TANGO

D2.1 – Cognition-aware explanations for HML

Submission Date: 29/11/2024

Due Date: 30/11/2024

DOCUMENT SUMMARY INFORMATION

Grant Agreement No	101120763	**Acronym**	TANGO
Full Title	It takes two to tango: a synergistic approach to human-machine decision-making		
Start Date	01/10/2023	**Duration**	48 Months
Project Type	Research and Innovation Action (RIA)	**Funding Programme**	Horizon Europe
deliverable	D2.1 – Cognition-aware explanations for HML		
Type	R	**Dissemination Level**	PU
Lead Beneficiary	UNIVERSITA' DI PISA (UNIPI)		
Authors	Anna Monreale (UNIPI); Stefano Teso (UNITN)		
Co-authors	Riccardo Guidotti (UNIPI); Francesca Naretto (UNIPI); Tommaso Turchi (UNIPI); Roberto Pellungrini (SNS), Andrea Beretta (CNR), Salvatore Rinzivillo (CNR), Alan Dix (UoS); Ben Wilson (UoS), Andrea Passerini (UNITN); Novi Quadrianto (BCAM)		
Reviewers	Wolfgang Stammer (TUDA); Tim Tobiasch (TUDA); Nick Chater (UoW); Simon Myers (UoW)		

Funded by EU

DISCLAIMER

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO

COPYRIGHT

DOCUMENT HISTORY

Version	Date	Changes	Contributor(s)
V0.1	2024-09-27	Initial ToC	A. Monreale, S. Teso
V0.2	2024-10-15	First Draft	A. Monreale , S. Teso, A. Beretta, R. Guidotti
V0.3	2024-10-30	Second version	A. Monreale, S. Teso, S. Rinzivillo, R. Pellungrini
V0.4	2024-11-15	Third version	A. Monreale , S. Teso, F. Naretto, B. Wilson, A. Dix, T. Turchi, B. Sayin

V0.5	2024-11-22	Peer review	W. Stammer, T. Tobiasch, N. Chater, S. Myers
V0.6	2024-11-26	Address peer review comments	A. Monreale and S. Teso
V1.0	2024-11-29	Revised and corrected the final version	A. Passerini, S. Rinzivillo

PROJECT PARTNERS

Partner	Country	Short name
UNIVERSITA DEGLI STUDI DI TRENTO	IT	UNITN
TECHNISCHE UNIVERSITAT DARMSTADT	DE	TUDa
UNIVERSITA DI PISA	IT	UNIPI
CONSIGLIO NAZIONALE DELLE RICERCHE	IT	CNR
SCUOLA NORMALE SUPERIORE	IT	SNS
UNIVERSITE PARIS CITE	FR	UPC
FONDAZIONE BRUNO KESSLER	IT	FBK
CARR COMMUNICATIONS LIMITED	IE	CARR
ISTRAZIVACKO-RAZVOJNI INSTITUT ZA VESTACKU INTELIGENCIJU SRBIJE	RS	IVI
SURGICAL SCIENCE SWEDEN AB	SE	SuS
UNIVERSITATSKLINIKUM HEIDELBERG	DE	UKHD
CENTRE FOR EUROPEAN POLICY STUDIES	BE	CEPS
BCAM - BASQUE CENTER FOR APPLIED MATHEMATICS	ES	BCAM
U-HOPPER SRL	IT	UH
INTESA SANPAOLO SPA	IT	ISP
EIT DIGITAL	BE	EITD
SHARE FONDACIJA	RS	SHARE
A11-INITIATIVE FOR ECONOMIC AND SOCIAL RIGHTS	RS	A11
MINISTARSTVO ZA BRIGU O PORODICI I DEMOGRAFIJU	RS	MFWD
AZIENDA PROVINCIALE PER I SERVIZI SANITARI	IT	APSS
SWANSEA UNIVERSITY	UK	UoS
THE UNIVERSITY OF WARWICK	UK	UoW

LIST OF ACRONYMS

Acronym	Definition
DoA	Description of Action
EC	European Commission
HaDEA	European Health and Digital Executive Agency
WP	Work Package

Executive Summary

One of the strategic objectives of the TANGO project is to develop paradigms and algorithms for synergistic human-machine learning leveraging the theoretical framework of mutual understanding supported and enabled by the use of cognitive-aware explanations. In this deliverable, we define and discuss how different types of explanation processes can be integrated into interactive machine learning strategies to enable *synergistic* human-machine collaboration. To this end, we leverage the high-level desiderata of explanation processes, defined by WP1 in Deliverable D1.1, which suggests *interactive explanations* able to explain and justify the model decisions, without the need to open the black box and exploiting continuous interaction with human stakeholders. In particular, we propose to complete and implement this perspective through the Situation Awareness model for decision making integrated with interactive explanations. As a consequence, this document presents different types of explanation processes and how they correlate with the desiderata of D1.1 and the Situation Awareness model, and which kind of interaction strategies they enable. Finally, we discuss some ethical and legal issues that can rise from the introduction of explanations and from the interactive processes.

Contents

List of Figures

1 Introduction

The fields of eXplainable Artificial Intelligence (XAI) and Interactive Machine Learning (IML) have traditionally been explored separately. On one hand, XAI aims at making Artificial Intelligence (AI) and Machine Learning (ML) systems more understandable, by equipping them with algorithms to explain their own decisions [Bodria et al., 2023]. Such explanations are fundamental for enabling humans to inspect the system's knowledge and reasoning patterns. However in traditional ML/AI systems humans participate only as passive observers and have no control over the system or its behavior. On the other hand, IML focuses primarily on communication between machines and humans, and it is specifically concerned with eliciting and incorporating human feedback into the training process via intelligent user interfaces [Michael et al., 2020]. Unfortunately, despite covering a broad range of techniques for in-the-loop interaction between humans and machines, most research in IML does not explicitly consider explanations of ML models. Recently, few works focused on the integration of XAI within the IML loop. The core idea behind the need of interacting through explanations is that it represents a human-centric solution to the problem of acquiring an aware human feedback, and therefore leads to higher-quality AI and ML systems, in a manner that is effective and transparent for both users and machines[Teso et al., 2023]. By observing the machine's explanations, users have the opportunity to build a better understanding of the machine's overall logic, which not only has the potential to facilitate trust calibration [Amershi et al., 2014], but it also supports and amplifies our natural capacity of providing appropriate feedback [Kulesza et al., 2010]. Machine explanations are also key for identifying imperfections and incorrect reasoning of ML models enabling potential correction from human experts. At the same time, human interventions, after a reasonable understanding, are a very rich source of supervision for models and are also very natural for us to provide. In fact, explanations tap directly into our innate learning and teaching abilities [Aodha et al., 2018] and are often spontaneously provided by human annotators when given the chance [Stumpf et al., 2007].

Integrating explanations into interactive machine learning strategies helps make the human-machine collaboration more reliable, trustful and based on awareness of the situation where both humans and machines are operating. This integration (of explanations with interactions) is a fundamental building block of "synergy" between humans and AI systems. In fact, the role of explanations in this interaction is dual:

- Human-to-System Understanding: Users interact with explanations to build their understanding of the system's reasoning, decisions, and underlying models.
- System-to-Human Understanding: Through users' interaction patterns with explanations, the system gains valuable insights into users' comprehension levels, mental models, and decision-making processes.

This bidirectional understanding through explanation interaction aligns with the epistemic interaction paradigm Dix et al. [2024], which suggests that subtle modifications to interaction patterns can substantially increase the information available for system adaptation. In the context of XAI and IML integration, this means designing explanation interfaces not just to convey information, but also to gather rich feedback about users' understanding and reasoning processes through their interaction patterns.

However, as highlighted in Deliverable D1.1, explanation processes must meet some important

desiderata that are fundamental for making explanations effective, suitable for humans, adequate to contexts where they are operating and able to construct common ground for enabling the shared understanding of the main factors useful for making aware decisions. We can outline such desiderata as follows:

D1 – **interactivity**: AI systems should be able to verbally, visually or diagrammatically coherently explain, justify and defend their predictions or suggestions.

D2 – **contextuality**: Explainability requires local, context-specific responses built on mutual understanding between an AI system and the questioner.

D3 – **rationalization**: AI systems should be explainable in the same way as humans are, without the need for opening the black box.

In order to design and implement synergistic interactive machine learning models, we propose the use of the Situation Awareness model [Endsley, 1995] for decision making integrated with explanation processes which incorporate the above desiderata and enable the support of humans' understanding and awareness.

