DARE: An Explainable AI-visualization Framework for Ill-defined Decision Making

Angelos Chatzimparmpas and Evanthia Dimara

***Abstract*—Real-world decision making often unfolds in fluid, uncertain, and ill-defined contexts where objectives shift, data are incomplete, and non-quantifiable factors such as social values, ethics, and institutional constraints play critical roles. Conventional AI and decision-support systems assume fixed criteria and stable data, leaving these contexts underserved. Building on an interdisciplinary definition of decision making attentive to its ill-defined forms, we introduce DARE, an explainable AI and visualization framework that complements the FAIR data principles with the DARE principles: *Deliberation*, *Agency*, *Resilience*, and *Empathy*, which emphasize dialogue, human control, adaptability, and human sensitivity in design. DARE conceptualizes decision making as an iterative alignment of human-defined criteria with algorithmic representations through which decision structure gradually emerges. We revisit existing AI paradigms through this lens and illustrate how weak supervision and concept-based modeling exemplify this process by connecting heuristic human reasoning to interpretable model concepts. Input visualization serves as the expressive layer that captures evolving, qualitative, and uncertain reasoning through interaction, allowing humans to externalize and refine decision logic before formalization. Explainability in DARE arises not from post-hoc justification but from the continuous visibility of how human and algorithmic reasoning co-develop. Uncertainty is treated as an inherent dimension of deliberation, something to represent, navigate, and learn from within the decision process, while human and algorithmic heuristics are regarded not as truths or biases but as evolving hypotheses to examine and refine through interaction. Together, these elements support human–AI decision making that remains transparent, adaptable, and grounded in human judgment across value-laden and ill-defined contexts.**

***Index Terms*—Explainable AI (XAI), Human-centered AI (HCAI), Responsible AI, Visualization, Visual Analytics, Decision Making, Ill-defined Decision Making, Multi-Criteria Decision Making (MCDM), Hybrid Intelligence, Artificial Intelligence (AI), Machine Learning (ML)**

I. Introduction

ILL-DEFINED, high-stakes decisions, such as coordinating city-wide vaccination strategies under evolving pandemic constraints or allocating resources in response to a volatile humanitarian crisis, often demand more than blind trust in automated outputs. Yet mainstream AI solutions have historically favored fully automated models, sidelining human expertise in the name of efficiency. As noted in Shneiderman's Human-centered AI (HCAI) framework [1], successful systems empower rather than replace people by ensuring meaningful human control over algorithmic processes. The European Commission's proposed AI Act [2] translates this principle into binding regulatory requirements, specifying human oversight and transparency as mandatory for high-risk AI systems. Despite these initiatives, real-world practice still relies heavily on black-box approaches [3], especially in tasks that remain inherently ill-defined [4], where objectives may shift midstream, data are fragmented or incomplete, and organizational or societal considerations outweigh data-driven logic [5], [6].

Yet, much of today's AI and visualization infrastructure remains anchored in assumptions of stability. State-of-the-art AI pipelines, including deep neural networks and reinforcement learning [7], typically presuppose fixed objectives and abundant labeled data. Visual Analytics (VA) frameworks for human-in-the-loop ML largely build on this premise, supporting well-bounded analytic tasks [8] or emphasizing post-hoc inspection of trained models [9]. Even decision-support tools grounded in Multi-Criteria Decision Making (MCDM) [10]–[12] assume that objectives, criteria, and trade-offs can be specified upfront. These assumptions break down in practice when legal interpretations shift, organizational priorities change, or social values reframe decision goals. As a result, supporting decision makers in contexts where both evidence and evaluative frames evolve over time remains unaddressed.

In response to calls to address the challenges of AI-assisted decision making [13]–[16], we propose an Explainable AI (XAI) and visualization framework designed to keep humans in control of ill-defined decision processes. To ground system design, we revisit foundational assumptions about what constitutes a decision and introduce a working definition of ill-defined decision making. Rather than treating uncertainty as a limitation, our approach embraces it as a structural feature, proposing systems that help surface, negotiate, and iteratively refine decision criteria. We introduce the DARE principles: ***D**eliberation* (dialogue between humans and models), ***A**gency* (meaningful human control over decision logic), ***R**esilience* (adaptability to evolving data and goals), and ***E**mpathy* (sensitivity to people and contexts affected by decisions) as conceptual requirements for decision-support systems. The DARE framework connects these principles through a continuous loop between human and algorithmic reasoning, where decision criteria are expressed, tested, and refined through interpretable representations. Techniques such as weak supervision [17], concept-based modeling [18], and input visualization [19] exemplify how this loop can be realized in practice, linking heuristic human reasoning with transparent computational models. In doing so, DARE moves beyond static human-in-the-loop pipelines toward genuinely collaborative reasoning that enables decision logic to emerge and evolve in ill-defined contexts.

