TANGO

D2.2 – Initial version of methods for HML

Submission Date: 31/03/2025

Due Date: 31/03/2025

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	D2.2 – Initial version of methods for HML		
Type	R	**Dissemination Level**	PU
Lead Beneficiary	Technical University Darmstadt (TUDA)		
Authors	Wolfgang Stammer (TUDA); Tim Tobiasch (TUDA); Stefano Teso (UNITN)		
Co-authors	Thomas Kehrenberg (BCAM); Clara Punzi (SNS); Roberto Pellungrini (SNS); Samuele Bortolotti (UNITN); Andrea Passerini (UNITN)		
Reviewers	Novi Quadrianto (BCAM), Jose Alonso (BCAM)		

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	2025-02-10	Initial ToC	T. Tobiasch (TUDA), W. Stammer (TUDA)
V0.2	2025-02-28	Contributions from all partners	All partners
V0.3	2025-03-04	Revision and harmonization	T. Tobiasch (TUDA), W. Stammer (TUDA), S. Teso (UNITN)
V0.4	2025-03-13	Comments from internal reviewers	N. Quadrianto (BCAM), J. Alonso (BCAM)
V0.5	2025-03-18	Integration of comments	Tobiasch (TUDA), W. Stammer (TUDA)
V0.6	2025-03-27	Final adjustments	R. Acosta (UNITN) A. Passerini (UNITN)

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 present the research activities of WP2, which focuses on the development of cognition-aware explanations and machine learning architectures to support synergistic human-machine learning (HML). Specifically, WP2 addresses five key research directions: (1) developing explanations that enhance human situational awareness; (2) advancing neuro-symbolic models that enable transparent and revisable reasoning; (3) designing interactive learning strategies that incorporate human feedback; (4) developing learning-to-defer algorithms that dynamically manage decision-making responsibilities between humans and AI; and (5) ensuring fairness through bias detection, fairness verification, and adaptive learning techniques. Building on the theoretical insights of WP1 and the explanatory frameworks introduced in D2.1, we highlight the progress made in developing initial machine learning methods for synergistic HML. This deliverable provides an overview of the key methodological contributions from T2.2 to T2.5 and shows how the work of WP2 aligns with the guiding principles of interactivity, contextuality, and rationalization. The deliverable synthesizes results from neuro-symbolic reasoning, (explanation-based) interactive learning, and fairness-aware modeling, contributing to the overall goal of supporting explainable, adaptive, and trustworthy human-AI collaboration in the TANGO project.

Contents

List of Figures

1 Introduction

The TANGO perspective is specifically conceived to target AI-based human empowerment. A key strategic objective (SO2.1) of TANGO is thus 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. Work package 2 (WP2) supports this objective by researching and developing cognition-aware explanations and machine learning architectures that enhance mutual understanding in human-AI teams. It builds on theoretical insights from WP1 and specifically addresses the project's strategic objective SO2.1.

This strategic objective is structured around five key research directions within WP2 which, in turn, are instantiated into five tasks: (1) developing cognition-aware explanations that improve human situational awareness, ensuring that AI-generated insights align with human expectations; (2) designing neuro-symbolic models that learn from human corrections, incorporate prior knowledge, and balance reliability, inspectability, and revisability; (3) advancing synergistic interactive ML algorithms that evolve with human decisions and enhance prediction robustness through explanations and corrective interventions; (4) designing learning-to-defer algorithms that allow AI systems to effectively complement human decision-making, integrating insights from hybrid decision-making paradigms and Large Language Models LLMs; and (5) ensuring fairness in AI models by developing conceptual frameworks, methodologies, and formal verification techniques to assess fairness in interactive learning and learning-to-complement settings. These tasks all contribute toward measurable objective R3, namely the availability of a set of algorithms for synergistic human-machine learning.

To tackle these research directions WP2 builds on the theoretical foundations established in WP1, which in D1.1 identified three key principles essential for integrating interaction and explanations into AI systems to ensure trustworthiness:

> 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 the same way as humans are, without the need to open the black box.

These properties serve as important guiding principles for WP2.

The previous deliverable, D2.1, further integrated the Situation Awareness model [Endsley, 1995] into decision-making frameworks, aligning it with explanation processes that embody the outlined desiderata. This approach aims to enhance human understanding and awareness, ensuring that AI-generated explanations contribute meaningfully to situational comprehension and informed decision-making. A key contribution of D2.1 was a categorization of machine explanations into actionable classes that support such synergistic human-machine learning.

Hereby, D2.1 outlined three primary categories of explanations: feature-driven explanations, which leverage input-output relationships to highlight influential factors in model predictions; concept-based explanations, which organize knowledge into interpretable symbolic structures

that align with human reasoning processes; and natural language explanations, which generate textual justifications tailored to different levels of user expertise. These explanation types were discussed in the context of the three core desiderata identified in WP1. Moreover, by introducing cognition-aware explanation methodologies within TANGO and thereby tying the findings of WP1 to actionable machine learning architectures, deliverable D2.1 had tackled the first aspect of SO2.1.

The goal of this current deliverable D2.2 is to specifically highlight the progress that has been made in terms of developing initial machine learning methods for synergistic HML and, in this way, tackling additional aspects of SO2.1. It relates both to the properties disseminated from WP1, as well as the mapping provided in deliverable D2.1. This deliverable thus discusses in detail key technical contributions from tasks of WP2 that leverage said explanation types to implement different forms of human-machine synergy.

Work in the context of task T2.2 advances neuro-symbolic models that provide transparent, revisable reasoning through structured representations, facilitating human interpretability and control. T2.3 focuses on interactive learning strategies, integrating concept-based representations and memory mechanisms to ensure that AI systems evolve through human guidance and feedback. T2.4 introduces learning-to-defer and learning-to-reject methods, equipping AI with the ability to recognize uncertainty and selectively defer decisions to human experts when necessary. Lastly, T2.5 develops fairness-aware machine learning techniques, incorporating bias detection, formal fairness verification, and adaptive learning mechanisms to promote equitable and trustworthy AI decision-making. Collectively, these contributions reinforce WP1's principles of interactivity, contextuality, and rationalization, ensuring that AI systems in TANGO operate as explainable, adaptive, and human-aligned collaborators.

The deliverable is structured along the tasks of WP2 of the original proposal (*i.e.*, tasks T2.2 - T2.5, whereby task T2.1 was already covered within deliverable D2.1). Per task, we hereby highlight the methodological details of selected publications and briefly summarize further publications at the end of each task section. We conclude this deliverable with a final summary on the current progress of WP2 in the context of initial versions of methods for synergistic human-machine learning.

2 Neuro-Symbolic Models for Understandability and Reliability (T2.2)

In this section, we provide technical details on two specific highlights related to understandable and reliable neuro-symbolic (NeSy) models for hybrid decision systems. We distinguish two main project branches: (1) understandability and generalizability, and (2) reliability specifically in the context of shortcut learning. We conclude this section with a short summary of other important project publications in the context of Task T2.2, as well as the relations to previous project results.

2.1 NeSy Methods for Understandability and Generalizability

Learning abstract concepts from images in an unsupervised manner is particularly challenging in high-stakes decision-making scenarios, where AI models must integrate visual perception with generalizable relational reasoning while ensuring interpretability and trustworthiness.

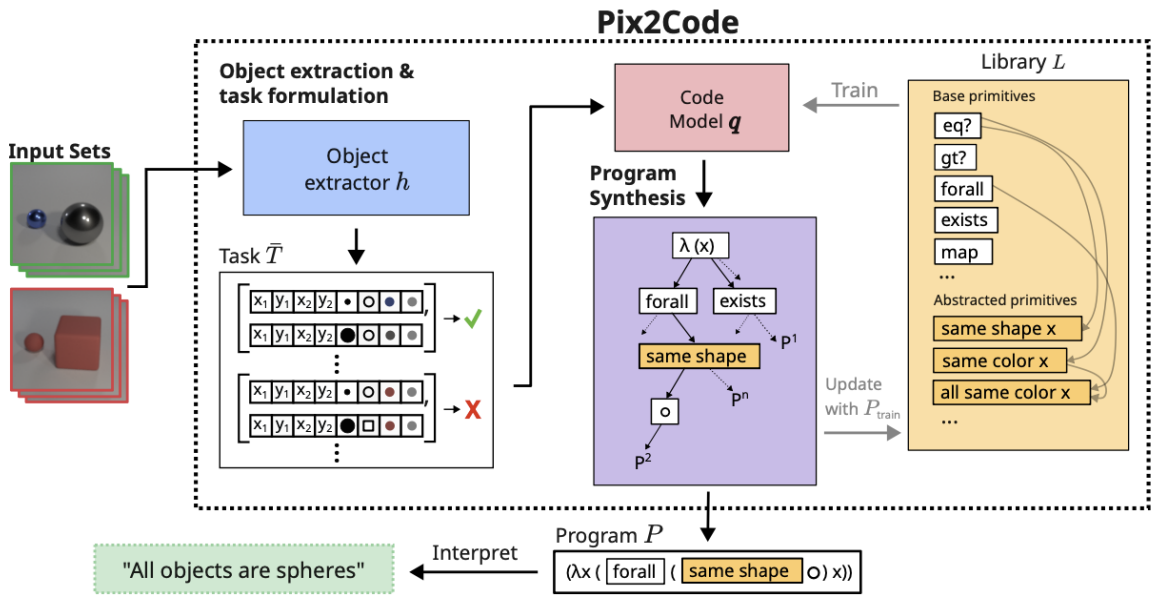

Figure 1: *The Pix2Code architecture.* Object representations are extracted from positive and negative examples and converted into a classification task. Programs to solve them are searched with a probabilistic model over a library of primitives that is learned and enhanced during training.

While humans naturally recognize and apply abstract visual concepts in diverse contexts, machine learning models often struggle with the complexity and variability of visual scenes [Kim et al., 2018, Stabinger et al., 2021] or the combinatorial nature of concept formation [Vedantam et al., 2021]. Moreover, such models largely rely on opaque implicit representations, making it difficult to understand their learned knowledge. This lack of interpretability is a significant barrier to adoption in hybrid decision support systems (HDSS), where transparency and alignment between human and AI reasoning are critical.