In this deliverable, we first describe the Situation Awareness model and discuss how different explanations can support this model (Section 2). Moreover, we also present and discuss the integration of explanations in different interactive learning protocols to enable a synergistic collaboration (Section 3). Finally, we discuss some ethical and legal implications of such synergistic learning strategies (Section 4).

2 Situation Awareness for Mutual Understanding

The implementation of explanation processes which satisfy the desiderata D1–interactivity, D2–contextuality and D3–rationalization is fundamental for establishing trust and common understanding between humans and AI-based decision systems, during both the model learning and decision making phases. In TANGO, we propose to use the situation awareness (SA) model, elaborated by Endsley in [Endsley, 1995], enriched with explanation processes (satisfying the above requirements) to support humans' understanding and awareness and to foster more effective interaction with AI-based decision-making systems. Our proposal enables the implementation of both synergistic machine learning and interactive decision making. Clearly, the main difference in the two contexts is the objective of the collaboration. In the first case, human-AI collaboration aims at improving the decision model resulting from the learning process; while in the second case, human-AI collaboration aims at improving the decision making process. In the synergistic machine learning, the understanding enforced by SA enriched with explanations empowers humans to effectively guide with awareness the model in learning how to decide at inference phase. In the interactive decision making, this understanding supports humans to make better and more aware decisions.

2.1 Situation Awareness

Situation awareness (SA) is a model supporting mutual understanding and user-centered design in decision making processes [Endsley, 1995]. The SA definition stems from naturalistic decision-making (NDM) [Lipshitz et al., 2001, Klein, 2008], a descriptive framework on how people decide in real-world situations. It has become an important component in command

and/or control decision support systems, where a designated commander has to make decisions and act to accomplish specific goals. Indeed, in these cases, a clear understanding of the current situation, along with an accurate projection of future states, is essential for effective decision-making. According to [Endsley, 1995], SA refers to "*the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future*", and focuses on the modeling of a user's environment to enhance their awareness of the current situation. More specifically, Endsley provides a three-level definition of SA where each level corresponds to a phase of the decision-making process:

- **Level 1 – Perception**: The perception of key information relevant to the decision maker's needs, which can include information gathered directly via a person's senses, or through electronic sensors, reports, communications with others, or other means. Thus, it includes information gathered from natural, engineered, and human sources.
- **Level 2 – Comprehension**: The comprehension or understanding of the significance of that information with regard to the decision makers' goals. SA involves knowing more than just data; it also includes being able to put together disparate pieces of data to inform relevant decisions, and interpreting what that data means.
- **Level 3 – Projection**: Projection of the current situation to inform likely or possible future situations. By constantly projecting ahead, decision-makers can act proactively instead of just reactively.

In alignment with the previous literature, SA can support the transparency of a system [Chen et al., 2018] while maintaining the user in control in a complex and dynamic environment. To better support transparency and human trust, Lee and See [2004] propose to present information to humans related to *purpose*, *process*, and *performance* (3Ps) of automation. The purpose represents the degree to which the automation is used according to the designer's intent. The *purpose* information describes *why automation was developed.* The *process* is defined as the degree to which a system appropriately solves a specified task, and the process information describes *how the automation operates*. The *performance* is defined as a measure of reliability of the system, or how the system has been efficient in the past, showing a history of previous interactions.

On these bases, Chen et al. [2018], propose a SA framework supporting efficient interaction through the communication of specific information in each of the above phases of the SA (*perception*, *comprehension*, and *projection*). This information process has the goal of facilitating the shared understanding required to perform effective human–agent teamwork. In the context of human-agent collaboration, the transparency of artificial agents plays a pivotal role in fostering three critical aspects of interaction: the mutual predictability of team members, the development of shared understanding, and the capacity to redirect and adapt to one another. A similar approach has been applied by Sanneman and Shah [2022] but with a focus on how explainability can be integrated with SA.

2.2 Explanation processes for supporting Situation Awareness in TANGO

Sanneman and Shah [2022] proposed SAFE-AI, a framework intended to provide a structured human-centered approach for defining requirements for explanation systems, providing

guidance for the development of explanation techniques and the definition of their informational requirements. In this deliverable, taking into consideration the interactive nature of TANGO decision-making systems, we describe how explainability processes in such setting, implementing the above desiderata **D1 – interactivity**, **D2 – contextuality** and **D3 – rationalization**, can effectively support SA upgrading human-AI interactions to "synergistic" human-AI interactions. Before describing the integration explanations in different types of interactive processes for machine learning, we discuss the requirements of each of the three SA levels for defining an appropriate human-AI collaboration.

In the first level, **Perception** of contextual elements, the human operator must not be overwhelmed by excessive information. It is crucial in this phase to establish shared goals between humans and AI to support their alignment and so, a first form of mutual understanding. To achieve this goal, the main questions to be answered are: *What's the AI trying to achieve in the current situation? Are humans' goals and AI aligned in a specific context?* To support SA effectively in this phase, humans need information that is useful for evaluating a first form of alignment with the AI systems. They need to understand the context and elements that characterize the situation where the AI suggestion/decision is acting. Hence, the explanation process should provide information about the input and output of the AI system, potentially with its level of uncertainty, for highlighting critical information for the context in which the decision occurs. The presented information should have the objective of enabling humans to recognize already known elements. If we consider the 3Ps information process, presented by Lee and See [2004], this means that humans have to be informed about the *purpose* (task addressed by the AI), a simple form, and or an intuition about the *process* for making the decision (input, output of the models and importance of some factors/features) and the *performance* (confidence and or uncertainty of the AI model). The presented information should support the humans' judgment of the current situation. Given the requirement to limit the human workload and that the role of the explanation is to enable and potentially confirm the humans' perception and their judgment, *in this phase simple and single-step explanation is sufficient*. Consequently, score-based explanations can be presented to humans e.g., feature importance explanations and saliency map-based explanations for images and text [Bodria et al., 2023].

In the second level, where the **Comprehension** of the situation is necessary, a shared reasoning between humans and AI systems is essential for an effective collaboration. In other words, humans should understand the AI's rationale behind the suggestion and update their mental models with the knowledge provided by the AI to improve their understanding of the situation. The question to be answered to support this phase is: *Why does the AI system suggest that advice?* In this phase, we consider that an important element of the explanation process is its **interactivity**, since it should help in constructing a shared common ground to enhance the *mutual understanding* in such a way that humans improve their ability to make informed decisions in a more appropriate way. We highlight that in case of an interactive process for machine learning the human awareness in "making a decision" has the impact of improving the learning process because the human operator could exploit the acquired knowledge for deciding whether updating accordingly the machine leaning model. In case the explanation is used for justifying a model decision at inference phase (explainable interactive decision making) the human awareness will impact directly the final decision making process. Referring again to the 3Ps information process supporting the comprehension phase means providing humans with information about the *process*, i.e., how the AI system reaches the

specific decision/suggestion. Since the goal is to construct a common knowledge and the explanation process should also be able to adapt to human needs, preferences, and skills, *a more complex explanation process is necessary in this phase.* In particular, this is necessary for an interactive explanation process that provides explanations by a **progressive disclosure approach**, which step-by-step provides explanations for incrementing progressively the human comprehension [Springer and Whittaker, 2020]. The need for deeper comprehension can be derived from different factors: humans are not aligned with the AI suggestion/decision or do not completely trust the AI suggestion. In these cases, explanations provided during the interactive process can also be more elaborate than a simple feature importance. For example, explanation processes exploiting logic-based explanations (factual and counterfactual rules), example-based explanations (examples, counter-examples), concept-based explanations, or context-based explanations might be more appropriate to trigger a deliberate reasoning. However, considering the desiderata D1-D3 highlighted above, for facilitating mutual understanding the process should prioritize local explanations based on surrogate models facilitating the **rationalization** and **contextuality** (Section 3.1), concept-based explanations providing explanations expressed with more high-level details enabling **contextuality** (Section 3.2) and textual-based explanations for enhancing expressiveness and comprehensibility of explanations and for facilitating a conversational approach and so, **interactivity** (Section 3.3).

In the third level, the **Projection** of possible outcomes according to different actions, the human operator should engage in counterfactual thinking to envision potential actions and the following outcomes. AI can support counterfactual thinking by engaging in a negotiation process based on different decisions, and the human operator can evaluate which alternative is the best. The question that guides this phase is: *what could happen if humans follow AI?* The understanding of how AI reached the proposed decision/suggestion is paramount for humans to project and evaluate the potential outcomes and make more informed decisions through a thoughtful weighting process of the suggestions, thus improving users' situation awareness. To support this phase, the explanation process should support user prediction of AI behavior by enabling an understanding of what the system would do if its inputs changed or if the model were to change in any way. In other words, explanations able to answer "what if" questions about system behavior are adequate for this phase. Any kind of explanation process based on counterfactuals could enable this type of support [Guidotti, 2024].