Angelos Chatzimparmpas and Evanthia Dimara are with the Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, NL. E-mail: {a.chatzimparmpas@uu.nl, evanthia.dimara@gmail.com}

II. Related Work

We review how decision-focused visualizations (II-A) and AI-visualization frameworks (II-B) support human decisions.

A. Decision-focused Visualization Tools

While many visualizations can indirectly support decision making by helping users understand relevant data, prior work has argued for visualization systems explicitly designed around decision tasks [4], [6], [20]–[22]. These decision-focused tools differ from general analytic systems in that they incorporate explicit support for *choice*, rather than stopping at exploration or insight generation [23]. Typical examples include visualizations for multi-attribute comparison [24], preference elicitation [10], ranking [11], and outcome-based trade-off analysis [25].

Most such systems build on MCDM techniques (see [23] for a review). Tools such as ValueCharts [26] and LineUp [11] visualize alternatives across multiple attributes and let users adjust weights to observe ranking changes. These approaches are effective for structured decision problems where criteria and objectives are specified in advance. In *ill-defined* contexts, however, criteria are often incomplete, contested, or evolving and must be discovered or revised during the decision process itself, a form of support most existing decision-support visualizations provide only to a limited extent [6], [27]–[29].

Despite the complexity of real-world decision making, algorithmic support in decision-focused visualization remains limited, typically relying on summation-based ranking or simple additive weighting, with only a few systems integrating AI methods (5 out of 88, in a recent review [23]). For example, FairSight [30] targets high-stakes ranking scenarios using Logistic Regression, Support Vector Machines, and RankSVM to detect bias. Andrienko et al. [31] apply genetic algorithms to evacuation planning, while Müller et al. [32] use Bayesian networks for cancer therapy decisions. Guo et al. [33] assist marketing decisions by predicting future behaviors with deep recurrent neural networks, and Podium [34] enables partial rankings and infers preference weights via Ranking SVM. While these systems demonstrate productive combinations of visualization and computation, they also assume fixed or early-specified decision criteria, offering limited support for *ill-defined* decisions where criteria emerge or shift mid-process. The use of large language models (e.g., ChatGPT-4o [35], DeepSeek-R1 [36]) or vision-language models (e.g., CLIP [37], Qwen2.5-VL [38]) remains largely unexplored for integrating adaptive AI guidance with fluid, user-defined objectives.

B. AI-visualization Frameworks

Prior VA frameworks for XAI have been developed to support global and local model interpretability [39], [40]. Angelini et al. [39] present a VA framework that allows ML and AI experts to explore and refine Partial Dependence Plot (PDP) computations [41], enabling deeper global understanding of feature dependencies in predictive models. Similarly, ExpLIMEable [40] extends LIME [42] through an interactive dashboard that lets researchers, model evaluators, and practitioners (e.g., medical imaging specialists) adjust parameters and examine their effect on local explanations. However, these frameworks are tightly coupled to specific techniques such as LIME or PDP, limiting their generalizability across XAI approaches and their usefulness for human-driven, ill-defined decision tasks.

Broader frameworks for expert users provide extensive support for explainability by integrating multiple XAI techniques rather than focusing solely on PDPs or LIME. The explAIner framework [43] facilitates iterative understanding, diagnosis, and refinement of AI models by integrating various explainability techniques into an interactive VA system for domain experts and model developers. VIS4ML [44] uses an ontology-driven approach to categorize the role of VA in ML workflows and identify where interactive visualization can enhance interpretability for expert users, such as ML engineers and model developers. None of these frameworks incorporates a human decision making task; instead, they focus on interpretability rather than guiding non-experts' decisions by aligning humans with AI decision criteria.