To address these challenges, we introduce Pix2Code in Wüst et al. [2024a], a framework that uses program synthesis to learn and represent visual concepts in an explicit, interpretable, and generalizable way. Not only do programs allow extrapolation to novel, unseen inputs, *e.g.*, regardless of the number of objects, but their compositional nature is particularly useful for learning to reuse existing knowledge in new ways [Ellis et al., 2021, Stengel-Eskin et al., 2024, Balog et al., 2017]. In addition, even the longest programs are readable by human users [Cambronero et al., 2023], providing an inherent form of interpretability. Finally, program synthesis approaches offer easy human revision [Trivedi et al., 2021], such as rewriting or updating the internal program representations. Thus, by embedding explainable, programmatic representations in hybrid decision workflows, Pix2Code contributes to TANGO's vision of cognitive AI that adapts to user needs and facilitates bidirectional, explanation-augmented human-AI interaction. We provide details below.

2.1.1 Methodology

Pix2Code (*c.f.* an overview in Fig. 1) extends program synthesis to *visual* relational reasoning by integrating neural object extraction with symbolic program synthesis. This allows the

system to learn structured visual concepts. It consists of three main stages: object extraction, symbolic program synthesis, and interpretability-driven revision.

In the first stage, a pre-trained neural network processes input images to extract symbolic object representations, identifying key attributes for each object, such as shape, color, and position. These extracted representations are converted into a structured format suitable for further processing.

The second stage involves synthesizing programs based on this structured input, where a domain-specific language (DSL) is used to construct relational visual concepts as compositional programs. Given the symbolic object representations, the program synthesis component searches for programs that best separate positive and negative image examples of a concept. This is achieved by an enumerative search guided by a probabilistic model that prioritizes likely program structures and uses a probabilistic wake-sleep learning framework to continuously refine its program library.

During this process, a learned library of program primitives is continuously updated, allowing Pix2Code to evolve its representations dynamically. The program library captures increasingly complex visual concepts over time, enabling better generalization to new tasks. The DSL includes basic primitives such as logical operators (e.g., and, or, not), relational predicates (eq?, gt?), and higher-order functions (forall, exists, map, fold). These primitives allow Pix2Code to compose expressive and interpretable rules that define abstract visual concepts, such as “all objects are spheres” or “there exists at least one red triangle.”

The third stage, interpretability-driven revision, allows users to inspect and refine learned concepts. Because concepts are represented as explicit programs, they can be revised by modifying program primitives, introducing new domain-specific rules, or removing spurious correlations. This process allows for human-in-the-loop adjustments to ensure that the learned concepts are consistent with the intended visual reasoning tasks. In contrast to end-to-end neural approaches, which often produce opaque representations, by leveraging this hybrid neuro-symbolic approach, Pix2Code provides a transparent and adaptive framework for unsupervised visual concept learning while ensuring that learned concepts remain interpretable, generalizable, and revisable.

Beyond its core methodology, Pix2Code supports compositional learning, meaning that once a concept is learned, it can be reused to generate new concepts without retraining the entire system. This modularity significantly increases the framework’s ability to generalize to unseen data configurations, such as different task domains and changed entity settings. In addition, the use of program synthesis ensures scalability, as complex visual concepts can be efficiently represented and manipulated through hierarchical program structures.

Unlike traditional deep learning models, which often rely on extensive labeled data, Pix2Code can learn these structured representations with minimal supervision by leveraging weakly labeled data. Thus, instead of requiring explicit annotations for each relational concept, the framework derives its programmatic representations from loosely specified concept examples, using positive and negative image sets as indirect supervision. This weak supervision allows Pix2Code to discover and refine abstract visual concepts without exhaustive manual labeling, making it particularly suitable for real-world applications where labeled data is scarce or expensive to obtain.

Original Concept	Retrieved Program	GPT4 Natural Language Translation
All objects are cyan	(λ (x) ((λ (y z) (forall (λ (u) (eq? (index y u) z)))) color cyan x))	"All objects are cyan in color."
All objects are cubes	(λ (x) ((λ (y z) (forall (λ (u) (eq? (index y u) z)))) shape cube x))	"All objects are cubes."
All objects are purple and all objects are spheres	(λ (x) (and ((λ (y z) (forall (λ (u) (eq? (index y u) z)))) shape sphere x) ((λ (v w) (forall (λ (a) (eq? (index v a) w)))) color purple x)))	"All objects are spheres and all objects are purple."
All objects are brown or all objects are cubes	(λ (x) (or ((λ (y z) (forall (λ (u) (eq? (index y u) z)))) shape cube x) ((λ (v w) (forall (λ (a) (eq? (index v a) w)))) color brown x)))	"All objects are either cubes or all objects are brown."
All objects are small and there exists a purple object	(λ (x) (and ((λ (y z) (forall (λ (u) (eq? (index y u) z)))) size small x) (exists (λ (v) ((λ (w a b) (eq? (index b w) a)) v purple color)) x)))	"All objects are small in size, and there is at least one purple object."
There are three gray objects	(λ (x) (eq? ((λ (y) (count (map (λ (z) ((λ (u v) (index u v)) color z)) y))) x gray) 3))	"There are three objects that are gray."
There exists an arbitrary object and there exist three other objects that are blue	(λ (x) (gt? ((λ (y) (count (map (λ (z) ((λ (u v) (index u v)) color z)) y))) x blue) 2))	"There are more than two objects that are blue in color."

Table 5: Pix2Code's concepts are transparent programs and can be translated to natural language by LLMs. CURI concepts (left), Pix2Code's program representations (middle) and (right) natural language translations (via gpt-4-turbo).

2.1.2 Results

To evaluate the effectiveness of Pix2Code, we apply it to two benchmark datasets: Kandinsky Patterns [Müller and Holzinger, 2021] and CURI [Vedantam et al., 2021]. These datasets challenge the model to recognize and reason about structured relational concepts in both 2D and 3D visual settings.

Our evaluation focused on several key performance aspects: overall concept discovery, compositional generalization, scalability, and interpretability. Pix2Code consistently outperformed the baseline neural models on all of these measures. It demonstrated a superior ability to discriminate between positive and negative concept examples and achieved high accuracy in recognizing compositional structures in images. Importantly, compared to purely neural models that struggled with unseen concept configurations, Pix2Code effectively generalized to novel relational structures and varying numbers of objects.

A critical aspect of Pix2Code's effectiveness is its transparency and verifiability. Unlike black-box deep learning models that produce opaque feature representations, Pix2Code generates human-readable programs that explicitly define the recognized visual concepts (*c.f.* Tab. 5). This interpretability allows users to inspect and refine the synthesized programs if necessary. We conducted additional experiments to explore the revisability of learned concepts. By modifying Pix2Code's program library, users were able to correct shortcut learning behavior, ensuring that the model learned the intended visual relationships rather than spurious correlations in the training data.

Another key strength of Pix2Code is its scalability and modularity. Once a relational concept is learned, it can be reused to construct more complex representations without the need for

complete retraining. This aspect significantly improves efficiency and allows the system to dynamically expand its knowledge base. Our results show that Pix2Code maintains high performance even as the complexity of visual scenes increases, supporting its potential as a robust approach to unsupervised visual concept learning.

In summary, our results suggest that Pix2Code offers a promising direction for interpretable and robust visual concept learning. By using program synthesis, it combines the strengths of neural and symbolic reasoning, allowing for improved generalization, human interpretability, and revisability. This approach has significant potential for future applications in AI systems that require transparent, adaptive, and scalable visual reasoning capabilities.

2.2 Addressing Reasoning Shortcuts for Reliable Neuro-Symbolic AI

The *raison d'être* of NeSy predictors is to encourage [Xu et al., 2018] or guarantee [Manhaeve et al., 2018, Giunchiglia and Lukasiewicz, 2020, Ahmed et al., 2022, Hoernle et al., 2022] that their predictions are consistent with the prior knowledge provided upfront, which encodes essential structural or safety constraints, thus providing improved reliability over neural baselines [Di Liello et al., 2020, Hoernle et al., 2022, Giunchiglia and Lukasiewicz, 2020]. NeSy architectures typically consist of two stages: a first stage performs perception, i.e., it converts the input (e.g., image) into a set of high-level concepts; the second stage applies logical reasoning to the extracted concepts to infer a prediction. It is at this stage that prior knowledge is taken into account.

However, it has recently been shown that NeSy architectures can achieve high prediction accuracy by learning concepts with *unintended semantics* [Marconato et al., 2024b, Li et al., 2023], with serious consequences for trustworthiness. Consider Fig. 2: here, a NeSy model is applied to the `BDD-OIA` [Xu et al., 2020, Sawada and Nakamura, 2022] autonomous driving task. The model must predict safe actions based on the content of a dashcam image, under the constraint that the vehicle must stop whenever it detects a pedestrian or a red light. The model can achieve perfect accuracy *and* satisfy the constraint by mistaking pedestrians for red lights, precisely because both entail the correct (stop) action [Marconato et al., 2024b].

These so-called ***reasoning shortcuts*** (RSs) occur because prior knowledge and data may be insufficient to determine the intended semantics of all concepts, and – as our example shows – cannot be avoided by maximizing prediction accuracy alone. They compromise in-distribution [Wang et al., 2023] and out-of-distribution generalization [Marconato et al., 2024b, Li et al., 2023], continual learning [Marconato et al., 2023a], reliability of neuro-symbolic verification tools [Xie et al., 2022], and concept-based interpretability [Koh et al., 2020, Marconato et al., 2023b] and debugging [Teso et al., 2023]. Importantly, unsupervised strategies for mitigating RSs either offer no guarantees or work under restrictive assumptions, while supervised ones rely on expensive annotations (e.g., concept supervision) [Marconato et al., 2024b]. Worse still, state-of-the-art NeSy architectures predict concepts affected by RSs *with high confidence*, making it impossible to distinguish between reliable and unreliable concept predictions.