We highlight that in case an AI decision model is already trained and used for supporting a user in making decisions, potentially for any new instance for which the user has to make a prediction the system will present the explanation adequate to support the corresponding SA phase. The availability of an explanation (more or less deep and exhaustive) for each instance makes the user able to make more *aware* decisions. This type of processes will be the focus of WP3 and Deliverable D3.1.
In the learning phase, we claim that (interactive) explanations cannot be presented for each instance, involved in the learning process, because this could make the learning process unfeasible. In particular, it could quickly exceed any available time and cognitive load budgets. In this setting, interactive explanation processes, adopting the progressive disclosure approach, should be enabled only for those instances for which the interactive learning protocol identifies some critical factors, which require human interventions with their expertise. In this situation humans need a deep comprehension of the different factors of the situation for understanding whether updating either their mental model or the AI model. As a

consequence, in synergistic machine learning an explanation process typically supports the comprehension and projection phase of SA.

3 Explanations for Synergistic Machine Learning

In this section we discuss three main categories of explanation processes that can be used for supporting the SA phases while implementing the desiderata D1-D3 identified in WP1. In particular, we describe feature-driven explanations, concept-based explanations and natural language explanations and their integration into an interactive machine learning protocol for enabling a synergistic human-AI collaboration. Our discuss highlights how each category of explanation presents characteristics that enable one or more desiderata.

3.1 Feature-driven explanations for enabling synergistic learning

Surrogate-based explainers. Some approaches for deriving explanations of black box models apply a process called surrogacy. The core idea behind this technique is to obtain a decision model capable to replicate the black-box model predictions as best as possible, and at the same time, provide interpretability. Clearly, the challenge in this context is to derive an interpretable model with similar accuracy (similar predictions) with respect to the complex black-box model. Indeed, complex models such as a deep learning model are more suitable to learn the complex nature of the decision domain in presence of complex data than simple models. In general, the training of an interpretable surrogate model requires a model-agnostic method, since it does not need any information about the inner workings of the black box model, only access to data and the prediction function is enough. As a consequence the choice of the black box model type and of the surrogate model type is decoupled. Explanations derived by surrogate models meet the desiderata **D3 – rationalization**, i.e., this type of explanations are making AI models explainable without opening the knowledge of the internal working of the AI model but deriving the reasons of the decisions by observing only input data and output. An important dimension which can be considered for distinguish different explainers is their ability in providing *local* or *global* explanations. Thus, in the literature, some surrogate-based explainers use **global** surrogate models while others exploit *local* ones.
Global surrogate models aim to explain *all the predictions* made by the black-box model. Local surrogate models, on the other hand, aim to explain single instance predictions made by the black-box model. We highlight that in general, explaining globally a complex decision boundary with a simple model (e.g., a linear model or a rule-based model), which tends to oversimplify a complex reality, might be very challenging and sometimes impossible. Local surrogate models addressed the complexity of the decision boundary focusing their work on a small portion of the complex decision boundary, i.e., reasoning locally and around the instance to be explained. We highlight that the *locality of an explanation* is a property that makes an explanation process compatible with the desiderata **D2 – contextuality**. In the context of post-hoc explanations we can identify three types of explanations that can be derived by using a surrogate model: *rule-based* explanations, *example-based* explanations and *score-based* explanations. These types of explanations make an AI decision model explainable by using the same set of features received in input by the decision model, so they are also called feature-driven explanations.

Rule-based explanations. A rule-based explanation is a predicate defining a region in the feature space that is sufficient for classifying a given point. In other words, a rule is defined in terms of the features of the Ai decision model that both cover the given point and is sufficient for its classification. Rule-based explanations are perturbation resistant in the sense that the explanation applies to a neighborhood about a given point. This type of explanation is usually extracted from surrogate models, such as decision trees or rule-based classifiers, that mimic the black-box model in the vicinity of the record being explained [Guidotti et al., 2024]. Such surrogate models often enable the extraction of either *factual rules*, which represent the path satisfied by the analyzed record and describe the reasons of the model classification and *counterfactual rules*, which outline paths that lead to the opposite prediction. This kind of explanation can be employed as local or global rules, depending on the surrogate model generated. In the context of local surrogate based models in the literature two explanations approach exist LORE [Guidotti et al., 2024] and Anchor [Ribeiro et al., 2018], while for global surrogate models TREPAN [Craven and Shavlik, 1995] is one of the most popular.

Example-based explanations. Example-based explanation methods select particular instances (real or synthetic) of the dataset to explain the behavior of AI decision models or to explain the underlying data distribution. Example-based explanations are mostly model-agnostic, because they do not use the knowledge of the model internals for generating such examples. We highlight that these methods do not create summaries of features (such as score-based explanations and rule-based approaches) but present as explanations an instance with the complete set of features. Example-based explanations only make sense if we can represent an instance of the data in a humanly understandable way. In general, example-based methods work well if the feature values of an instance carry more context, meaning the data has a structure, like images or texts do, or in case of the tabular data which do not have an high number of features or if we are able to find a way to summarize an instance. Example-based explanations help humans construct mental models of the machine learning model and the data the machine learning model has been trained on. It especially helps to understand complex data distributions. Example-based explanations help in simulating an approach typically humans adopt: *reasoning by examples or analogies*. Suppose, a physician sees a patient with trouble in breathing and staying awake and a high fever. The patient's symptoms remind her of another patient she had some time ago with similar symptoms. Thus, she can perform some tests to verify the presence of the same disease. Implicitly, some machine learning approaches, such as Decision Trees and k-nearest neighbors methods, apply an example-based reasoning. Some example-based explainers provide *examples* that are similar to the instance to be explained receiving the same AI model decision while others provide *counterexamples* (counterfactual explanations), i.e., instances similar to the one to be explained but with a different AI model decision.
The objective of counterfactual explanations is to understand the difference between the instance under analysis and those in the opposite class. There are some important considerations on counterfactuals, such as *feasibility* of counterfactuals, e.g., the changes suggested should be feasible and realistic; their *diversity*, e.g., in a set of counterfactuals there should be different changes in the variables even with the same outcome class, e.g., a counterfactual may require the changing of the "age" variable, another instead focuses on the changes on the "income" attribute; similarity in terms of feature values with respect to the instance but also should also change as few features as possible. Examples of methods returning counterfactual explanations are DICE, CEM and LORE [Mothilal et al., 2020, Dhurandhar

et al., 2018, Guidotti et al., 2024].
Other forms of example-based explanations are *prototypes*, i.e., a selection of representative instances from the data, and criticisms, i.e., instances that are not well represented by those prototypes Gurumoorthy et al. [2019], Kim et al. [2016].

Score-based explanations. Score-based explanations provide a numerical relevance value to each feature in the input record. In this setting we can have surrogate-based explainers, such as LIME [Ribeiro et al., 2016], or gradient based methods, such as SHAP [Lundberg and Lee, 2017] and DeepLIFT [Shrikumar et al., 2017]. These explanations are usually local, but there are also examples of global models.

3.1.1 Can feature-driven explanation upgrade interaction protocols to enable understanding?

Any of the above types of explanations can be integrated in an interactive machine learning protocol for providing humans with knowledge useful for understanding and eventually defining interventions for collaborating in the learning process.