Notably, XEdgeAI [45] targets industrial workers, such as field engineers, who make real-time judgments on images to refine ML models via interactive exploration and rate text-based explanations. Yet, its focus is limited to defect analysis in visual quality inspection for manufacturing and maintenance. The framework is designed for supervised classification problems and relies on domain experts with AI expertise to augment data with time-consuming manual annotations. Gathani et al. [46] observed that such experts, data scientists, and business analysts are costly to employ and often scarce or unavailable when decision makers need timely support.

Many visualization frameworks support analytical tasks such as model refinement, fairness evaluation, or discourse analysis [9], [47]–[50]. For instance, Sacha et al. [47] enable users to refine ML models through interactive exploration. This framework supports exploratory analysis rather than decision support, as it does not address ill-defined tasks, weigh trade-offs, or help define decision criteria. VisArgue [48] integrates computational linguistics and VA to analyze argumentation structures and discourse in deliberative discussions, helping political and social scientists understand how arguments evolve. While it supports deliberation, it does not help decision makers align human and AI decision criteria. FairCompass [9] focuses on fairness-aware auditing of AI models, guiding users through structured fairness assessments using subgroup discovery and decision tree-based analysis. Although it allows users to evaluate model bias, it does not support evolving or ambiguous decision contexts where human objectives remain uncertain. Lotse [49] aids data analysts and ML experts by optimizing workflows and improving sensemaking in VA environments, but it likewise does not address ill-defined decision making or the comparison of emerging criteria. Although AI is used in high-risk domains such as clinical decision support and hiring [51], [52], to the best of our knowledge, no existing framework combines AI-supported human decision making for ill-defined, evolving tasks, as most AI-visualization frameworks remain focused on data exploration [53], progressive knowledge generation [54], [55], or agent-based approaches assuming structured analytic environments [50].

III. Conceptual Scope: Grounding Decision-Oriented AI Design

This section outlines foundations for real-world decision making support. It defines ill-defined decisions (III-A) and guiding principles (III-B) and reviews AI paradigms through this lens (III-C), setting the stage for the DARE framework.

A. Which Decision Making?

Decision making has been characterized as a task defined by the explicit intention to select one or more options from a set of alternatives, based on decision criteria [4], [23], [27]. In this sense, *choice* is the minimal condition distinguishing decision making from activities like judgment, which involves assessing information without committing to an option [22], [56]; sensemaking, which focuses on constructing cognitive frameworks and representations from data [4], [57]; or analysis aimed at exploring information spaces rather than directly informing a choice [4], [20], [23]. Beyond the act of choosing, decision making (i) implies permanent or temporary loss of unselected alternatives, (ii) is future-oriented, meaning outcomes follow from choice, (iii) involves overt or covert actions, and (iv) typically carries some degree of personal stake or responsibility for the outcomes [22]. The process surrounding such choice can unfold at various levels of abstraction: for instance, at a high level, through iterative stages leading to choice, namely intelligence (gathering and structuring information) and design (creating alternatives) [6], [23], [58]; and at a lower level, through micro-operations that create, filter, and select among candidate options [20].

Building on this definition of decision making, we focus on a subcategory we term ill-defined decision making, where the structure of the decision problem is not fully specified. In such settings, objectives, alternatives, constraints, criteria, or evaluation standards may be fluid, ambiguous, absent, or evolve during the decision process [6], [27]. Ill-defined decisions therefore require active problem formulation, in which decision makers identify, construct, or revise key elements of the decision as they proceed. They often unfold under shifting conditions, incomplete information, limited procedural clarity, and the need to integrate subjective insights, qualitative considerations, or emergent external influences [6], [59]. This is characteristic of many real-world decisions, which often begin not as clearly bounded choice problems but as open, complex, and partially structured situations [58], [60], [61]. Although such decisions often involve uncertainty and incomplete information, they are not defined by uncertainty alone; rather, they differ from well-defined decisions under uncertainty, where the decision structure is fixed but outcomes remain unknown [59]. This class of decisions has been largely overlooked in visualization research [29]. We extend our working definition below through contrasts with adjacent decision paradigms.