A second key issue for understanding and addressing RSs is that appropriate datasets – i.e., NeSy tasks known to lead to RSs – are scarce and scattered throughout the literature, hindering research on this challenging problem. Current benchmark suites for learning and reasoning neglect RSs altogether [Vermeulen et al., 2023] and lack OOD data suitable for investigating their impact, while others are limited to larger models [Yu et al., 2023, Yue et al.,

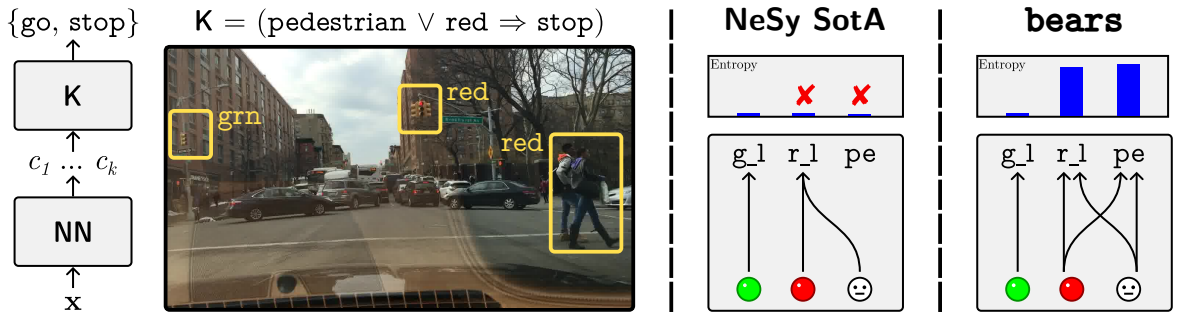

Figure 2: Illustration of `bears` on an autonomous driving task. A Neuro-Symbolic (NeSy) model processes a traffic scene by extracting concepts via a neural network (NN) and reasoning over them using symbolic knowledge K. While state-of-the-art NeSy models suffer from reasoning shortcuts (e.g., associating pedestrians and red lights with stopping), `bears` mitigates this issue by introducing uncertainty calibration, reducing overconfidence and improving interpretability.

2023, Hendrycks et al., 2021]. At the same time, available datasets annotated with concept supervision (*e.g.* CUB200 [Wah et al., 2011]), which is essential for assessing concept quality, do not require logical reasoning or prior knowledge.

2.2.1 Methodology

We tackle the first issue – overconfidence about poor concepts – as follows. Instead of trying to avoid RSs altogether, we suggest that NeSy predictors should be *aware of their reasoning shortcuts*, i.e., they should assign lower confidence to concepts affected by RSs, thus allowing users to identify and avoid low-quality predictions while maintaining high accuracy. We say that a NeSy predictor is *reasoning shortcut-aware* if it satisfies the following desiderata:

I. **Calibration**: For all concepts *not* affected by RSs, the system should achieve high accuracy and be highly confident. Vice versa, for all concepts that *are* affected by RSs, the model should have low confidence.

II. **Performance**: The NeSy predictor should achieve high *label accuracy* even if RSs are present.

III. **Cost effectiveness**: The system should not rely on expensive mitigation strategies.

Calibration (**I**) captures the essence of our proposal: a model that knows which of the learned concepts are affected by RSs can prevent its users from blindly trusting and reusing them. Of course, this should not come at the expense of prediction accuracy (**II**) or expensive concept-level annotations (**III**), so as not to hinder applicability.

Unfortunately, as we show empirically in [Marconato et al., 2024a], state-of-the-art NeSy architectures are *not* RS-aware. We address this problem by introducing `bears` (BE Aware of Reasoning Shortcuts), a simple but effective method for making NeSy predictors RS-aware – in the sense that they satisfy **I**-**III** – which does not rely on costly dense monitoring. `bears` replaces the concept extraction module with a diversified *ensemble* specifically trained to promote concepts whose uncertainty is proportional to how strongly they are affected by RSs.

Table 6: `bears` *dramatically improves RS-awareness in the real-world.* Results on `BDD-OIA` with `DeepProbLog` show substantial expected calibration error ECE improvements both jointly (mECEC) and for different classes of concepts (F= `forward`, S= `stop`, R= `turn right`, L= `turn left`).

	mECE_C	$\mathrm{ECE}_C(F,S)$	$\mathrm{ECE}_C(R)$	$\mathrm{ECE}_C(L)$
DPL	0.84 ± 0.01	0.75 ± 0.17	0.79 ± 0.05	0.59 ± 0.32
+ MCDrop	0.83 ± 0.01	0.72 ± 0.19	0.76 ± 0.08	0.55 ± 0.33
+ Laplace	0.85 ± 0.01	0.84 ± 0.10	0.87 ± 0.04	0.67 ± 0.19
+ ProbCBM	0.68 ± 0.01	0.26 ± 0.01	0.26 ± 0.02	0.11 ± 0.02
+ DeepEns	0.79 ± 0.01	0.62 ± 0.03	0.71 ± 0.10	0.37 ± 0.12
+ `bears`	$\mathbf{0.58 \pm 0.01}$	$\mathbf{0.14 \pm 0.01}$	$\mathbf{0.10 \pm 0.01}$	$\mathbf{0.02 \pm 0.01}$

The intuition behind `bears` is illustrated in Fig. 2 (right): in `bears`, perception is performed by an ensemble of neural concept extractors that are specifically trained so that each extractor captures a different RS. The overall effect is this: *i*) For concepts that are not affected by RSs (like green lights in the example), all extractors will agree on their semantics and the ensemble will assign them high confidence; *ii*) For concepts that *are* affected by RSs, the extractors will disagree on their semantics, resulting in low confidence concept predictions. This strategy has solid theoretical support [Marconato et al., 2024a].

Note that `bears` is model-agnostic, in the sense that it can be applied to many state-of-the-art NeSy architectures, including DeepProbLog [Manhaeve et al., 2018], logic tensor networks [Badreddine et al., 2022], semantic-probabilistic layers [Ahmed et al., 2022], and in general to all NeSy models that follow the aforementioned two-stage setup.

2.2.2 Results

Our experiments show that `bears` successfully improves the RS-awareness of three state-of-the-art NeSy architectures on four NeSy datasets, including a high-stakes autonomous driving task, and enables us to design a simple but effective active learning strategy for acquiring concept annotations for mitigation purposes. Table 6 shows that in `BDD-OIA`, the expected calibration error (ECE) and the calibration error for the actions *forward*, *stop*, *move right*, and *move left* are much lower compared to the various competitors. Thus, by improving RS-awareness, `bears` helps users of NeSy predictors to identify low-quality concepts and models, increasing the overall trustworthiness of this class of architectures. Moreover, it enables the model itself to request human anntoators to supply targeted supervision for those low-quality concepts that – thanks to `bears` – they are not confident in anymore.

Regarding the dataset issue, we have further proposed `rsbench`, an integrated benchmark suite that provides all the ingredients needed to systematically evaluate the impact of RS and the effectiveness of mitigation strategies. It includes a curated collection of learning and reasoning tasks that have been shown to be affected by RS. In addition, `rsbench` includes entirely new and already established tasks of different flavors – arithmetic, logical, and high-stakes – along with associated data sets and data generators for evaluating OOD scenarios. It also includes `Python` implementations of quality metrics useful for assessing the impact of RSs on NeSy models and, more generally, the reliability of concepts learned (explicitly or

implicitly) by other concept-based architectures and end-to-end neural networks. Finally, a novel algorithm, `countrss`, that uses automated reasoning techniques [Biere et al., 2009] to verify a priori whether a task is affected by RSs and to count them.

`rsbench` consists of three arithmetic datasets based on `MNIST` [LeCun, 1998] digits: two where the default logic is sum, and one where customization is possible. There are also three logic datasets: one based on `MNIST`, one based on `Kandinsky patterns` [Müller and Holzinger, 2021], and one based on `CLEVR` [Johnson et al., 2017] where the logic is fully customizable. Finally, two high stakes records are included: `BDD-OIA` [Xu et al., 2020], the autonomous driving dataset, and `SDD-OIA`, a synthetic version of `BDD-OIA` where the rules are traffic laws. In addition, we proposed metrics to evaluate the impact of reasoning shortcuts (RSs), such as accuracy and F1 score of concepts, as well as *collapse*, which measures the tendency of the model to use fewer concepts than available. Finally, we developed a software called `countrss` that, given the learning specifications (knowledge, data, and concept monitoring), returns the number of reasoning shortcuts the problem allows.

2.3 Further Work

To improve the inspectability and revisability of unsupervised visual concept learning, particularly for single-object concepts, we additionally introduce the Neural Concept Binder (NCB) framework [Stammer et al., 2024]. NCB enables the derivation of both discrete and continuous concept representations, allowing the identification of object properties such as color and shape without explicit human supervision. It integrates recent advances in object-factor disentanglement [Singh et al., 2023], hierarchical clustering [Campello et al., 2013], and retrieval-based inference to extract structured object-centric concept representations. Unlike standard concept learning models, which struggle to balance expressiveness and interperability, NCB achieves this by combining a *soft-binding* mechanism for continuous representations with a retrieval-based *hard-binding* mechanism that leads to discrete counterparts of the continuous encodings. NCB's generalization robustness and ability to refine concepts post hoc were demonstrated using the CLEVR Sudoku benchmark, a newly introduced dataset that combines visual perception with complex symbolic reasoning tasks. The results highlight the potential of NCB to bridge the gap between neural and symbolic computation, and to provide an approach to learning basic concepts in an unsupervised manner that is interpretable and adaptive.

In [Wüst et al., 2024b], we have also investigated the ability of state-of-the-art vision-language models (VLMs) to handle complex visual reasoning tasks with open and evolving vocabularies. The primary goal is to assess how well modern VLMs, such as GPT-4o and LLaVA, can recognize abstract patterns and infer underlying rules in Bongard problems (BPs) — a classic benchmark designed to test human-like concept learning and reasoning. A key finding is that VLMs have significant difficulty with visual abstraction and generalization, often failing to recognize even simple concepts such as shape symmetry or directional patterns. Even when a relevant concept is explicitly presented, their performance remains inconsistent, suggesting a lack of robust symbolic reasoning. In addition, the study introduces a diagnostic multiple-choice setting that allows for a structured comparison of the models' ability to identify valid rule pairs among predefined choices. These controlled evaluations show that the overall success rates of current models remain remarkably low compared to human-level performance. By evaluating VLMs in these settings, the study aims to uncover limitations of

current large models and explore ways to integrate their perceptual and reasoning capabilities into neuro-symbolic architectures.