In a standard **active learning loop** whenever the machine queries the label of an instance, it also presents a prediction for that instance and potentially a local explanation for the prediction. Explanations can be extracted using feature-driven methods. At this point, the annotator supplies a label for the instance and, optionally, corrective feedback on the explanation. This means that the human annotator can, for instance, indicate what input variables the machine is wrongly relying on for making its prediction. Another possibility is to use counterfactual explanations highlighting why for some classes the model cannot discriminate well, and for each of them asks an annotator to provide feedback on what features enable them to distinguish between the two classes and then using the feedback for helping the decision model accordingly. A particular setting in active learning is the **interactive process of skeptical learning** designed for incremental interactive learning from potentially unreliable humans, where in each iteration the machine receives a sample and updates the model accordingly. For each sample the machine compares a quality score of the annotation with that of its own prediction, and in case the prediction results more reliable than the annotation by some factor, the machine asks the humans to double-check the labeling of sample. Feature-driven explanation process could help in justifying i.e., to explaining to the human annotator why the machine believes the outcome of the decision should be different. The understanding derived from the explainable mechanism could lead the annotator to change her mind or to override the machine's suggestion and force it to accept her decision. In return, the human annotator can signal to the machine that its explanation is correct or not by helping a better training of machine learning model.
Explanations used for helping humans in understanding the cases of skepticality can be of different types. Since the human annotator needs to reason on the skepticality of the machine on a specific instance prediction/labeling, in this case *local explanations* are considered more suitable than the global ones; this allows also to make the explainable interactive process compliant with **D2 – contextuality** desiderata. Moreover, to implement the **D3 – rationalization**, *local surrogate explainers*, enabling feature importance, example-based explanations (factual and counterfactual) and rule-based explanations, should be preferred. However, following the reasoning presented in Section 2.2, in case of skepticality humans clearly need to achieve a

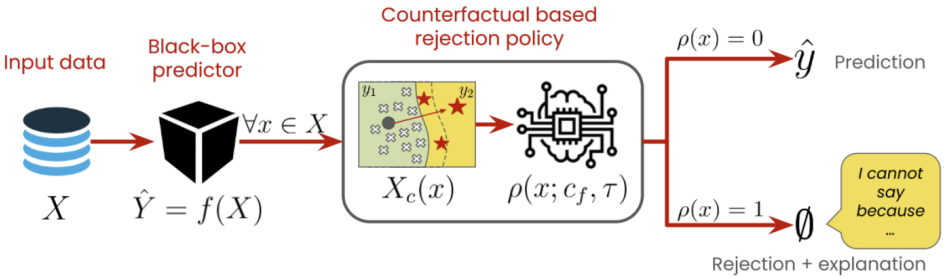


Figure 1: Explanations for learning to abstain.

deep understanding for identifying the correct decision (Comprehension phase of SA model) and eventually to update the model, thus rule-based explanations and counterfactual explanations can enable a more comprehensive awareness of the specific situation. Moreover, in this phase humans, needing to acquire awareness, can interact with the system asking with a progressive disclosure approach more and more sophisticated explanations for triggering a deeper reasoning. For example, the human can start with a feature importance and combine it with examples and counterfactuals. In TANGO project, feature-driven explanations have been integrated in an interactive process for skeptical learning, called FRANK, which also includes checks of fairness guarantees and consistency over the time during the incremental learning [Mazzoni et al., 2024].

Feature-driven explanations can be also incorporated in "**Learning to Abstain**"(L2A), which are designed for enhancing system reliability. These methods are based on the idea that when the AI system has low confidence in its predictions or when errors could have significant consequences, it may be more appropriate to abstain from making a prediction and instead pass the input to a more advanced system or a human being. Several methods have adopted this mechanism [Zerilli et al., 2022]. However, there has been limited focus on a critical limitation of L2A: the lack of transparency in the rejection policy, i.e., in the reasons why certain instances are rejected by the system, which can ultimately erode human trust and satisfaction with the automated system. In this context, model-agnostic interpretable abstention policies based on *local* counterfactual explanations, extracted by exploiting a local surrogate model, might enable humans to understand the reasoning behind the abstentions of the classifiers (Fig. 1). As above, locality of the explanation is fundamental because humans need to understand the rejection reason for each specific instance and this enables also the Desiderata **D2 – contextuality**. Moreover, the use of surrogate model can also enable to fulfill the **D3 – rationalization** requirement.

3.2 Concept-based Explanations for Synergistic Machine Learning

Whereas the explanations discussed in Section 3.1 are defined in terms of individual attributes or training examples, human communication most often relies on higher-level constructs, such as words and symbols [Kambhampati et al., 2022, Rudin, 2019]. This simple observation has lead to the development of *concept-based explanations*. While the notion of “concept” is notoriously difficult to pin down, in Explainable AI concepts are (usually concrete) properties useful for describing the decision-making process of a model at a higher – and more human-friendly – level of detail. In an image classification task, for instance, concepts would describe

what *objects* are visible in the image as well as their *attributes*. A conceptual description of an image (a set of pixel values) would be that it pictures a "`red ball`" and a "`fluffy dog`": here, "`red`", "`fluffy`", "`ball`" and "`dog`" are concepts, captured by binary variables, that summarize the contents of the image without referencing individual pixels. AIs able to express themselves using concepts would enable simpler and richer communication with human decision makers and stakeholders while avoiding possible ambiguities.

Concepts hold much promise for synergy. To see this, consider the prediction task illustrated in Fig. 2 (taken from [Stammer et al., 2021]). Here, a predictor has to recognize handwritten digits, with different digits colored using distinct colors. In practice, machine learning predictors quickly learn to rely on color information to predict the digits. This is an instance of *spurious correlation* [Geirhos et al., 2020, Lapuschkin et al., 2019] and it is highly undesirable: affected models are unable to accurately predict digits that have been recolored, desaturated, etc. Low-level explanations, such as saliency maps, capture what pixels the model's predictions depend on, but *cannot disambiguate between positional and color information*. This makes it impossible for human observers to understand that the machine relies on color for its predictions, completely hiding the issue. In stark contrast, explanations that justify the model's predictions in terms of high-level concepts including "`color`" and "`shape`" avoid this ambiguity, as humans could readily and unambiguously verify that the model is misusing color information.

Concepts in Explainable AI. Producing concept-based explanations like the above requires understanding what concepts the model has learned and how they affect the model's predictions. There are two popular strategies for tackling these problems:

- *Extracting concepts in a post-hoc fashion*. In other words, given access to the latent representations (aka embeddings) computed by a neural network, these techniques attempt to figure out what concepts are encoded into the latent representation and how these ultimately impact the model's decisions [Kim et al., 2018, Schwalbe, 2022].
- *Learning concepts in an ante-hoc fashion*. This amounts to designing machine learning models in which concepts are first-order citizens: the model first describes the input in terms of high-level concepts, and then uses these to infer a prediction. In this sense, the concepts are explicitly modelled in the predictor's architecture; they act as a conduit

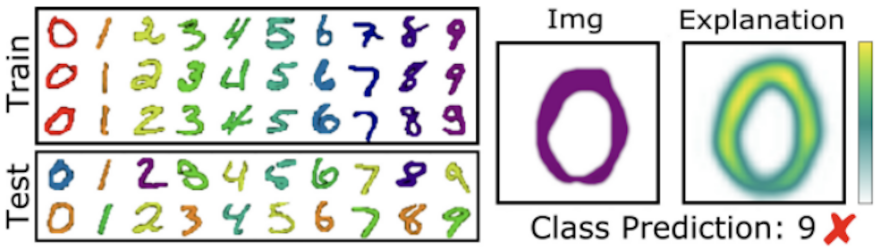
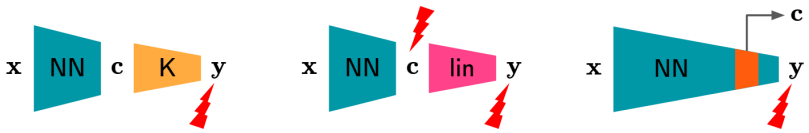


Figure 2: Visual Explanations for ColorMNIST: (Left) the general data distribution between train and test split; (Right) a typical visual explanation of a CNN. Notice digit pixels are considered as important for the wrong prediction.



(a) NeSy (DPL & LTN) (b) CBM (c) NN & CLIP

Figure 3: Schematic comparison between concept-based models, neuro-symbolic models, and regular neural networks paired with a post-hoc concept-based explainer.

between the input and the output, and their value can be accessed directly, without requiring any sort of post-processing.

One issue with post-hoc approaches is that they provide no guarantee that the concepts they extract perfectly match those that are learned and used by the model itself. This *faithfulness gap* can mislead users into thinking that the model relies on concepts that it is in fact not using. While this issue can be viewed as a natural consequence of **D3 – rationalization**, and as such it is not a show-stopper for synergy, it can also be side-stepped by working with concept-based models instead. This is why in the following we chiefly focus on ante-hoc approaches.

Two families of concept-based models. Ante-hoc concept-based models can be grouped into two families: *concept-based models* "proper" (CBMs) and *Neuro-Symbolic models* (NeSy). They follow the same architectural blueprint, as shown in Fig. 3. There, we compare a NeSy model (left) to a CBM (middle) and a post-hoc concept extraction procedure (right). (The lightning bolts indicate what variables receive supervision during training.)