Automated Decision Making: Automated decision making has deep roots in computer science, where more automation was traditionally equated with better performance. Our notion of ill-defined decision making builds on HCAI frameworks, which advocate for retaining human agency, especially in complex or ethically sensitive contexts [1], [2]. Visualization, too, exists precisely when human judgment is indispensable; as Munzner notes, if fully automatic solutions suffice, "there is no need for human judgment, and thus no need for you to design a vis tool" [62]. Yet even in human-centered venues like VIS, the term "decision" (e.g., in Area 6: Analytics & Decisions [63]) is sometimes ambiguous, whether it refers to a human process or an algorithmic operation. Our work emphasizes the former, focusing on ill-defined decisions that resist full automation by nature and demand interpretive, iterative human involvement. This does not preclude the use of automation: such decisions can still benefit from selective machine assistance. Yet identifying subtasks [20] where automation can support but not replace human decisions remains an open challenge.

Data-Driven Decision Making & Decision Support Systems: These systems can be seen as early automation tools for managerial decision processes, relying on structured data and rules to assist or partially automate choices. Our notion of ill-defined decision making departs from the assumptions of data-driven decision making (DDDM), where choices are guided by quantitative analysis supported by dashboards, predefined metrics, and Key Performance Indicators (KPIs) [64]. Rooted in early Decision Support Systems (DSS) research [65], DDDM depends on ETL pipelines, rule-based logic, and scenario modeling to benchmark performance and forecast outcomes, while AI-based DSS incorporate machine learning, optimization, and knowledge-based methods to support decision making [66]. Business professionals use what-if analysis to adjust parameters and assess effects on target KPIs, yet even in such structured settings decision makers often struggle to define objectives, maintain metrics, or adapt models to change [6], [46]. In ill-defined contexts, these challenges intensify: assumptions stay implicit, criteria evolve, and constraints emerge dynamically. Such decisions still benefit from structured data and metrics, but they elude purely data-driven solutions.

Multi-Criteria Decision Making & Recommender Systems: Both MCDM techniques and recommender systems automate parts of the decision pipeline. As discussed earlier (Sec. II-A), MCDM provides formal methods for ranking alternatives using explicit criteria and weighted scoring models [67]–[70]. Recommender systems [71]–[73] similarly aim to prioritize options but typically infer preferences from behavioral data rather than structured input. Both families support structured decision processes yet assume stable objectives and quantifiable trade-offs, conditions rarely met in ill-defined contexts. Nevertheless, they remain useful for specific phases, such as when criteria stabilize or when surfacing and aligning stakeholder goals.

Heuristic, Strategic, & Naturalistic Decision Making: Several decision paradigms operate outside fully structured models [74] yet remain distinct from ill-defined decision making. These include heuristic traditions as reviewed within behavioral decision research and decision science [75], as well as related perspectives on strategic and naturalistic decision making. Heuristic decision making, unlike rational economic models based on utility maximization and formal logic [76], emphasizes cognitive shortcuts and bounded rationality [77], [78]. Strategic decision making concerns consequential, infrequent choices under uncertainty, influenced by organizational dynamics,

political negotiation, and ambiguity [79], and has also been formalized in game-theoretic models of interdependent action [80]. Naturalistic decision making examines how experts act in time-pressured, real-world contexts by recognizing patterns and mentally simulating outcomes rather than systematically comparing alternatives [81]. Although all these paradigms accommodate ambiguity and depend on human expertise, they generally rely on relatively stable goals, evaluative criteria, roles, or accumulated knowledge. Ill-defined decision making, by contrast, is characterized more centrally by the need to iteratively construct and refine what constitutes a good decision.

In summary, ill-defined decision making defies standardization. It draws on inputs ranging from structured data and documents to social signals, expert knowledge, and personal reflection, each varying by context. For example, in financial fraud investigations, analysts decide whether to escalate anomalous transactions based on records, regulatory norms, and ambiguous behavioral cues [82]. In public health policy, officials make high-stakes choices such as imposing lockdowns or allocating resources, relying on incomplete evidence, stakeholder positions, and shifting societal values [83]. In job searches, individuals commit to career paths under uncertainty, balancing aspirations with constraints and evolving priorities [28], [84]. These decisions differ not only by domain but also by form: some involve selecting among predefined options, others require creating new ones; some rely on measurable objectives, others evolve through tacit negotiation. Preferences may be explicit or absent, constraints fixed or emergent, and even alternatives may lack clear structure. Together, these characteristics clarify the nature of ill-defined decision making but do not prescribe how to support it. This diversity makes it difficult – if not impossible – to impose a unified framework for AI-assisted decision support. Yet one pattern persists: the need for an evolving articulation of decision criteria. Even when objectives begin vague or contradictory, decision makers still articulate evaluative anchors that guide sensemaking, comparison, and commitment [61]. Our approach builds on this trajectory by proposing a framework that places criterion formation at the core of ill-defined decision support.