At the same time, we have been working to make LLMs themselves more reliable. While these models are a promising venue for natural language understanding and generation, they often generate untrue information and contradict themselves in reasoning tasks. To address this problem, we have developed a novel loss, rooted in probabilistic reasoning and knowledge compilation, that dramatically improves the factuality and self-consistency of an LLM with respect to arbitrary reasoning tasks [Calanzone et al., 2024]. A more advanced version of this work has been accepted at ICLR'25 [Calanzone et al., 2025]. This new fine-tuning approach teaches the LLM to respect logical constraints - encoded as propositional formulas - when reasoning about facts, allowing it to remain consistent even when only a small set of facts is available. By integrating these constraints directly into the training, the model learns to satisfy rules such as implication or negation consistency without the need for external reasoning tools at inference time. This results in logically consistent LLMs (LoCo-LLMs) that outperform standard fine-tuning and solver-based approaches, especially in low-data regimes and for unseen but semantically related facts. This is a step toward embedding reasoning directly into the internal beliefs of LLMs, enabling more reliable, interpretable, and scalable reasoning.

We have also introduced *Uncertainty Matters* [Barrainkua et al., 2024], a framework for more robust reliability assessment. It is a probabilistic approach designed to perform uncertainty-aware, multi-objective algorithmic comparisons, including, for example, estimating uncertainty-aware differences in performance and fairness guarantees between two algorithms. This framework can be effectively applied to various algorithmic evaluation scenarios, such as hold-out and K-fold cross-validation, explicitly modeling the correlation between folds to account for dependencies introduced by overlapping training sets. Using a Bayesian approach, *Uncertainty Matters* treats key performance and fairness metrics as random variables derived from confusion matrices, allowing joint estimation of their posterior distributions. This results in more stable, informative, and reliable comparisons, reduces the sensitivity of inferences to arbitrary data splits, and provides a clearer understanding of both algorithmic performance and fairness in contexts with limited data and multiple objectives.

Building on the foundations of causal modeling, we have extended this framework by introducing meta-causal types and models [Willig et al., 2024]. Causal modeling is closely related to neuro-symbolic learning, as it enables the representation and discovery of symbolic structures from data. The proposed extension allows the representation of more complex and dynamic mechanisms in evolving systems, where causal relationships are not fixed but shift due to external factors or agent interventions. A key advance in this work is the introduction of meta-causal states, which categorize different causal graphs into equivalence classes based on their qualitative behavior. These states capture transitions between causal structures, allowing for a more principled analysis of how causal relationships emerge, change, or disappear due to external interventions or internal system dynamics. In addition, the framework supports the identification of root causes at the meta-level, which is particularly useful for analyzing dynamic environments where agents actively shape causal dependencies. By incorporating these insights, meta-causal reasoning goes beyond classical causal inference to provide a new paradigm for algorithmic reasoning about complex, evolving causal structures.

2.4 Summary

Task T2.2 advances understandable and reliable neuro-symbolic (NeSy) models for hybrid decision-making, in line with the WP1 principles of interactivity, contextuality, and rationalization. Methods such as Pix2Code enhance interactivity (D1) by using explicit program-based representations that provide inherent justifications for model reasoning, while rule-based rationalization ensures transparency and supports rationalization (D3). By incorporating interpretability-driven revision, these models enable real-time human-in-the-loop interaction, ensuring that AI reasoning remains inspectable and revisable. In addition, this method incorporates the actionable explanation framework from D2.1, using concept-based representations and natural language justifications to enhance contextuality (D2). Hereby, symbolic program synthesis ensures alignment with human reasoning, while natural language translation makes NeSy output more interpretable and accessible.

Furthermore, in the context of WP1 principles, `bears` enhances interactivity (D1) by ensuring that models provide calibrated concept-level explanations, allowing users to identify and distrust uncertain predictions. It improves contextuality (D2) by dynamically adapting confidence levels to task-specific reasoning challenges, ensuring that explanations remain sensitive to local variations in the data. `bears` also improves rationalization (D3) by improving uncertainty calibration. These contributions are consistent with the explanation framework of D2.1, in particular by improving concept-based representations through a diversified ensemble of neural concept extractors, which improves semantic consistency in AI reasoning.

3 Synergistic Interactive Machine Learning Strategies (T2.3)

In this section, we provide details on novel methods that enable synergistic learning strategies. We focus on synergistic reinforcement learning based on concept bottleneck models and conclude with short summaries of other relevant publications and their relation to previous project results.

3.1 Interactive Learning Strategies with CBMs for Reinforcement Learning

Deep reinforcement learning (RL) agents face numerous challenges that hinder their ability to learn optimal or generalizable policies. These include reward sparsity [Andrychowicz et al., 2017], difficult credit assignment [Raposo et al., 2021], and goal misalignment [di Langosco et al., 2022], where agents learn unintended shortcut strategies that do not generalize well to new scenarios. One of the most critical obstacles is the black-box nature of deep neural networks, which prevents domain experts from inspecting and revising learned policies.

Traditional explainable AI (XAI) methods often fail to faithfully represent an RL agent's decision-making process, making it difficult to diagnose and mitigate these issues. To address these limitations, we introduce Successive Concept Bottleneck Agents (SCoBots) [Delfosse et al., 2024]—a novel framework that enhances transparency and interpretability in RL by integrating hierarchical concept bottleneck (CB) layers. Unlike existing CB models [Koh et al., 2020, Stammer et al., 2021] that focus on individual object properties, SCoBots capture both object attributes and relational concepts, allowing for a more structured and explainable RL decision process.

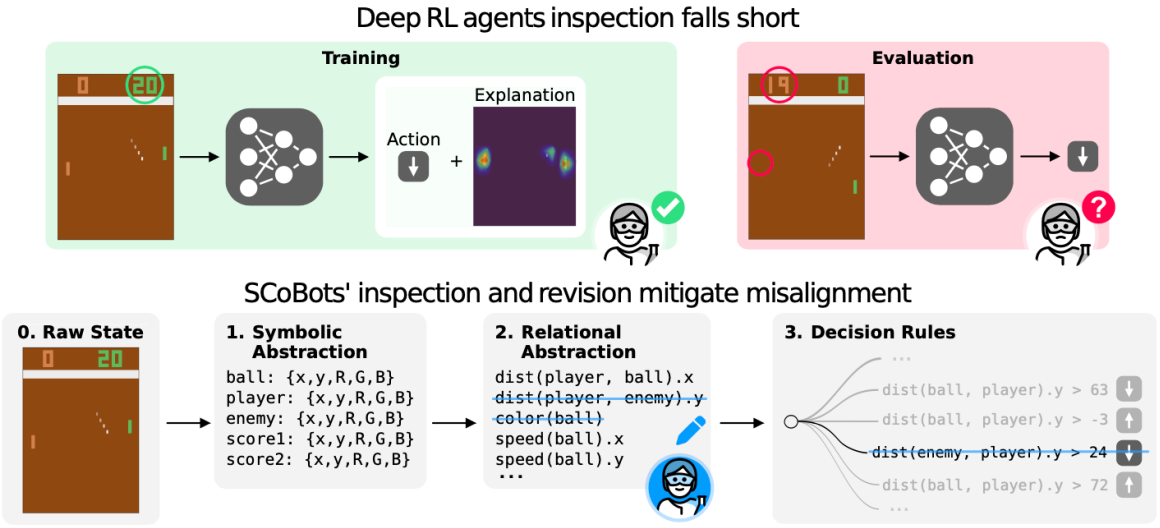

Figure 3: *Top:* Deep RL agents trained on Pong. When the enemy is hidden, the deep RL agent fails to catch the ball. *Bottom:* Successive Concept Bottlenecks Agents (SCoBots) allow for multi-level inspection and revision of reasoning preventing the agents from focusing on potentially misleading concepts.

3.1.1 Methodology

SCoBots introduce a hierarchical approach to concept-based RL by structuring decision-making through successive concept bottleneck layers. The framework consists of symbolic object abstraction, relational concept extraction, and decision rule formulation, forming a pipeline that enhances the interpretability of RL agents while maintaining strong performance.

The first stage, symbolic object abstraction, involves extracting structured representations of objects from raw environmental states. These representations include essential object properties such as position and color, ensuring that the agent's reasoning is based on meaningful and interpretable components rather than low-level visual input. This stage allows RL models to transition from opaque neural representations to explicit symbolic encodings, providing the basis for comprehensible subsequent reasoning steps.

Next, relational concept extraction builds on this structured representation by inferring higher-order relationships between objects. Using predefined relational functions, SCoBots extract key interaction patterns such as distances, motion correlations, and object dependencies. In contrast to standard bottleneck models that only consider individual object attributes, SCoBots emphasize the importance of contextual relationships that are crucial for decision-making in dynamic environments. These relational abstractions allow agents to capture structured dependencies, resulting in more robust and generalizable policies.

Finally, the decision rule formulation phase uses these extracted concepts to determine actions. Rather than relying on opaque deep networks, SCoBots employ decision trees to represent learned policies, making the agent's choices easily inspectable and modifiable. Specifically, a decision tree is distilled from a trained neural action model such that the decision tree action model accurately reflects its neural counterpart [Çağlar Aytekin, 2022]. Overall, this step ensures that the decision process remains interpretable at every stage. The

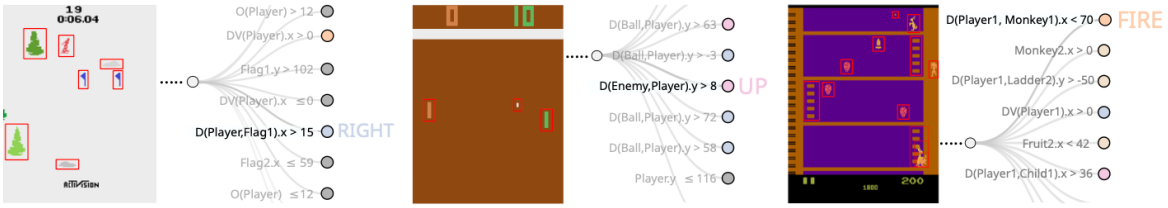

Figure 4: ***SCoBots allow to follow their decision process, because of their object centricity and interpretable concepts.*** **The decision processes of SCoBots (extracted from the decision tree action models) with the associated states from a guided SCoBot on Skiing (left), and from normal SCoBots on Pong (middle) and Kangaroo (right).**

ability to intervene at multiple levels of abstraction substantially enhances the adaptability of RL models, allowing for targeted corrections without requiring full retraining.

A key feature of SCoBots is their interactive refinement capability. Experts can review learned concepts, remove irrelevant or misleading features, and introduce domain knowledge through structured constraints. This interaction allows SCoBots to address common RL-specific problems such as shortcut learning and reward sparsity. By leveraging human-aligned relational abstractions, SCoBots ensure that their learned behaviors remain transparent and adaptive, mitigating the risks associated with opaque reinforcement learning approaches.