CBMs consist of two computational blocks: a neural network (NN in the figure) extracts concepts from the low-level input and an interpretable model (e.g., a sparse linear layer; "lin" in the figure) infers a prediction from the concepts alone [Poeta et al., 2023]. Many CBM architectures exist, including Self-explainable Neural Networks [Alvarez-Melis and Jaakkola, 2018], Part-Prototype Networks [Chen et al., 2019] Concept-Bottleneck Models [Koh et al., 2020], Concept Embedding Models [Zarlenga et al., 2022], and GlanceNets [Marconato et al., 2022]. They chiefly differ in how they implement the concepts and in what kind of supervision they require. For instance, Part-Prototype Networks implement concepts as prototypes in embedding space and are designed for unsupervised concept discovery, while Concept-Bottleneck Models predict concepts via a feed-forward neural backbone trained using concept-level supervision.

NeSy models share the same two-step architecture, but differ in how they infer predictions. Specifically, these are computed by applying *logical reasoning* to the extracted concepts (K in the figure). This is critical for enabling high-stakes applications in which predictions have to comply with existing regulations, such as structural or safety constraints. For instance, consider a model responsible for determining whether an autonomous vehicle should "`stop`" or "`go`" based on what obstacles it sees on the road: the prior knowledge might state that whenever it can see a "`pedestrian`" or a "`red light`", the car should "`stop`". While regular neural networks can in principle learn such safety constraints from data, there are no guarantees that they will do so nor that they will output predictions that comply with the learned

rules, hindering trustworthiness. Instead, one can natively steer the predictions of NeSy models by encoding any such rules into a logic format: the reasoning step will automatically output predictions consistent with these constraints, either with guarantees or in expectation, depending on the model. Well known NeSy architectures include DeepProbLog [Manhaeve et al., 2018], Logic Tensor Networks [Badreddine et al., 2022], the Semantic Loss [Xu et al., 2018], and Semantic Probabilistic Layers [Ahmed et al., 2022]. A key difference between them is how they implement reasoning: in order to ensure learning is end-to-end differentiable, they often "relax" the reasoning step, for instance by imbuing the prior knowledge with fuzzy [Badreddine et al., 2022] or probabilistic [Xu et al., 2018, Ahmed et al., 2022] semantics. Crucially, in NeSy the concepts are most often treated as latent variables, in that they receive no concept-level supervision during training.

What is the role of concepts in synergistic machine learning? A first answer is that concepts can help *understand what the model is truly thinking*. Indeed, in the Color MNIST example above, concept-based explanations can inform users that the predictor is mistakenly relying on "`color`" rather than "`shape`", and therefore that it will likely misbehave in real world applications [Stammer et al., 2021].

Another important remark is that concepts are a natural medium for implementing **progressive disclosure**, in the sense that a single concept (say, "`car`") can be often broken down into constituents (for instance, "`body`", "`wheel`", "`windshield`", etc.) or detailed further ("`red car`", "`sporty car`"), where all resulting elements can be viewed as concepts themselves. In principle, this decomposition can be made interactable: after receiving a concept-based explanation, stakeholders could ask the model to further break down the responsibility of a concept into its constituents, thus supporting incremental, human-centered, and personalized understanding of the model's decision process.

A second answer is that concepts can be used to **upgrade interaction protocols** (as per **D1 – interactivity**), facilitating reactivity and control. To gain an intuition about what benefits concept-level explanations can bring, consider the following setup, adapted from [Bontempelli et al., 2023]. Imagine having to debug a predictor trained to classify birds from images. This task is naturally confounded. For instance, albatrosses are often found in sea regions, and sparrows in forest regions. This can drive the model to use background information (sea vs. forest) to discriminate between different species. In turn, this compromises prediction performance whenever this spurious correlation breaks (e.g., for pictures of birds in zoos). A popular solution is to use explanatory interactive learning [Schramowski et al., 2020]: in this setting, an expert user looks at model's saliency maps – which will show that the predictor is wrongly focusing on the background – and provides corrective feedback. This process is however *expensive*, as the user has to indicate all those pixels that the model should not be looking at. Moreover, such corrections do *not* generalize across examples, as the pixels that correspond to the background are likely to change across images. The situation is much better for concept-based models: in this case, the user can simply instruct the model nor to rely on *concepts* that are irrelevant for the label (in our example, "`sea`" and "`forest`"). This corrective feedback can be obtained with a single mouse click and applies seamlessly to all input images, irrespective of their specific layout [Stammer et al., 2021, Bontempelli et al., 2023, Teso et al., 2023].

A nice feature of concept-based explanations is that they are a *drop-in replacement* for

saliency maps, meaning that interaction protocols that rely on the latter can be readily augmented with the former. These include many of those used for establishing synergy within TANGO, such as skeptical learning (Section 3.1.1) and learning to abstain (Section 3.1.1). At the same time, concepts are no less *contextual* than attribute-level explanations (**D2 – contextuality**).

Another interaction pattern that benefits from concept-level feedback are *interventions*. Whenever a CBM or NeSy model mis-identifies the concepts present in the input, a sufficiently expert user can edit the predicted concept values. The model will then automatically propagate these changes to the final prediction [Koh et al., 2020].

Finally, thanks to their progressive disclosure properties, concepts can be employed to implement **argumentation**, which is precisely the kind of justification mechanism vouched for by **D1 – interactivity**. For instance, a user could ask for concept-based explanations to unveil what concepts caused a model to output a bad prediction for the purpose of understanding and potentially correcting them. Rather than stopping here, the process can continue: assuming that some concepts have been predicted poorly, the human could ask the model to explain those concepts in terms of more fine-grained concepts, and so forth. This would enable and prompt humans to supply corrective feedback *precisely for those finer-grained concepts that are responsible for the model's mistakes while allowing them to interact at a level of detail they feel comfortable with*.

Synergy and concept quality. In order for all the above to work, however, the machine's concepts have to be at least partially *aligned* with those of human stakeholders [Russell, 2019]. This is not a given. Recall that CBM and NeSy architectures learn concepts from limited amounts of data, which is often noisy or low-quality, and in many cases without direct supervision on the concepts themselves. This can introduce a gap, in the sense that humans and the model might not attach the same meanings to the same concepts. Naturally, this can thwart communication – with both ante-hoc and post-hoc approaches – in both directions.

Ensuring that machine concepts have human-interpretable semantics is not a trivial problem, and in practice machine concepts can be unaligned. For instance, NeSy models can suffer from *reasoning shortcuts*, whereby they learn to output predictions consistent with the prior knowledge by learning concepts with unintended semantics. In the context of autonomous driving, a NeSy model might be designed to predict "`stop`" whenever "`pedestrians`" or "`red lights`" are visible on the road. It turns out that, since both concepts yield the same correct outcome (that is, "`stop`"), NeSy models can achieve high performance on the training set by confusing pedestrians with red lights [Marconato et al., 2023a, 2024b].

A first line of defense against these issues is to provide models with annotations for the concepts themselves, to be used during training. This strategy is however expensive, as most data in the wild does not come with concept annotations and these have to be acquired separately. A possible solution is then to *leverage Large Vision-Language Models* (VLMs), such as Llava [Liu et al., 2024], to automatically generate concept-level annotations, either directly (by querying about concepts present in an input image) or indirectly (by extracting concepts activations from their latent representations by comparing the latter to neural encodings of textual descriptions of the concepts themselves). While in practice this strategy can be very effective [Yang et al., 2023, Oikarinen et al., 2022], it is still unclear whether it

is a perfect solution, in the sense that there is no guarantee that the annotations provided by VLMs are always high quality or aligned with the stakeholders' own concept vocabulary. At the same time, even concepts learned by CBMs using complete and clean concept-level supervision can still suffer from *concept leakage*, whereby some concepts may unintentionally encode information about unrelated concepts [Margeloiu et al., 2021, Mahinpei et al., 2021, Marconato et al., 2023b].

One further issue is that the human meaning is to some extent *subjective*: different stakeholders have different backgrounds and expertise, hence they will possibly associate different meanings to the same concept. Perhaps surprisingly, subjectivity is not a deal breaker, neither from an algorithmic perspective nor from a theoretical one. In fact, It turns it can be formalized mathematically using notions from *causality* [Pearl, 2009, Marconato et al., 2023b], and so can concepts themselves. Roughly speaking, we can say that concept alignment holds if it is the case that whenever provided with a certain input of interest (e.g., an image) a human observer and a machine provide the same conceptual description thereof (modulo simple transformations that do not impact interpretation). If this holds, then we know for a fact that the machine and the human associate the same meaning to the same concepts. This perspective also allows provides a recipe for handling concept leakage [Marconato et al., 2023b].