B. Which Principles?

We discussed ill-defined decision making as resisting full automation or purely data-driven resolution, characterized by incomplete, evolving, or contested decision criteria. This raises the question: what properties must decision-support systems exhibit to remain useful under such conditions? Standards like the FAIR principles (Findability, Accessibility, Interoperability, Reusability) have been transformative for data management and stewardship [85] and have been extended to software, workflows, and AI models [86]. Yet FAIR does not address how systems built on such data can support human decision makers in ill-defined contexts. To complement FAIR, building on the analysis in Sec. III-A, we introduce the **DARE principles** as requirements for real-world decision-support systems: ***D**eliberation*, ***A**gency*, ***R**esilience*, and ***E**mpathy*.

The choice of terms in the DARE principles takes inspiration from different disciplines. Agency connects to Shneiderman's HCAI vision, where systems should amplify rather than replace human judgment, preserving meaningful control [1] ("product"

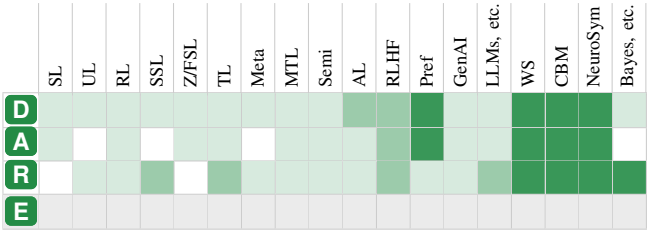
D **Deliberation**: How the dialogue unfolds
→ *How criteria are co-evolved and reconciled through dialogue.*
Deliberation means the system supports an ongoing dialogue around decision logic. Decision criteria are created, revisited, refined, and, when necessary, reconciled through interactions among stakeholders and with the AI system. Deliberation ensures that decision logic remains transparent, adaptive, and accountable by supporting co-evolution and negotiation.

A **Agency**: Who holds the pen
→ *Who the human or the system has authority over decision logic.*
Agency means that decision makers retain authorship and control over the operational logic. The system must enable humans to create, edit, and prioritize criteria, define proxies, set constraints, and negotiate trade-offs through interaction. Agency ensures that people, not algorithms, govern the decision logic.

R **Resilience**: How the system copes with reality
→ *How the system remains useful under messy conditions.*
Resilience means the system continues to support decision making under missing, noisy, or heterogeneous data, shifting goals, or uncertainty. It requires mechanisms for graceful degradation, uncertainty-aware reasoning, and substitution of unreliable inputs. Resilience ensures that decision support remains meaningful as real-world conditions evolve.

E **Empathy**: How humans are considered
→ *How the system attends to decision makers and those affected.*
Empathy means the system is designed with and for the humans involved, both those making decisions and those affected by them. It relies on participatory and empirically informed design practices and sensitivity to diverse contexts, ensuring the system does not abstract away lived realities. Empathy ensures that decision support remains not only functional but also humane.

Table I: Conceptual alignment of DARE principles and AI paradigms, reflecting our discussion in III-C; opacity shows relative tendencies that may evolve as new variants emerge.



Note: We use the term Artificial Intelligence (AI) rather than Machine Learning (ML) to reflect both scope and intent. Technically, several paradigms, such as logic-based, probabilistic, or conformal methods, extend beyond learning to include knowledge representation and reasoning which predate ML. Conceptually, our framework integrates interactive and visualization modules (input and explanation layers) that enable humans to co-shape decision logic by articulating, revising, and consolidating knowledge. This situates the approach within AI as a broader system of adaptive, explainable, and human-centered computation rather than ML as an isolated optimization process.

principle). Empathy echoes his "process" principle, informed by user-centered and participatory approaches [1] that involve real decision makers [6] as well as populations affected by decisions, resonating with ethical perspectives on how data practices can marginalize or empower people [87]. Deliberation aligns with calls for enhanced flexibility in interaction design [88] and with traditions of Socratic dialogue [89] and deliberative democracy that emphasize reasoned and iterative reconciliation of perspectives [90], [91]. Resilience responds to calls for acknowledging that data may be incomplete, uncertain, or unreliable [92], and also draws on ecological views of sustaining function under disturbance [93] and on policy research on robust decision making that addresses uncertainty and change [94]. We deliberately avoid terms like transparency, as the goal is not to expose all internal AI mechanics but to provide the right visibility for meaningful human control, much as a pilot reads the cockpit rather than the wiring. In our framing, such transparency is embedded within *Deliberation*, which emphasizes the interpretability and revisability of decision logic and criteria. Empathy complements this view by collecting empirical evidence on how much, and what kind of, information supports agency in context. In Sec. III-C, we shift from these concepts to a technical perspective, discussing how different AI paradigms potentially align with the DARE principles.