3.1.2 Results

To evaluate the effectiveness of SCoBots, we conducted experiments using Atari Learning Environments (ALE) [Bellemare et al., 2015], a widely used benchmark for RL. Specifically, we analyzed the ability of SCoBots to detect and mitigate policy misalignment in the classic game Pong, where traditional deep RL agents often learn to track the opponent's paddle instead of the ball.

Our results demonstrate that SCoBots achieve competitive performance with deep RL agents while greatly improving interpretability and revisability. Across multiple Atari games, including Pong, Kangaroo, and Skiing, SCoBots exhibited human-aligned decision processes, allowing domain experts to diagnose and correct potential failure modes.

In Pong, we discovered that deep RL agents often rely on the position of the opponent's paddle rather than the position of the ball, leading to shortcut learning behaviors that fail when the opponent's paddle is hidden. By examining relational concepts in SCoBots, we identified and mitigated this problem by removing the influence of the enemy paddle from the decision process. The revised agent successfully generalized to new variations of Pong where traditional deep RL agents failed.

In Skiing, where agents typically receive only sparse rewards based on final performance, we introduced an additional object-centric reward signal based on relational concepts (e.g., proximity to the next flag). This guided reward mechanism improved the credit assignment process, leading to faster convergence and higher final scores.

In Kangaroo, deep RL agents often maximize reward by attacking enemies rather than completing the intended goal of rescuing Joey. By modifying the relational concept represen-

tations, we aligned agent behavior with the true game objective, demonstrating the flexibility of SCoBots in guiding RL policies toward human-intended goals.

Overall, our results highlight that SCoBots provide a transparent and structured approach to reinforcement learning, allowing for both high-performance and expert-driven revisions. This work paves the way for more trustworthy and adaptive RL systems also in hybrid decision scenarios, i.e., in real-world applications that require alignment with human understanding and intent.

3.2 Further Work

We have also explored novel approaches that focus on the bidirectional nature of explanations between humans and AI systems. Through the development of theoretical frameworks and experimental platforms, we have explored how explanations provided by humans can meaningfully guide AI systems toward more human-centered decision-making. In Dix et al. [2025], we have the conceptualization and initial development of mechanisms that enable AI systems to incorporate human explanatory feedback, moving beyond traditional one-way explanations from AI to humans and allowing for more flexibility than traditional explanation-based interaction. The research explores different forms of human explanation - from structured feature constraints to free-text rationales - and proposes architectural approaches for integrating them into machine learning systems. This work provides important insights for developing more effective human-AI collaborative systems, where human reasoning can actively shape both the immediate responses and long-term learning of AI systems.

Moving to concept-based models, we have further developed new strategies for improving AI models through human corrections, specifically interventional corrections, ensuring that AI systems can *effectively* learn from such user feedback and adapt over time. Traditional Concept Bottleneck Models (CBMs) allow human users to review and correct concept-level decisions, but these interventions are typically applied once and discarded, limiting their long-term impact. To address this, we have introduced Concept Bottleneck Memory Models (CB2M) [Steinmann et al., 2024], a framework that enables AI systems to dynamically retain and reuse human feedback, ensuring that similar errors are automatically corrected in future decisions. A key technical advancement in CB2M is its dual memory mechanism, which consists of (1) an error memory that tracks past errors and (2) an intervention memory that stores and reuses human corrections. This allows hybrid AI-human systems to recognize recurring decision patterns and proactively apply relevant corrections, minimizing redundant human intervention. In addition, CB2M incorporates a similarity-based retrieval system using nearest-neighbor search in latent space, enabling it to identify potential decision errors before they occur and request human input only when necessary. Overall, this approach allows to reduce the cognitive load on users while improving decision quality in high-stakes applications.

3.3 Summary

Task T2.3 develops synergistic interactive machine learning strategies that enhance human-AI collaboration, consistent with the principles of WP1 and the explanatory framework of D2.1. Successive concept bottleneck agents (SCoBots) enhance interactivity (D1) by structuring reinforcement learning (RL) through hierarchical concept bottleneck layers, enabling transparent, revisable decision-making. Unlike opaque RL models, SCoBots use symbolic object

representations and relational concepts, ensuring contextuality (D2) by grounding AI reasoning in task-specific dependencies. This directly supports the concept-based explanations of D2.1, as SCoBots emphasize object-centric and relational reasoning over uninterpretable latent-space features.

In addition, Concept Bottleneck Memory Models (CB2M) enhance rationalization (D3) by introducing a dual memory mechanism that retains past mistakes and human corrections, ensuring that AI systems adapt dynamically rather than discarding feedback. By integrating nearest-neighbor retrieval, CB2M enables consistent, human-aligned learning, reinforcing D2.1's framework of concept-based, but, in principle, also natural language explanations.

4 Learning to Complement Humans (T2.4)

In this section, we present novel methods that enhance hybrid AI-human decision-making by enabling AI systems to effectively complement human expertise. We focus on self-aware learning approaches, particularly mechanisms for learning to reject and learning to defer, which allow AI models to recognize uncertainty and abstain when necessary. The section concludes with an overview of additional recent work and finishes with a summary of the task's methods in the context of TANGO's previous findings.

4.1 Self-aware Learning

A key challenge in trustworthy AI-based decision-making is dealing with uncertainty, especially in cases where the model is not sufficiently confident in its predictions. In such scenarios, AI models can employ an abstention mechanism, commonly known as "Learning to Abstain" (L2A). This allows the model to defer uncertain decisions to a human expert or a more advanced system. However, existing L2A approaches often lack transparency in their rejection policies, leading to reduced user trust and limited adoption.

Punzi et al. [2024a] introduces **Learning to Reject via Local Rule-based Explanations (L2loRe)**, a novel method for improving the explainability of rejection policies in AI models. L2loRe exploits the concept of counterfactual explanations, using the distance between data points and their counterfactual counterparts to determine confidence levels. By incorporating these distances into the rejection mechanism, the approach provides interpretable rejection criteria and increases trust in AI decision-making.

4.1.1 Methodology

L2loRe operates within the **Learning to Reject (L2R)** framework, which consists of a prediction model and a rejection policy. The overall structure of the method is illustrated in Fig. 5. The predictive model, denoted f, maps input features to an outcome, while the rejection policy, ρ, determines whether the model should make a prediction or defer the decision based on a confidence score.

Confidence is quantified by measuring the proximity of a data instance to its counterfactuals. The closer a data point is to its counterfactual, the lower its confidence score, indicating greater uncertainty. To achieve this, L2loRe first generates counterfactuals for each instance using rule-based explainability techniques, specifically Local Rule-based Explanations (LORE). The

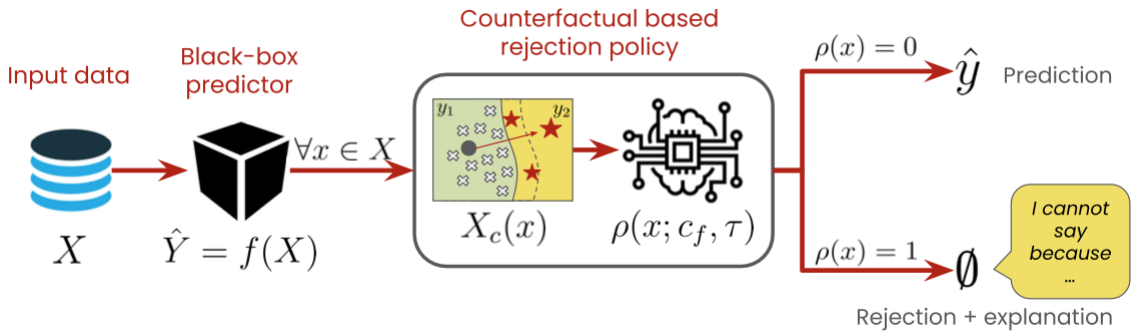

Figure 5: *Overview of L2loRe.* A black-box predictor $f(X)$ generates predictions, and a counterfactual-based rejection policy determines whether to accept or reject them. If the instance is uncertain ($\rho(x) = 1$), the model abstains from making a decision and provides an explanation for the rejection.

confidence function c_f is then computed as the distance between an instance and its closest counterfactual. If this confidence value falls below a rejection threshold τ_r, the instance is rejected, where the rejection threshold is determined by constrained optimization to balance classification accuracy with rejection quality. The rejection function is formally defined as follows:

$$\rho(x; c_f, \tau_r) = \begin{cases} 1, & \text{if } c_f(x) < \tau_r \text{ (Reject)} \\ 0, & \text{otherwise (Accept)} \end{cases}$$

Beyond classification and rejection, L2loRe provides human-readable justifications for both predictions and abstentions. This is achieved by extracting decision rules from LORE-based explanations, identifying influential features that contributed to the rejection decision, and presenting clear textual explanations of why a prediction was made or an instance was deferred. An additional technical contribution of L2loRe is its use of a **gamma-distributed sampling strategy** to optimize the rejection threshold τ_r, which ensures that the rejection mechanism adapts dynamically to varying levels of model uncertainty. In addition, L2loRe employs a **multi-metric evaluation framework** that uses rejection-aware performance measures to evaluate the trade-off between classification quality and abstention effectiveness.

L2loRe is a post-hoc, model-agnostic approach that can be applied to any pre-trained classifier, making it adaptable to different classification models and tabular datasets. The rejection threshold τ_r is chosen via a constrained optimization problem that maximizes classification performance while ensuring that the rejection rate remains below an upper bound r_{max} and that the proportion of false predictions does not exceed a defined threshold e_{max}. The method has been evaluated on binary classification (COMPAS, Adult, German Credit) and multiclass classification (Wine, Abalone, Student) datasets, where it demonstrated improved classification accuracy and rejection quality across different learning scenarios. This is discussed in more detail in the next section.

4.1.2 Results

Preliminary evaluation of L2loRe demonstrates its potential to improve model transparency, reliability, and human-AI interaction. One of the key benefits observed is its ability to provide interpretable rejection decisions, ensuring that when the model fails to make a prediction, the reason for this decision is clearly articulated. By integrating counterfactual-based confidence scoring, L2loRe enables users to understand model uncertainty in a structured and intuitive way, thereby increasing trust and transparency in AI decision-making.