The question which needs to be addressed is how to ensure – algorithmically – that machine concepts are not misaligned or, if this turns out to be too difficult a target, to at least encourage machines to be **uncertain about low-quality concepts**, so as to lessen stakeholders' trust in them [Marconato et al., 2024a]. A clear issue with the above definition of alignment is that it is "global": the human and the machine have to agree about *all possible inputs*. This condition is neither easily satisfied nor easily checked. One possible route forward is then to instead ensure that alignment holds "locally", that is, that the human and machine agree at least on those concepts that are strictly necessary for communicating about the inputs actually occurring in a specific learning or decision task. This could enormously reduce the computational and labeling requirements for supporting concept-based communication - and thus synergy. In turn, this suggests that alignment might be built incrementally via interaction: as new inputs have to be processed, a system should ensure that the concepts necessary to handle those inputs - and only those - are indeed aligned with the human's. This setup would be compatible with both **D1 – interactivity** and **D2 – contextuality**, and in turn gradually converge toward mutual understanding, at least for what concerns the concepts in use.

TANGO is at the forefront of these efforts, and indeed has been investing in *designing challenging benchmarks* for assessing the quality of learned concepts, also in changing environments [Busch et al., 2024, Bortolotti et al., 2024].

Concepts and synergy. There are several important links between concept-based explanations and desiderata **D1**–**D3**. A first important link is that to **D3 – rationalization**: establishing synergy does not necessarily imply having to "open the black box"; rather, as with humans, high-level justifications of machine behavior may be enough. Concepts do exactly this: they reinterpret the machine's inner logic at a higher level of detail. In practice, CBMs and NeSy architectures make it straightforward to figure out what concepts lead to a certain prediction, without disclosing how those concepts were extracted.

Next, by disambiguating the model's reasoning to users – as illustrated by the Color MNIST

example – concept-based explanations can help models make their justifications, which are key for **D1 – interactivity**, less ambiguous. This also facilitates mutual understanding, which lies at the heart of **D2 – contextuality**.

Naturally, the requirement that concepts are aligned across agents – that is, that the same concept vocabulary is shared by all parties involved - supports **D1 – interactivity** when it holds. At the same time, interaction itself is a natural approach for establishing this alignment, in the sense that users and machine can interact – for instance, by exchanging concept-level supervision [Stammer et al., 2021, Bontempelli et al., 2023] or weak supervision [Stammer et al., 2022] – so as to progressively align the semantics of the concepts they use for communication [Marconato et al., 2023b].

3.3 Natural language explanations for enabling synergistic learning

Another excellent medium for expressing explanations is *natural language*. Natural language techniques, Large Language Models (LLMs) and Vision-Language Models (VLMs) can enhance explanations by providing text-based descriptions of decisions and semantically enriching the explanations. In case the goal is to explain in natural language NLP models, explanations are often cast in the form of *rationales* [Camburu et al., 2018]: these are (minimal) subsets of an input sentence that are sufficient for the model to output a certain prediction. For instance, in sentiment classification, a movie review could be classified as “negative” due to a particularly ironic sentence, which plays the role of rationale for the prediction.

Foundation models as an interface for synergy. However, LLMs and VLMs makes it progressively easier to generate meaningful natural language descriptions, opening the door to applications in explainable and interactive machine learning. As a matter of fact, LLMs and VLMs can be used as an interface between humans and machine learning models, and they can be integrated into an interactive dialogue system designed to explain ML models through conversations [Slack et al., 2023, Alonso et al., 2020].
Natural language explanations can be generated using an explanation generator layer providing text-based explanations of the black-box models decisions. An approach is to transform feature-driven explanations, presented in Section 3.1, into natural language explanations. For example, rule-based explanations can be converted into text-based explanations by defining *templates* that transform key meta-data from the rule premises into a textual format. The completed template is then provided as a prompt to a LLM, which uses the instructions to generate a natural language explanation of the decision rule. This approach can be also used for explaining local tree-based surrogate models [Alonso and Bugarín, 2019]. However, also score-based explanations can be explained in natural language providing additional knowledge about the meaning of features [Burton et al., 2023].

It is easy to see that this medium can bring similar benefits as using concept-level explanations, in that both families of explanations allow to unpack the model’s reasoning into high-level, human-interpretable terms, with benefits for both understanding and, eventually, control.

Considering the different interactive protocols used for establishing synergy within TANGO, such as skeptical learning, learning to abstain, and explanatory interactive learning, it is worth to note that text-based explanation processes obtained by the use of LLMs can help in improving both the interactive and explanation processes facilitating the implementation of

Desiderata **D1 – interactivity**, in addition to **D2 – contextuality** and **D3 – rationalization** already satisfied by the use of local and surrogate based explainers. A recent innovative approach, called Retrieval-Augmented Generation (RAG), combines the generative capabilities of LLMs with a retrieval system to ground responses in relevant external data, such as proprietary or domain-specific information. It is based on the context injection, i.e., the retrieved information is fed into the LLM as context, guiding it to generate more accurate, informed and more relevant responses for the application domain. The use of RAG models for enhancing expressiveness of the explanation can considerably improve the **D2 – contextuality** desiderata. It can really help in supporting the SA phases (especially the Comprehension phase) by enabling both **interactivity** but also responses, and thus explanations, which are **context-dependent** and which can facilitate humans in understanding the situation and the factors determining the model decision.
Moreover, in the context of synergistic decision making, one could conceive of ML models supplying useful guidance to human decision makers [Banerjee et al., 2023, 2024]. This is useful for upgrading learning to defer strategies to high-stakes applications. In short, learning to defer builds on a *separation of responsibilities* setup, whereby individual predictions are either taken by the machine in full autonomy or offloaded to a human expert. One issue with this setup is that, on the one hand, the expert may end up over-relying on the machine's decisions due to *anchoring bias*, thus losing the human oversight that is increasingly being required by regulatory agencies to ensure trustworthy AI. On the other hand, the expert is left entirely unassisted on the (typically hardest) decisions on which the model abstained. One option is then to build models capable of generating guidance capable of supporting human decision making. In this setup, the human is always in the loop and always responsible for the decisions they make. Natural language then is the ideal format for encoding guidance, for instance in medical diagnosis tasks in which information comes from multiple heterogeneous textual or tabular source (e.g., electronic health records).

How to deal with hallucinations? The downside of relying on foundation models is that of potentially foregoing faithfulness – meaning that explanations might not fully reflect the reasoning process of the model to be explained – for instance due to *hallucinations*. These can be classified into two types: factuality and faithfulness hallucinations. Factuality hallucinations occur when an LLM's response contradicts real-world facts, while faithfulness hallucinations arise when an LLM's response does not align with user instructions or the given context [Huang et al., 2023]. To mitigate factuality hallucinations, RAG [Lewis et al., 2020] can be employed, enabling LLMs to access external tools (e.g., a web search engine or Python interpreter) or resources (e.g., Wikipedia or Pubmed) to gather necessary information and ensure accurate responses. However, LLMs may still rely on ingrained knowledge, occasionally ignoring retrieved information [Xu et al., 2024a, Jin et al., 2024]. An alternative approach involves using the self-correction capabilities of LLMs to post-process and refine answers, enhancing their originality [Ji et al., 2023, Dhuliawala et al., 2024], or even integrating ideas from neuro-symbolic AI to encourage LLMs to satisfy logical constraints and prior knowledge [Calanzone et al., 2024]. Addressing faithfulness hallucinations requires improving the alignment of LLM responses with user instructions and context [Tian et al., 2020, van der Poel et al., 2022, Wan et al., 2023, Shi et al., 2024, Gema et al., 2024]. This might involve ensuring logical coherence in multi-step reasoning tasks [Wang et al., 2023a], or creating a reward model that adaptively promotes hallucination-free generations [Amini et al., 2024, Lu et al., 2023, 2022,

Deng and Raffel, 2023]. Despite these advancements, completely eliminating hallucinations remains a challenge [Xu et al., 2024b]. Consequently, it is essential to also enhance human awareness of these limitations. Strategies could include enabling humans to manage LLM outputs by accessing model logits to estimate uncertainty [Guo et al., 2017, Kuhn et al., 2023], prompting LLMs to verbally express their level of uncertainty [Mielke et al., 2022, Lin et al., 2022, Xiong et al., 2024, Kadavath et al., 2022], or by asking the same question multiple times to select the most consistent response [Lin et al., 2023, Chen and Mueller, 2024, Wang et al., 2023b, Zhao et al., 2023a].