C. Which Paradigm of AI?

We examine how common AI paradigms [7], [95]–[97] relate to the DARE principles, focusing on Deliberation (D), Agency (A), and Resilience (R) (see Table I for a summary of our reflections). Empathy (E) is outside the scope of this paper, as it concerns empirical and contextual aspects of human involvement, such as UI design validation, domain intricacies, and societal impact, that cannot be meaningfully generalized across paradigms. Excluding it also acknowledges that future empirical work may refine or challenge the conceptual tendencies discussed here. For example, decision makers in specific domains may exhibit different preferences or AI interpretations.

Data-driven Paradigms: These learning-based approaches focus on statistical patterns derived from labeled or unlabeled data through fixed optimization objectives that minimize error by maximizing predictive accuracy (or similar performance metrics). They perform best when human goals are well specified and data are rich enough to reveal stable patterns, yet their design allows little flexibility to revisit or negotiate decision criteria once training begins. Supervised (SL) and unsupervised (UL) learning [98] exemplify this logic: the former relies on labeled examples encoding a single interpretation of correctness, the latter infers latent structure autonomously without human feedback. In both, learning proceeds mostly without dialogue or reinterpretation, limiting ***R**esilience* and human ***A**gency* to dataset construction. Reinforcement learning (RL) [7], though more iterative, operates under similar assumptions when reward functions are predefined. The model optimizes for a mathematically fixed notion of success that may diverge from evolving human criteria, especially in variants like RL from AI feedback (RLAIF) [99], thus constraining ***D**eliberation*. Self-supervised learning (SSL) [7] and contrastive methods (e.g., SimCLR, CLIP) generate pseudo-labels or relational constraints directly from data, achieving powerful representations but learning notions of similarity, importance, or relevance implicitly rather than through ***D**eliberation*. Zero- and few-shot (Z/FSL) learning ([7]) extend this pattern by leveraging pretrained embeddings to generalize to new classes with minimal or no labeled data (e.g., GPT-4o, CLIP). They exhibit technical ***R**esilience* through data efficiency, but remain bound to fixed representational spaces that cannot adapt when problem framings or decision semantics evolve. Transfer (TL), meta- (Meta), and multi-task learning (MTL) [7] reuse or jointly optimize representations across tasks to improve efficiency and generalization, yet their adaptability is algorithmic: they adjust parameters, not principles, and lack mechanisms to negotiate competing objectives or redefine goals. Data-driven paradigms thus excel under well-posed conditions but show limited ***D**eliberation* because objectives are defined ex ante, minimal ***A**gency* since reasoning cannot be revised during learning, and conditional ***R**esilience* confined to statistical rather than conceptual variation. In ill-defined or value-sensitive contexts, they risk producing consistent yet misaligned outcomes, precisely solving problems whose meaning may no longer be agreed upon.

Feedback-driven Paradigms: These approaches are another form of learning by example that incorporate human or external feedback during learning, reintroducing limited interaction into otherwise autonomous optimization. They enhance *deliberation* by allowing adjustment through evaluative signals and can restore partial *agency* via user interventions, yet human influence remains indirect, mediated by labels, preference scores, or corrective feedback rather than direct edits to decision logic. A first group combines labeled and unlabeled data or relies on selective querying, as in semi-supervised (Semi) and active learning (AL) [7]. These approaches improve data efficiency by involving humans in labeling the most informative or uncertain examples, briefly increasing ***A**gency* since users affect what the model learns from, but ***D**eliberation* remains confined to label selection. Task definition and optimization objectives stay fixed, and the system provides no means to reinterpret its learned representations. A second group introduces preference-based feedback, where humans express comparative judgments rather than categorical labels. RL with human feedback (RLHF) [100] and related preference-learning methods use such signals to align model outputs with human values. This mechanism improves perceived alignment but mainly at the outcome level: humans steer learning direction without access to or control over internal logic. Thus, these systems achieve limited ***D**eliberation*, since negotiation of decision criteria becomes approval or disapproval signals, and partial ***A**gency*, as users cannot author or revise underlying logic. Feedback-driven paradigms lie between static, data-driven learning and fully knowledge-based modeling. They offer moderate gains in ***D**eliberation* and ***A**gency* by integrating human evaluation, while iterative feedback improves technical ***R**esilience*. Yet their objectives remain predefined, and adaptation rationales are often opaque. They enable interaction without genuine co-authorship in decision logic: humans can influence the direction of learning, but not redefine what is learned.