Beyond explainability, L2loRe also improves classification performance by selectively rejecting uncertain instances, resulting in higher accuracy among accepted predictions. The model successfully adapts its rejection policy to the specific dataset by learning a distribution-aware abstention threshold, ensuring that rejection decisions remain contextually meaningful. These capabilities align with the TANGO project's goal of developing cognition-aware AI systems, as L2loRe provides a structured approach to hybrid AI-human decision-making, allowing AI models to defer uncertain cases to human experts when necessary. Future work will further refine L2loRe by exploring alternative counterfactual generation methods and improving distance metrics for confidence estimation, ensuring that its rejection mechanism remains robust across diverse application domains.

4.2 Further Work

In addition to work on L2loRe, we have further reviewed existing paradigms for hybrid decision-making. Punzi et al. [2024b] examines the current state of the art for machine learning-based systems that integrate a human component, focusing on the learning aspects. It identifies three distinct paradigms with increasing levels of human interaction: i) human oversight, ii) learning to abstain, and iii) learning together. In human oversight - the paradigm adopted by most Explainable AI - an AI-based system is supervised in its operations by a human. The human uses explanations to inspect and understand the AI, and decides whether to accept or modify its outcome. In abstention learning, the machine can refuse to process certain data, such as those in which it is less confident or, more recently, those in which the estimated human performance is higher than its own. In co-learning, humans interactively integrate their knowledge into the AI, for example by sharing explanations of decision rules.

Additional work has been on evaluating the effectiveness of learning delay strategies. To this end, in Palomba et al. [2025], we linked the potential outcomes framework for causal inference to learning delay. This allows them to identify the causal impact of introducing a deferral option on predictive accuracy, distinguishing between two scenarios. In the first, they assume that the deferral policy has access to both human and ML decisions, in which case the individual causal effects can be identified for deferred instances and aggregations of them. In the second, it has access only to human decisions, and causal effects can be estimated via regression discontinuity design.

Lastly, we have also explored the implementation of hybrid decision-making using large language models (LLMs). These have the innate ability to take instructions and make suggestions in textual form, making them a prime venue for synergistic machine learning. WP2 has explored data-driven fine-tuning strategies to improve the quality of textual guidance provided by LLMs in the context of medical decision-making, to optimize the quality of downstream human decisions [Banerjee et al., 2025].

4.3 Summary

Task T2.4 develops learning-to-defer and learning-to-reject strategies that enhance human-AI collaboration by ensuring that AI models recognize when to act autonomously and when to defer to human experts. The L2loRe framework supports interactivity (D1) by integrating counterfactual-based explanations into deferral policies that provide explicit justifications for abstention. In contrast to traditional confidence-based deferral, L2loRe enhances rationalization (D3) by providing rule-based explanations that mirror human reasoning, consistent with the feature-based explanations of D2.1.

Research in the context of T2.4 also enhances contextuality (D2) by adapting deferral strategies to task-specific needs and user expertise levels. The surveyed paradigms — human oversight, learning to abstain, and learning together — demonstrate different levels of human involvement, supporting the actionable explanation framework of D2.1. In addition, causal inference techniques improve AI's ability to reason about when abstention is beneficial, while LLMs generate natural language justifications that enhance human decision-making.

5 Learning Fair Models (T2.5)

In this section, we explore novel methods for fairness-aware machine learning to ensure that AI systems operate equitably across diverse populations. A key focus is on mitigating bias arising from incomplete demographic representation, where training data may systematically exclude certain groups, leading to spurious correlations and unfair decisions. Finally, we review other recent work and conclude with an analysis of how these contributions align with TANGO's broader goals of cognitively aware and fair AI-human collaboration.

5.1 Invisible Demographics

In Kehrenberg et al. [2024], we formalized the problem of systematic bias or dataset curation resulting in one or more groups having zero labeled data – *invisible demographics*.

When trained on diverse labelled data, machine learning models have proven themselves to be a powerful tool in all facets of society. However, due to budget limitations, deliberate or non-deliberate censorship, and other problems during data collection, certain groups may be under-represented in the labelled training set. Inspired by the idea of protected attributes from algorithmic fairness, we consider generalised secondary "attributes" that subdivide the classes into smaller partitions. We refer to the partitions defined by the combination of an attribute and a class label, or leaf nodes in aforementioned hierarchy, as *groups*. To characterise the problem, we introduce the concept of classes with *incomplete attribute support*. The representational bias in the training set can give rise to spurious correlations between the classes and the attributes which cause standard classification models to generalise poorly to unseen groups.

We illustrate our problem setup in Fig. 6, using Coloured MNIST digits as examples; here, the first level of the hierarchy captures digit class, the second level, colour. While the (unlabeled) deployment set contains all digit-colour combinations (or groups), half of these combinations are missing from the (labelled) training set. A classifier trained using only this labelled data would wrongly learn to classify 2's based on their being purple and 4's, based on their being green or red (instead of based on shape) and when deployed would perform no better than

random due to the new groups being coloured contrary to their class (relative to the training set).

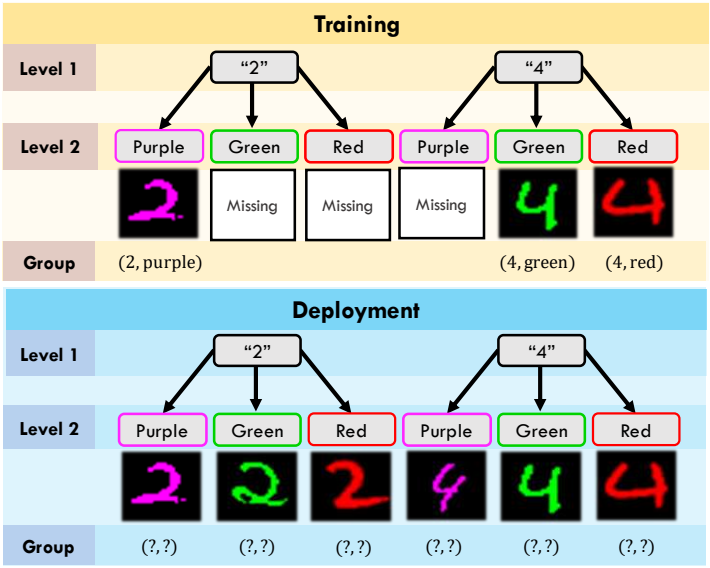

Figure 6: Illustration of our general problem setup of invisible demographics using neutral examples in terms of digits and colours.

5.1.1 Methodology

We cast the problem of learning an attribute-invariant representation as one of *support-matching* between a dataset that is *labelled* but has *incomplete* support over groups, and one, conversely, that has *complete* support over, but is *unlabeled*. The idea is to produce a representation that is invariant to this difference in support, and thus invariant to the attribute. However, it is easy to learn the wrong invariance if one is not careful. To measure the discrepancy in support between the two distributions, we adopt an adversarial approach, but one where the adversary is operating on small sets – which we call *bags* – instead of individual samples. These bags need to be balanced with respect to attribute a and class label y, *i.e.*, (a, y), such that we can interpret them as approximating groups as opposed to the joint probability distribution.

Specifically, our goal is to disentangle x into two subspaces: a subspace z, representing the class, and a subspace $\tilde{a}$, representing the attribute. For the problem to be well-posed, it is essential that bags differ only in terms of which groups are present. We thus sample the bags according to the following set of rules :

1. Bags of the deployment set are sampled so as to be approximately balanced with respect to a and y (all combinations of a and y should appear in equal number).
2. For bags from the training set, all possible values of y should appear with equal frequency. Without this constraint, there is the risk of y being encoded in $\tilde{a}$.
3. Bags of the training set should furthermore exhibit equal representation of each attribute within classes so long as rule 2 is not violated. For classes that do not have complete $\mathcal{A}$-support, the missing combinations of (a, y) need to be substituted with a sample from the same class.

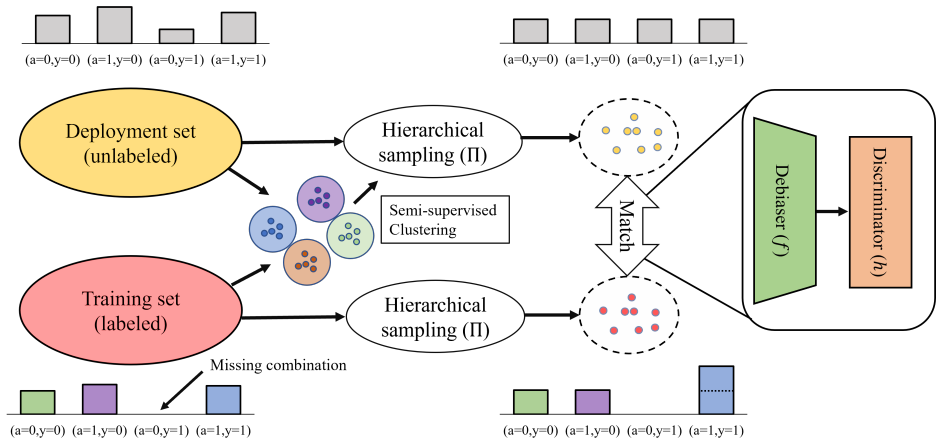

Figure 7: *Visualisation of our support-matching pipeline.* Bags are sampled from the training and deployment sets using the hierarchical sampling procedure. The debiaser is adversarially trained to produce representations from which the source dataset cannot be reliably inferred by the discriminator.

To summarise, our method can be broken down into two steps: 1) sample perfect bags (bags in which groups appear in equal proportions) from an unlabelled deployment set; and 2) produce disentangled representations using the perfect bags via adversarial support-matching. For the first step, the data points in the deployment set are labelled with cluster assignments generated by clustering algorithm, so that we can form perfect bags by sampling all clusters at equal rates. We note that we do *not* have to associate the clusters with specific a or y labels as the labels are not directly used for supervision. Balancing bags based on clusters instead of the true labels introduces an error, which we provide an error bound. For the second step, the model is composed of three core modules: 1) two encoder functions, f (which we refer to as the "debiaser") and t, which share weights and map x to z and $\tilde{a}$, respectively; 2) a decoder function that learns to invert f and t; and 3) a discriminator function that predicts which dataset a bag of samples was sampled from.

We present in Fig. 7 a high-level, visual overview of our support-matching pipeline capturing both the hierarchical sampling and modelling aspects that characterise our proposed method.