4 Ethical Considerations

In this section, we discuss some potential ethical and legal risks that it is important to assess and mitigate in TANGO explainable AI systems. Making AI-based decision systems explainable requires to take into account potential risks that can derive from both the interpretability layer that we are adding to the system and the human interaction that can lead to human interventions during the decision model learning. In particular, we will discuss potential risks of: *(i) unfairness*, which could lead to the amplification of gender and social bias; *(ii)* privacy violations, which could lead to jeopardize important human rights; and *(iii)* unacceptable *human persuasion*, which could lead to undermine the active role of humans in the decision making process.

Fairness The impressive capabilities of AI models (both predictive and generative) are coupled with substantial concerns about their fairness [Mehrabi et al., 2022, Teo et al., 2023] and sometimes these issues are also inherited by explanation models [Balagopalan et al., 2022]. These issues mainly derive from the training of the model on vast amounts of data, obtained from open sources on the web and often uncurated and incomplete, which may include content misaligned with current societal values, including discriminatory language based on gender, ethnicity, sexual orientation, age and others, often manifesting as stereotypes or prejudices, and harmful or hateful content. Recent research has demonstrated that generative language models and text-to-image generative architectures frequently produce toxic and biased content, raising ethical issues for their application in real-world scenarios. Considering the interaction learning strategies studied within TANGO which focus in the strong involvement of humans in the learning phase it is of fundamental importance to (1) avoid humans reinforcing potential biases; (2) exploit this interactivity during the learning phase as a new driver for discovering potential bias and mitigating by a more tailored approach. In TANGO we have already started to integrate fairness checks and mitigation of unfair behaviors in FRANK [Mazzoni et al., 2024]. However, it has been observed that fair models do not necessarily produce *fair explanations* [Yang and Howe, 2024]. Some work in this latter direction compare explanation quality metrics (fidelity, stability, sparsity) across different social groups [Balagopalan et al., 2022]. This suggests the need of learning strategies able to simultaneously fulfill fairness of decisions, satisfying explanation fairness, and maintaining the utility performance [Zhao et al., 2023b]. In particular, we highlight that it is necessary to develop methods capable of returning explanations for decisions that provide insights that do not disadvantage some social groups against others. This clearly requires to take into consideration by-design that AI systems and their explanations are developed in societal structures that are already unequal and are influenced by common patterns of inequality and

power in agency. For example, women and marginalised genders, people from a low income background and/or people of different ethnic backgrounds, such as migrants, are considered as main categories of underrepresented groups. These categories are part of the six protected grounds from discrimination. Thus, in the suggestions enabling actionable recourse (how to change the model's prediction), it should be guaranteed that explanations will require an effort proportional to the capabilities and opportunities of the different individuals and groups under analysis. Moreover, two important aspects which should not be underestimated are the issues about *intersectional fairness*, i.e., forms of discrimination towards groups of people with different, sensitive characteristics (e.g. gender, race, age, sexuality, religion, disability, etc.), and *fairness in changing environments* where we can have a data or domain distribution drift with respect to the training context [Barrainkua et al., 2023]. In addition, the use of LLMs for generating natural language explanations in form of a dialogue could be also another source of bias and unfairness. For example, the language used in the generated explanation could be inappropriate and unfair with respect to some underrepresented groups.

Privacy Another important issue to be taken into consideration is privacy of individuals. The risk of privacy exposure can depend on (1) *data*, which are not anonymized, (2) *AI model*, which is not trained by a privacy-preserving learning strategy, (3) *explainers and explanations*, which while adding a layer of transparency can reveal personal data of people whose data have been used to train the model or end-user of the model. In TANGO we have developed a framework which enables the assessment of privacy exposure of AI models explained by global or local surrogate models [Naretto et al., 2024]. Detecting the empirical privacy risk is fundamental because it enables to identify and apply mitigation strategy for example based on well-known models such as differential private that was successful applied for some specific feature based explanation methods [Patel et al., 2022]. Privacy-aware explanation mechanisms for decision making processes can also be developed in a federated learning setting where the absence of data sharing helps in maintaining under control the exposure of personal and sensitive information [Haffar et al., 2024].

Persuasion Persuasion, i.e., the process of changing someone's belief, position, or opinion on a specific matter, is a widely studied topic in the social sciences [Keynes, 1932]. LLMs can be considered as models trained to mimic human language and reasoning by ingesting vast amounts of textual data. They have very good performance in generating coherent and contextually relevant text with fluency and versatility. Some experts have widely expressed concerns about the risk of LLMs being used to manipulate online conversations and pollute the information ecosystem by spreading misinformation, exacerbating political polarization, reinforcing echo chambers, and persuading individuals to adopt new beliefs [Salvi et al., 2024]. Moreover, the ability of personalization, obtained conditioning the models' generations on personal attributes and psychological profiles [Bommasani et al., 2021] makes the problem more worrying. The persuasiveness of an LLM in the TANGO context can lead to a biased collaboration where the AI system covers a more important role with respect to humans. Indeed, although humans receive explanations with the goal of a mutual collaboration, the persuasive power can cancel the role of humans as active collaborators that can influence the model learning. Persuasion indeed can lead to a situation where the human opinion is not taken into consideration.

5 Conclusion

In this deliverable, we have defined and discussed how different types of explanation processes should be integrated into interactive machine learning strategies for enabling *synergistic* collaboration between humans and machine learning model. Deliverable D1.1 (WP1) establishes high-level desiderata for enabling synergistic machine learning and decision making, which involves explanations (that might not need to open the black box) and continuous interaction with human stakeholders. In this deliverable, we suggest that this perspective can be completed and implemented through the Situation Awareness model for decision making integrated with explanations. We have discussed how *low-level* and *high-level* explanations correlate with the desiderata of D1.1 and the Situation Awareness model, and what kind of interaction protocols they enable, with different explanations having different pros and cons. Finally, we have also discussed some important ethical and legal issues that can arise from the introduction of explanations and from the interactive process.

References

Kareem Ahmed, Stefano Teso, Kai-Wei Chang, Guy Van den Broeck, and Antonio Vergari. Semantic probabilistic layers for neuro-symbolic learning. *Advances in Neural Information Processing Systems*, 35:29944–29959, 2022.

Jose M. Alonso and A. Bugarín. Expliclas: Automatic generation of explanations in natural language for weka classifiers. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2019. doi: 10.1109/FUZZ-IEEE.2019.8859018.

Jose M. Alonso, Senén Barro, Alberto Bugarín, Kees van Deemter, Claire Gardent, Albert Gatt, Ehud Reiter, Carles Sierra, Mariët Theune, Nava Tintarev, Hitoshi Yano, and Katarzyna Budzynska. Interactive natural language technology for explainable artificial intelligence. In *Trustworthy AI - Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers*, page 63–70, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-73958-4. doi: 10.1007/978-3-030-73959-1_5. URL https://doi.org/10.1007/978-3-030-73959-1_5.

David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*, 2018.

Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Mag.*, 35(4):105–120, 2014.

Afra Amini, Tim Vieira, and Ryan Cotterell. Variational best-of-n alignment. *arXiv*, abs/2407.06057, 2024. URL https://arxiv.org/abs/2407.06057.

Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching categories to human learners with visual explanations. In *CVPR*, pages 3820–3828. Computer Vision Foundation / IEEE Computer Society, 2018.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz,

and Marzyeh Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *FAccT*, pages 1194–1206. ACM, 2022.

Debodeep Banerjee, Stefano Teso, and Andrea Passerini. Learning to guide human experts via personalized large language models. *arXiv*, abs/2308.06039, 2023. URL https://arxiv.org/abs/2308.06039.

Debodeep Banerjee, Stefano Teso, Burcu Sayin, and Andrea Passerini. Learning to guide human decision makers with vision-language models. *arXiv*, 2403.16501, 2024. URL https://arxiv.org/abs/2403.16501.

Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. Preserving the fairness guarantees of classifiers in changing environments: a survey. *ACM Comput. Surv.*, 2023. Just Accepted.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery Journal*, 2023.

Rishi Bommasani, Drew A. Hudson, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021.

Andrea Bontempelli, Stefano Teso, Fausto Giunchiglia, and Andrea Passerini. Concept-level debugging of part-prototype networks. In *International Conference on Learning Representations*, 2023.

Samuele Bortolotti, Emanuele Marconato, Tommaso Carraro, Paolo Morettin, Emile van Krieken, Antonio Vergari, Stefano Teso, and Andrea Passerini. A benchmark suite for systematically evaluating reasoning shortcuts. In *NeurIPS 2024 Datasets & Benchmarks track*, 2024.

James Burton, Noura Al Moubayed, and Amir Enshaei. Natural language explanations for machine learning classification decisions. In *IJCNN*, pages 1–9. IEEE, 2023.