Generative- and Foundation-based Paradigms: These paradigms learn data distributions to produce new, coherent samples or representations rather than predict fixed targets. They include models such as Variational Autoencoders (VAEs) [101], Generative Adversarial Networks (GANs) [102], and diffusion models [103], which synthesize realistic outputs by approximating or reversing data-generating processes. Recent foundation models, including large language, vision, and multimodal systems (LLMs, VLMs, LVMs) [104]–[106], extend this paradigm by scaling self-supervised and generative objectives to unprecedented levels. Even emerging *agentic AI* solutions – which add procedural behaviors atop these models – still rely on models trained on massive datasets that exhibit generalization and emergent reasoning-like capabilities. Yet, their flexibility remains statistical, not conceptual: they cannot expose, justify, or revise their reasoning, offering no genuine ***D**eliberation* or human ***A**gency*. Their apparent ***R**esilience* stems from coverage and scale rather than interpretive robustness, and their tendency to produce plausible yet unverifiable outputs ("hallucinations") undermines trust in sensitive domains. Generative and foundation paradigms thus illustrate the limits of purely data-driven creativity and highlight the need for conceptually transparent and editable reasoning structures, motivating hybrid approaches like CB-LLM [107] that combine foundation models with concept-based transparency for human oversight.

Knowledge- and Concept-driven Paradigms: These paradigms explicitly encode human reasoning, domain logic, or interpretable concepts into the learning process, shifting the focus from optimizing performance to making decision logic transparent and revisable. They align most closely with ***D**eliberation* and ***A**gency*, as humans can directly inspect, question, and modify the criteria a model applies. Weak supervision (WS) [108] exemplifies a knowledge-driven approach where experts articulate partial heuristics, textual cues, or labeling functions that together approximate ground truth. By formalizing domain intuition in a programmatic and revisable form, WS enables editable ***D**eliberation*, allowing criteria to evolve as understanding deepens, while preserving human ***A**gency* over labeling logic. Concept-based modeling (CBM) extends this idea by structuring internal representations around human-understandable concepts or symbolic rules [18], [109]. Such models enable decision makers to inspect intermediate reasoning, adjust relationships among concepts, and directly co-author the model's internal logic to maintain ***R**esilience* under changing conditions. Likewise, neurosymbolic AI fuses symbolic reasoning with neural learning, combining verifiable logic with adaptive behavior. Collectively, these paradigms embody a shift from parameter tuning to reasoning design, trading some scalability for shared interpretability and control. They show that effective AI for ill-defined or value-sensitive decisions depends on systems whose reasoning can be discussed, contested, and revised, making them the closest existing family to the DARE-aligned vision advanced in this work.

We acknowledge complementary approaches such as *Bayesian learning*, *swarm intelligence*, *genetic algorithms*, *multiobjective evolutionary algorithms*, and *uncertainty quantification* (e.g., conformal prediction). These methods can enhance ***R**esilience* by modeling and communicating uncertainty, yet do not constitute standalone AI paradigms. Rather, they act as cross-cutting layers that can be integrated across paradigms without changing their forms of ***D**eliberation* or ***A**gency*. Similarly, *federated learning* functions as a meta-framework across paradigms, emphasizing privacy-preserving model training on decentralized data while maintaining local autonomy.