5.1.2 Results

Within Kehrenberg et al. [2024], we provide theoretical results concerning the validity of our support-matching objective. Hereby, we prove that the "debiaser" f satisfying the objective is invariant to a, at least in those cases where the class does not have full a-support (which is exactly the case with *invisible demographics*).

We also empirically validate our framework on a range of datasets: Coloured MNIST as a synthetic image dataset (with both binary and multiclass variants), CelebA and NICO++ as real-world image datasets (binary and multiclass, respectively), and ACS as a tabular dataset showcasing generality. We empirically show that it is possible to maintain high performance for samples with classes with incomplete support in a robust fashion.

Figure 8: Example sample-wise attention maps for bags of CelebA (left) and Coloured MNIST (right) images sampled from a balanced deployment set.

Fig. 8 shows attention maps for bags from the deployment set. The attention weight assigned to each sample is proportional to the lightness of its frame, with black signifying a weight of 0, white a weight of 1. The training set is biased according to the *attribute bias* setting where for CelebA "smiling females" constitute the missing group and for Coloured MNIST purple fours constitute the missing group. The attention weights are used during the discriminator's aggregation step to compute a weighted sum over the bag. Those samples belonging to the missing group are assigned the highest weight as they signal from which dataset (training vs. deployment) the bag containing them was drawn from. We observe that the model pays special attention to those samples that are not included in the training set.

5.2 Further Work

In other work, we have continued to focus on advances in evaluation and monitoring. In an effort to advance evaluation and monitoring in continuous learning, we developed ConCon, a dataset designed to systematically analyze the impact of confounders in continuous learning settings [Busch et al., 2024]. In such scenarios, models often develop a bias toward spurious correlations, leading to unfair or unreliable decisions, especially as confounders change over time. ConCon provides a controlled setting to study how confounders affect model behavior, ensuring that continuous learning methods can be evaluated not only for catastrophic forgetting, but also for their ability to disentangle true causal factors from spurious ones. Empirical results show that even continuous learning methods that successfully prevent forgetting struggle with confounding, highlighting the need for novel approaches that explicitly address spurious correlations in dynamic learning environments.

In addition, we have developed an innovative framework for probabilistic formal verification of machine learning models [Morettin et al., 2024]. This framework is very general because it can assess whether models satisfy a variety of criteria, such as fairness constraints, with a sufficiently high probability, given a distribution of input data. It supports hybrid domains of discrete and continuous variables by using Weighted Model Integration (WMI), a sophisticated method for probabilistic inference. It has been successfully demonstrated on a variety of tasks, with one notable application being the monitoring of demographic parity, which requires that a predictor provide members of a protected group (e.g., women) with an equal chance of a favorable outcome (e.g., being selected for a job) compared to the broader population - a key fairness constraint.

With Lenders et al. [2024], we have developed a system that mitigates some of the fairness problems of existing abstention classifiers. Our solution is to train a rule-based system to highlight and reject unfair decisions in conjunction with a confidence-based abstention approach. This approach achieves the same accuracy as regular abstention classifiers, while providing an interpretable method for exploring rejection instances by fair rules, and providing

stakeholders with useful suggestions for possible corrective actions. Finally, with Ng et al. [2024] we have developed an active learning framework for addressing bias in the dataset curation process, first in a simplified setting where sensitive information is available in the data and can be acquired from human annotators.

5.3 Summary

Task T2.5 develops fairness-aware machine learning methods that mitigate biases due to incomplete demographic representation, ensuring transparent and equitable AI decision-making. The Invisible Demographics framework addresses systematic underrepresentation using adversarial support matching, enforcing contextuality (D2) by adapting fairness interventions to dataset-specific biases, and rationalization (D3) by ensuring that bias corrections are interpretable.

In addition, formal fairness verification techniques enhance interactivity (D1) by providing transparent fairness judgments rather than treating bias mitigation as an opaque process. Fairness-aware abstinence classifiers and active learning for bias mitigation are consistent with the actionable explanation framework of D2.1, producing explicit fairness justifications that refine model behavior.

6 Ethical Considerations

The development of human-machine learning systems within TANGO entails significant ethical responsibilities, in particular to ensure that such systems operate fairly, respect individual privacy, preserve human autonomy, and maintain high levels of transparency. As TANGO progresses towards cognition-aware explanations and hybrid decision-making, these ethical pillars guide the design and evaluation of neuro-symbolic models, interactive learning strategies, and fairness-conscious techniques. In this section, we highlight key ethical challenges related to fairness, privacy, persuasion, and transparency, and illustrate how they apply to the AI models and methods developed in TANGO.

6.1 Fairness

Fairness is critical to ensuring that AI systems do not propagate or reinforce societal biases. The Invisible Demographics work in TANGO directly addresses the risks of biased learning caused by incomplete demographic representation in training data. The adversarial support-matching techniques applied in this context aim to mitigate such biases by aligning attribute distributions across training and deployment phases. This is particularly important for concept-based models such as Pix2Code and SCoBots, whose representations and explanations rely on the clarity and correctness of the underlying learned concepts. If biased data informs these concepts, the resulting explanations could reinforce harmful stereotypes.

In addition, TANGO's neuro-symbolic reasoning approaches (e.g., `bears`) aim to make models aware of their own reasoning shortcuts, thereby revealing the conditions under which their decisions might be influenced by biased or spurious correlations. This self-awareness does not automatically guarantee fairness, but it does enable both the system and human supervisors to detect and correct potential biases in real time. However, explanation fairness—ensuring

that explanations do not disproportionately misrepresent or obscure reasoning paths for certain groups—remains an open challenge that TANGO continues to address.

6.2 Privacy

Privacy risks exist throughout the AI lifecycle, from data collection to training, inference, and explanation. TANGO's privacy-aware approaches, particularly its exploration of privacy-preserving learning strategies, seek to minimize these risks without compromising model performance or explainability. This is particularly challenging for neuro-symbolic approaches such as Pix2Code, which explicitly surface structured object attributes and relational concepts. While these are beneficial for interpretability, they can also expose sensitive attribute correlations if not handled carefully.

In addition, Learning to Reject via Local Rule-based Explanations (L2loRe) raises nuanced privacy concerns when explanations of rejected instances reveal characteristics of the training data. The balance between transparency (providing interpretable rejection rules) and privacy (protecting sensitive data elements) becomes particularly delicate in these cases. Future privacy work in TANGO will focus on developing differential privacy-inspired techniques for both predictions and explanations, ensuring that transparency does not come at the expense of individual privacy.

6.3 Persuasion

As TANGO systems increasingly engage in interactive explanations, the persuasive potential of AI-generated justifications must be critically managed. Models such as SCoBots, which allow users to inspect multi-level decision processes and intervene in learned policies, offer significant transparency but also pose a risk: human users may overly trust these explanations, especially if the explanations convey a false sense of certainty or objectivity.

This risk is compounded when explanations are generated in natural language by large language models (LLMs), as TANGO also explores. Such explanations may appear convincing simply because of their fluency and coherence, even if the underlying reasoning is flawed. To mitigate this, TANGO emphasizes the importance of calibrated explanations - where the system actively communicates its uncertainty - and dual accountability, where both human and AI contributions to decision-making are explicitly documented.

6.4 Transparency

As highlighted in previous reports, transparency is fundamental to ethical AI, especially in collaborative human-machine learning systems such as those envisioned in TANGO. The approaches presented in this deliverable enhance transparency in multiple ways, ensuring that AI models are not only interpretable but also allow for human oversight and intervention.

The Pix2Code framework improves transparency by synthesizing explicit programmatic representations of visual concepts, allowing humans to inspect, understand, and even modify these representations if necessary. Similarly, SCoBots structure decision-making through interpretable concept bottlenecks, where each stage of processing remains accessible for inspection, questioning, and correction, supporting both explainability and collaborative learning. The `bears` framework enhances transparency by explicitly identifying concepts vulnerable to

reasoning shortcuts, associating confidence scores with concept reliability, and enabling users to assess when and where to trust the model. Additionally, L2loRe embeds transparency into rejection decisions by providing local, rule-based explanations for every abstained prediction, ensuring that rejection policies remain understandable and context-specific, rather than functioning as black-box decisions.

Transparency in TANGO is not just about revealing what the AI did—it extends to why it did it, under what conditions it might fail, and how humans can intervene. This proactive and collaborative form of transparency aligns with TANGO's broader vision of cognition-aware and human-centered AI. It is, however, important to note that transparency alone is insufficient if the revealed information is not cognitively accessible to human users. Therefore, TANGO prioritizes explanations that match the expertise and mental models of different users (*e.g.*, translating complex programmatic explanations in Pix2Code to natural language statements), ensuring that transparency benefits technical developers, domain experts, and lay users alike.

7 Conclusion

This deliverable D2.2 presents the initial version of methods for synergistic human-machine learning (HML) developed within TANGO's Work Package 2 (WP2). It builds on the theoretical foundations established in WP1, incorporating its principles of interactivity, contextuality, and rationalization into practical machine learning architectures. Additionally, it extends the methodological groundwork laid out in D2.1, where different types of machine explanations were mapped to actionable categories that facilitate human-AI collaboration. Throughout this deliverable, we have detailed progress across T2.2–T2.5, covering advancements in cognition-aware explanations, neuro-symbolic models, interactive learning strategies, hybrid decision-making, and fairness-aware AI systems.

Each task contributes to TANGO's strategic objective SO2.1, which aims to develop paradigms and algorithms that ensure AI systems enhance, rather than replace, human expertise. Task T2.2 focuses on understandable and reliable neuro-symbolic models, ensuring that AI systems provide structured, inspectable, and revisable reasoning processes. Task T2.3 advances synergistic interactive machine learning, enabling models to dynamically adapt based on human feedback and intervention. Task T2.4 introduces learning-to-reject and learning-to-defer strategies, ensuring that AI can identify uncertainty, abstain when necessary, and complement human decision-making. Finally, Task T2.5 explores fairness-aware machine learning, mitigating biases and ensuring AI models function equitably across diverse populations.

Beyond these technical contributions, this deliverable provides a cohesive and methodologically detailed perspective on WP2's progress, complementing the M18 Technical Report, which overviewed individual research contributions. Moving forward, WP2 will continue to refine these approaches, emphasizing explainability, reliability, and fairness as central pillars of trustworthy and effective AI-human collaboration.

