2168, 2019.

[496] Pengfei Yang, J. Li, J. Liu, C-C. Huang, R. Li, L. Chen, X. Huang, and Lijun Zhang. Enhancing robustness verification for deep neural networks via symbolic propagation. *Formal Aspects of Computing*, 33, 06 2021.

[497] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September 2015. Assoc. for Comp. Linguistics.

[498] Yichen Yang, Xiaosen Wang, and Kun He. Robust textual embedding against word-level adversarial attacks, 2022.

[499] Yichen Yang, Xiaosen Wang, and Kun He. Robust textual embedding against word-level adversarial attacks, 2022.

[500] Yiwei Yang, Eser Kandogan, Yunyao Li, Prithviraj Sen, and Walter S. Lasecki. A study on interaction in human-in-the-loop machine learning for text analytics. In *IUI Workshops*, 2019.

[501] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Assoc. for Comp. Linguistics.

[502] Muneki Yasuda, Hironori Sakata, Seung-Il Cho, Tomochika Harada, Atushi Tanaka, and Michio Yokoyama. An efficient test method for noise robustness of deep neural networks. *IEICE*, 10:221–235, 01 2019.

[503] Dengpan Ye, Chuanxi Chen, Changrui Liu, Hao Wang, and Shunzhi Jiang. Detection defense against adversarial attacks with saliency map. *Intl. Journal of Intelligent Systems*, 2021.

[504] Qinyuan Ye, Xiao Huang, Elizabeth Boschee, and Xiang Ren. Teaching machine comprehension with compositional explanations, 2020.

[505] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. Interpreting and evaluating neural network robustness, 2019.

[506] Haizi Yu, Heinrich Taube, James A. Evans, and Lav R. Varshney. Human evaluation of interpretability: The case of ai-generated music knowledge. *CoRR*, abs/2004.06894, 2020.

[507] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*, 2019.

[508] Omar Zaidan, Jason Eisner, and Christine D. Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Assoc. for Comp. Linguistics*, 2007.

[509] Liang-Jun Zang, Cong Cao, Ya-Nan Cao, Yu-Ming Wu, and Cun-Gen Cao. A survey of commonsense knowledge acquisition. *Journal of Computer Science and Technology*, 28(4):689–719, 2013.

[510] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[511] Xianlong Zeng, Fanghao Song, Zhongen Li, Krerkkiat Chusap, and Chang Liu. Human-in-the-loop model explanation via verbatim boundary identification in generated neighborhoods. In Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 309–327, Cham, 2021. Springer International Publishing.

[512] Amy X Zhang, Michael Muller, and Dakuo Wang. How do data science workers collaborate? roles, workflows, and tools. *ACM on Human-Computer Interaction*, 4(CSCW1):1–23, 2020.

[513] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Trans. on Image Processing*, 30:1291–1304, 2021.

[514] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions, 2018.

[515] Li Zhang and Haiping Lu. A feature-importance-aware and robust aggregator for gcn. In *CIKM*, page 1813–1822. ACM, 2020.

[516] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 38629–38642. Curran Associates, Inc., 2022.

[517] Mengdi Zhang, Jun Sun, and Jingyi Wang. Which neural network makes more explainable decisions? an approach towards measuring explainability. *Automated Software Engineering*, 29(2):39, Apr 2022.

[518] Q Zhang and al. Interpretable convolutional neural networks. In *CVPR*, 2018.

[519] Xiao Zhang and David Evans. Understanding intrinsic robustness using label uncertainty, 2021.

[520] Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni. Certified robustness to programmable transformations in LSTMs. In *EMNLP*, pages 1068–1083. ACL, November 2021.

[521] Zijian Zhang, Koustav Rudra, and Avishek Anand. Faxplainac: A fact-checking tool based on explainable models with human correction in the loop. *CoRR*, abs/2110.10144, 2021.

[522] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. Dissonance between human and machine understanding. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[523] Wenqi Zhao, Satoshi Oyama, and Masahito Kurihara. Generating natural counterfactual visual explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[524] Yuxiang (Chris) Zhao and Qinghua Zhu. Effects of extrinsic and intrinsic motivation on participation in crowdsourcing contest: A perspective of self-determination theory. *Online Inf. Rev.*, 38(7):896–917, 2014.

[525] Zhenge Zhao, Panpan Xu, Carlos Scheidegger, and Liu Ren. Human-in-the-loop extraction of interpretable concepts in deep learning models. *CoRR*, abs/2108.03738, 2021.

[526] Qinkai Zheng, Xu Zou, Yuxiao Dong, Yukuo Cen, Da Yin, Jiarong Xu, Yang Yang, and Jie Tang. Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning. In *NeurIPS*, 2021.

[527] Xiaoqing Zheng, J. Zeng, Y. Zhou, C-J. Hsieh, Minhao Cheng, and Xuanjing Huang. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples. In *ACL*, pages 6600–6610, July 2020.

[528] Yiqi Zhong, Lei Wu, Xianming Liu, and Junjun Jiang. Exploiting the potential of datasets: A data-centric approach for model robustness, 2022.

[529] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[530] Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. Improving robustness of neural machine translation with multi-task learning. In *Conf. on Machine Translation*, pages 565–571, 2019.

[531] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *ICCV*, pages 16443–16452, October 2021.

[532] Gabe Zichermann. Intrinsic and extrinsic motivation in gamification. `https://www.gamification.co/2011/10/27/intrinsic-and-extrinsic-motivation-in-gamification/`, 2011. Last Accessed: 05 June 2024.

[533] Vadim Ziyadinov and Maxim Tereshonok. Noise immunity and robustness study of image recognition using a convolutional neural network. *Sensors*, 22(3), 2022.

[534] Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. Towards robust image classification using sequential attention models. In *CVPR*, June 2020.

[535] Marc-André Zöller, Waldemar Titov, Thomas Schlegel, and Marco F. Huber. Xautoml: A visual analytics tool for establishing trust in automated machine learning, 2022.