Florian Peter Busch, Roshni Kamath, Rupert Mitchell, Wolfgang Stammer, Kristian Kersting, and Martin Mundt. Where is the truth? the risk of getting confounded in a continual world. *arXiv preprint arXiv:2402.06434*, 2024.

Diego Calanzone, Stefano Teso, and Antonio Vergari. Logically consistent language models via neuro-symbolic integration. *arXiv preprint arXiv:2409.13724*, 2024.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. *NeurIPS*, 2019.

Jessie Y. C. Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3):259–282, 2018. doi: 10.1080/1463922X.2017.1315750.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.283. URL https://aclanthology.org/2024.acl-long.283.

Mark W. Craven and Jude W. Shavlik. Extracting tree-structured representations of trained networks. In *NIPS*, pages 24–30. MIT Press, 1995.

Haikang Deng and Colin Raffel. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11781–11791, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.721. URL https://aclanthology.org/2023.emnlp-main.721.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Pai-Shun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*, pages 590–601, 2018.

Alan Dix, Ben Wilson, Matt Roach, Tommaso Turchi, and Alessio Malizia. Epistemic interaction–tuning interfaces to provide information for ai support. 2024.

Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. *Hum. Factors*, 37(1):32–64, 1995.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. Decore: Decoding by contrasting retrieval heads to mitigate hallucinations. *arXiv*, abs/2410.18860, 2024. URL https://arxiv.org/abs/2410.18860.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.*, 38(5):2770–2824, 2024.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Francesca Naretto, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Min. Knowl. Discov.*, 38(5):2825–2862, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org, 2017.

Karthik S. Gurumoorthy, Amit Dhurandhar, Guillermo A. Cecchi, and Charu C. Aggarwal.

Efficient data representation by selecting prototypes with importance weights. In *ICDM*, pages 260–269. IEEE, 2019.

Rami Haffar, Francesca Naretto, David Sánchez, Anna Monreale, and Josep Domingo-Ferrer. GLOR-FLEX: local to global rule-based explanations for federated learning. In *FUZZ*, pages 1–9. IEEE, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv*, abs/2311.05232, 2023.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123. URL https://aclanthology.org/2023.findings-emnlp.123.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1466.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221, 2022. URL https://api.semanticscholar.org/CorpusID:250451161.

Subbarao Kambhampati, Sarath Sreedharan, Mudit Verma, Yantian Zha, and Lin Guan. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12262–12267, 2022.

John Maynard Keynes. Essays in persuasion. *Pacific Affairs*, 5:905, 1932.

Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*, pages 2280–2288, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Int. conf. on machine learning*, pages 2668–2677. PMLR, 2018.

Gary Klein. Naturalistic decision making. *Hum. Factors*, 50(3):456–460, 2008.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been

Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv*, abs/2302.09664, 2023. URL https://arxiv.org/abs/2302.09664.

Todd Kulesza, Simone Stumpf, Margaret M. Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *VL/HCC*, pages 41–48. IEEE Computer Society, 2010.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Hum. Factors*, 46(1):50–80, 2004.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022. URL https://arxiv.org/abs/2205.14334.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2023. URL https://api.semanticscholar.org/CorpusID:258967487.

Raanan Lipshitz, Gary Klein, Judith Orasanu, and Eduardo Salas. Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, 14(5):331–352, 2001. doi: https://doi.org/10.1002/bdm.381.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.57. URL https://aclanthology.org/2022.naacl-main.57.

Ximing Lu, Faeze Brahman, Peter West, Jaehun Jung, Khyathi Chandu, Abhilasha Ravichander, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, Jillian Fisher, Bill Lin, Skyler Hallinan, Lianhui Qin, Xiang Ren, Sean Welleck, and Yejin Choi. Inference-time policy adapters (IPA): Tailoring extreme-scale LMs without fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages

6863–6883, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.424. URL https://aclanthology.org/2023.emnlp-main.424.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. In *International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, volume 1, pages 1–13, 2021.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural Probabilistic Logic Programming. *NeurIPS*, 2018.

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Glancenets: Interpretabile, leak-proof concept-based models. *NeurIPS*, 2022.

Emanuele Marconato, Gianpaolo Bontempo, Elisa Ficarra, Simone Calderara, Andrea Passerini, and Stefano Teso. Neuro-symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal. In *International Conference on Machine Learning*, pages 23915–23936. PMLR, 2023a.

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. *Entropy*, 25(12):1574, 2023b.

Emanuele Marconato, Samuele Bortolotti, Emile van Krieken, Antonio Vergari, Andrea Passerini, and Stefano Teso. Bears make neuro-symbolic models aware of their reasoning shortcuts. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024a.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. *Advances in Neural Information Processing Systems*, 36, 2024b.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

Federico Mazzoni, Riccardo Guidotti, and Alessio Malizia. A frank system for co-evolutionary hybrid decision-making. In *IDA (2)*, volume 14642 of *Lecture Notes in Computer Science*, pages 236–248. Springer, 2024.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, 2022.

Chris J. Michael, Dina M. Acklin, and Jaelle Scheuerman. On interactive machine learning and the potential of cognitive feedback. *CoRR*, abs/2003.10365, 2020.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50.

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT**, pages 607–617. ACM, 2020.

Francesca Naretto, Anna Monreale, and Fosca Giannotti. Evaluating the privacy exposure of interpretable global and local explainers. *Accepted at Trans. Data Priv.*, 2024.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*, 2022.

Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. In *FAccT*, pages 1895–1904. ACM, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*, 2023.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pages 1135–1144. ACM, 2016.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pages 1527–1535. AAAI Press, 2018.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of large language models: A randomized controlled trial. *CoRR*, abs/2403.14380, 2024.

Lindsay Sanneman and Julie A. Shah. The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems. *International Journal of Human–Computer Interaction*, 38(18-20):1772–1788, 2022. doi: 10.1080/10447318.2022.2081282.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

Gesina Schwalbe. Concept embedding analysis: A review. *arXiv preprint arXiv:2203.13909*, 2022.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico, June 2024.

Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR, 2017.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nat. Mac. Intell.*, 5(8):873–883, 2023.

Aaron Springer and Steve Whittaker. Progressive disclosure: When, why, and how do users want algorithmic transparency information? *ACM Trans. Interact. Intell. Syst.*, 10(4): 29:1–29:32, 2020.

Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proc. of the IEEE/CVF conf. on computer vision and pattern recognition*, pages 3619–3629, 2021.

Wolfgang Stammer, Marius Memmel, Patrick Schramowski, and Kristian Kersting. Interactive disentanglement: Learning concepts by interacting with their prototype representations. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022.

Simone Stumpf, Vidya Rajaram, Lida Li, Margaret M. Burnett, Thomas G. Dietterich, Erin Sullivan, Russell Drummond, and Jonathan L. Herlocker. Toward harnessing user feedback for machine learning. In *IUI*, pages 82–91. ACM, 2007.

Christopher T. H. Teo, Milad Abdollahzadeh, and Ngai-Man Cheung. On measuring fairness in generative models. In *NeurIPS*, 2023.

Stefano Teso, Öznur Alkan, Wolfang Stammer, and Elizabeth Daly. Leveraging explanations in interactive machine learning: An overview. *Frontiers in Artificial Intelligence*, 2023.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv*, abs/1910.08684, 2020. URL https://arxiv.org/abs/1910.08684.

Liam van der Poel, Ryan Cotterell, and Clara Meister. Mutual information alleviates hallucinations in abstractive summarization. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.399. URL https://aclanthology.org/2022.emnlp-main.399.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. Faithfulness-aware decoding strategies for abstractive summarization. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.210. URL https://aclanthology.org/2023.eacl-main.210.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 5546–5558, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.304. URL https://aclanthology.org/2023.acl-long.304.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pages 5502–5511. PMLR, 2018.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.486.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv*, abs/2401.11817, 2024b. URL https://arxiv.org/abs/2401.11817.

Yiwei Yang and Bill Howe. Does a fair model produce fair explanations? relating distributive and procedural fairness. In *HICSS*, pages 6868–6877. ScholarSpace, 2024.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models. *arXiv preprint arXiv:2209.09056*, 2022.

John Zerilli, Umang Bhatt, and Adrian Weller. How transparency modulates trust in artificial intelligence. *Patterns*, 3(4):100455, 2022.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.320. URL https://aclanthology.org/2023.acl-long.320.

Yuying Zhao, Yu Wang, and Tyler Derr. Fairness and explainability: Bridging the gap towards fair model explanations. In *AAAI*, pages 11363–11371. AAAI Press, 2023b.