Overall, the methodological advancements in WP2, as detailed in this deliverable, provide a strong foundation for the hybrid decision-making frameworks developed in WP3. WP2's contributions to cognition-aware explanations, neuro-symbolic reasoning, and interactive learning directly support WP3's goal of enabling AI-assisted decision-making across various domains, including individual, shared, team-based, and policy-making contexts. Specifically,

WP2's work on interpretable neuro-symbolic models (T2.2) ensures that AI decisions remain transparent and revisable, a crucial requirement for trustworthy hybrid decision-making in WP3. Additionally, WP2's interactive learning strategies (T2.3) and learning-to-defer mechanisms (T2.4) align with WP3's objective of optimally distributing decision-making responsibilities between humans and AI. WP2 also introduces fairness-aware AI methodologies (T2.5), which contribute to WP3's focus on ensuring ethical, unbiased, and accountable AI-driven decision support. These synergies ensure that WP3 can build upon WP2's advancements, integrating explainability, adaptivity, and fairness into hybrid decision-making systems that effectively complement human expertise.

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 (NeurIPS)*, 2022. URL https://openreview.net/forum?id=A4lAgAESt2.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://openreview.net/forum?id=H1bip_b_bH.

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

Matej Balog, Alexander L. Gaunt, Marc Brockschmidt, Sebastian Nowozin, and Daniel Tarlow. Deepcoder: Learning to write programs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=ByldLrqlx.

Debodeep Banerjee, Stefano Teso, Burcu Sayin, and Andrea Passerini. Learning to guide human decision makers with vision-language models. *arXiv*, 2025. URL https://arxiv.org/abs/2403.16501.

Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. Uncertainty matters: Stable conclusions under unstable assessment of fairness results. In *Proceedings of Machine Learning Research (PMLR)*, 2024. URL https://proceedings.mlr.press/v238/barrainkua24a.html.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents (extended abstract). In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2015. URL https://openreview.net/forum?id=r1EjHXGdWS.

Armin Biere, Marijn Heule, and Hans van Maaren. *Handbook of satisfiability*, volume 185. IOS press, 2009.

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*, 2024. URL https://arxiv.org/abs/2402.06434.

Diego Calanzone, Stefano Teso, and Antonio Vergari. Logically consistent language models via neuro-symbolic integration. *arXiv*, 2024. URL https://arxiv.org/abs/2409.13724.

Diego Calanzone, Stefano Teso, and Antonio Vergari. Logically consistent language models via neuro-symbolic integration. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=7PGluppo4k.

José Cambronero, Sumit Gulwani, Vu Le, Daniel Perelman, Arjun Radhakrishna, Clint Simon, and Ashish Tiwari. Flashfill++: Scaling programming by example by cutting to the chase. In *Principles of Programming Languages (POPL)*, 2023. URL https://www.microsoft.com/en-us/research/publication/flashfill-scaling-programming-by-example-by-cutting-to-the-chase/.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, 2013.

Quentin Delfosse, Sebastian Sztwiertnia, Mark Rothermel, Wolfgang Stammer, and Kristian Kersting. Interpretable concept bottlenecks to align reinforcement learning agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=ZC0PSk6Mc6.

Lauro Langosco di Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022. URL https://openreview.net/forum?id=TLhxAoJrg88.

Luca Di Liello, Pierfrancesco Ardino, Jacopo Gobbi, Paolo Morettin, Stefano Teso, and Andrea Passerini. Efficient generation of structured objects with constrained adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://openreview.net/forum?id=GgCOydV90FI.

A. Dix, T. Turchi, B. Wilson, A. Monreale, and M. Roach. Talking back – human input and explanations to interactive ai systems. In *Workshop on Adaptive XAI @ International Conference on Intelligent User Interfaces (IUI)*, 2025.

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke B. Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B. Tenenbaum. Dreamcoder: bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the International Conference on Programming Language Design and Implementation (PLDI)*, 2021. URL https://api.semanticscholar.org/CorpusID:235474001.

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

Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://openreview.net/forum?id=GgzscNf2k7l.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv*, 2021. URL https://arxiv.org/abs/2103.03874.

Nick Hoernle, Rafael Michael Karampatsis, Vaishak Belle, and Kobi Gal. Multiplexnet: Towards

fully satisfied logical constraints in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. URL https://openreview.net/forum?id=NhhdT4PpTa.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Thomas Kehrenberg, Myles Bartlett, Viktoriia Sharmanska, and Novi Quadrianto. Addressing attribute bias with adversarial support-matching. In *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://openreview.net/forum?id=JYbnJ92TJf.

Junkyung Kim, Matthew Ricci, and Thomas Serre. Not-so-clevr: learning same–different relations strains feedforward neural networks. *Interface Focus*, 8, 2018. URL https://api.semanticscholar.org/CorpusID:49223699.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Yann LeCun. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Daphne Lenders, Andrea Pugnana, Roberto Pellungrini, Toon Calders, Dino Pedreschi, and Fosca Giannotti. Interpretable and fair mechanisms for abstaining classifiers. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2024.

Zenan Li, Zehua Liu, Yuan Yao, Jingwei Xu, Taolue Chen, Xiaoxing Ma, L Jian, et al. Learning with logical constraints but without shortcut satisfaction. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural Probabilistic Logic Programming. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Emanuele Marconato, Gianpaolo Bontempo, Elisa Ficarra, Simone Calderara, Andrea Passerini, and Stefano Teso. Neuro-symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal. In *Proceedings of the International Conference on Machine Learning (ICML)*, 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 *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 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 (NeurIPS)*, 2024b.

Paolo Morettin, Andrea Passerini, and Roberto Sebastiani. Probabilistic ml verification via weighted model integration. *arXiv*, 2024. URL https://arxiv.org/abs/2402.04892.

Heimo Müller and Andreas Holzinger. Kandinsky patterns. *Artificial Intelligence*, 300:103546, 2021. URL https://www.sciencedirect.com/science/article/pii/S0004370221000977.

Yeat Jeng Ng, Viktoriia Sharmanska, Thomas Kehrenberg, Anastasia Pentina, and Novi Quadrianto. Disagreement-based Active Learning for Robustness Against Subpopulation Shifts. 9 2024. URL https://sussex.figshare.com/articles/conference_contribution/Disagreement-based_Active_Learning_for_Robustness_Against_Subpopulation_Shifts/27094891.

Filippo Palomba, Andrea Pugnana, Jose Manuel Alvarez, and Salvatore Ruggieri. A causal framework for evaluating deferring systems. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. URL https://openreview.net/forum?id=mkkFubLdNW.

Clara Punzi, Roberto Pellungrini, and Fosca Giannotti. L2lore: a method for explaining the reject option. In *Proceedings of the Internation Conference on Discovery Science (DS)*, 2024a.

Clara Punzi, Roberto Pellungrini, Mattia Setzu, Fosca Giannotti, and Dino Pedreschi. Ai, meet human: Learning paradigms for hybrid decision making systems. *arXiv*, 2024b. URL https://arxiv.org/abs/2402.06287.

David Raposo, Samuel Ritter, Adam Santoro, Greg Wayne, Théophane Weber, Matthew M. Botvinick, H. V. Hasselt, and Francis Song. Synthetic returns for long-term credit assignment. *arXiv*, 2021.

Yoshihide Sawada and Keigo Nakamura. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10:41758–41765, 2022.

Gautam Singh, Yeongbin Kim, and Sungjin Ahn. Neural systematic binder. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Sebastian Stabinger, David Peer, Justus Piater, and Antonio Rodríguez-Sánchez. Evaluating the progress of deep learning for visual relational concepts. *Journal of Vision*, 21(11):8–8, 10 2021. URL https://doi.org/10.1167/jov.21.11.8.

Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Wolfgang Stammer, Antonia Wüst, David Steinmann, and Kristian Kersting. Neural concept binder. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=ypPzyflbYs.

David Steinmann, Wolfgang Stammer, Felix Friedrich, and Kristian Kersting. Learning to intervene on concept bottlenecks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=gEbl6XNLK6.

Elias Stengel-Eskin, Archiki Prasad, and Mohit Bansal. Regal: Refactoring programs to discover generalizable abstractions. *arXiv*, 2024. URL https://arxiv.org/abs/2401.16467.

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

Dweep Trivedi, Jesse Zhang, Shao-Hua Sun, and Joseph J Lim. Learning to synthesize programs as interpretable and generalizable policies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://openreview.net/forum?id=wP9twkexC3V.

Ramakrishna Vedantam, Arthur Szlam, Maximillian Nickel, Ari Morcos, and Brenden M Lake. Curi: A benchmark for productive concept learning under uncertainty. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/vedantam21a.html.

Arne Vermeulen, Robin Manhaeve, and Giuseppe Marra. An experimental overview of neural-symbolic systems. In *Proceedings of the International Conference on Inductive Logic Programming (ILP)*, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. California Institute of Technology, Juli 2011.

Kaifu Wang, Efi Tsamoura, and Dan Roth. On learning latent models with multi-instance weak supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Moritz Willig, Tim Tobiasch, Florian Peter Busch, Jonas Seng, Devendra Singh Dhami, and Kristian Kersting. Systems with switching causal relations: A meta-causal perspective. In *Workshop on Causal Representation Learning Workshop @ NeurIPS*, 2024. URL https://openreview.net/forum?id=aMuFZK7Tl8.

Antonia Wüst, Wolfgang Stammer, Quentin Delfosse, Devendra Singh Dhami, and Kristian Kersting. Pix2code: Learning to compose neural visual concepts as programs. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024a. URL https://openreview.net/forum?id=EE4ikEQnOT.

Antonia Wüst, Tim Tobiasch, Lukas Helff, Devendra Singh Dhami, Constantin A. Rothkopf, and Kristian Kersting. Bongard in wonderland: Visual puzzles that still make AI go mad? In *Workshop on System-2 Reasoning at Scale @ NeurIPS*, 2024b. URL https://openreview.net/forum?id=4Yv9tFHDwX.

Xuan Xie, Kristian Kersting, and Daniel Neider. Neuro-symbolic verification of deep neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok,

Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv*, 2023. URL https://arxiv.org/abs/2309.12284.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv*, 2023. URL https://arxiv.org/abs/2309.05653.

Çağlar Aytekin. Neural networks are decision trees. *arXiv*, 2022. URL https://arxiv.org/abs/2210.05189.