IV. DARE Conceptual Framework

Building on the analysis in Sec. III, here we introduce the DARE framework (Fig. 1) for XAI and visualization in ill-defined decision making. The DARE framework is defined as a conceptual architecture that embodies the four DARE principles: ***D**eliberation*, ***A**gency*, ***R**esilience*, and ***E**mpathy*. A DARE framework can be conceived and instantiated through any AI paradigm, provided the theoretical design or implementation align with these principles. Our conceptualization of the DARE framework draws inspiration from recent frameworks on joint human–AI decision processes [13], [110], [111] and the core capabilities of two AI paradigms: *Weak Supervision* (WS), which treats decision criteria as evolving and partially specified, and *Concept-based Modeling* (CBM), which maintains interpretable conceptual structure among those criteria, preserving meaning across iterations. Sec. V shows how the AI-assisted decision making process described here can be instantiated through WS- and CBM-based mechanisms.

Sec. IV-A presents two running use cases illustrating different forms of ill-defined decision making. Building on Sec. III-A, we center the framework around the *choice* phase as the minimal condition that distinguishes decision making from other cognitive activities such as judgment, sensemaking, and analysis, and around evolving decision criteria as a stable anchor across ill-defined contexts. **Sec. IV-B** therefore presents the conceptual core of DARE, namely the iterative human–AI loop through which provisional criteria are articulated, tested, refined, and organized into higher-level concepts. **Sec. IV-C** situates this loop within the broader decision making process, consisting of intelligence, design, choice, and review, clarifying that the loop spans the full process but is most explicit in the choice phase. The remainder of the section explains how visualization and AI support this process. We distinguish between *core* support for the choice-centered loop (**Sec. IV-D**) and *supportive* functions across the other phases (**Sec. IV-E**). **Fig. 1** summarizes DARE's central support functions, input visualization, data exploration, information synthesis, XAI, and choice support, which the text develops as complementary mechanisms to articulate, interpret, compare, and refine decision logic. The section introduces additional forms of support, not shown in Fig. 1, that qualify, extend, and stabilize this process, including review and feedback mechanisms that close the loop, uncertainty representations that make criteria interpretable, collaborative deliberation that distributes reasoning across stakeholders, and mechanisms to externalize and consolidate decision knowledge. Sec. IV primarily operationalizes ***D**eliberation*, ***A**gency*, and ***R**esilience*; ***E**mpathy* remains a guiding principle, but because its realization depends on domain-specific empirical grounding and affected communities' perspectives, we discuss it in Sec. VI.

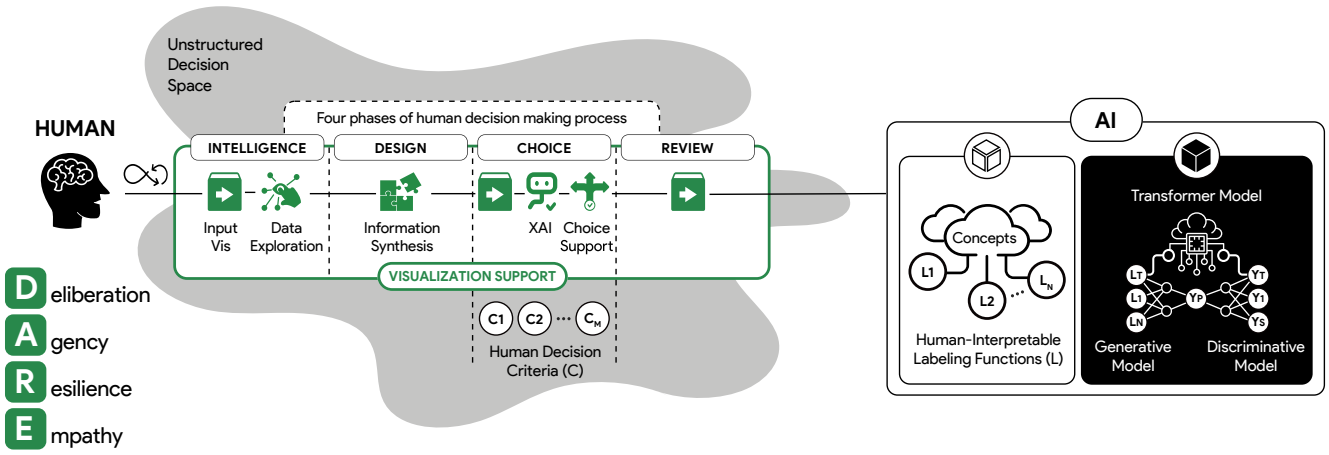


Figure 1: Overview of the DARE framework. Decision making unfolds within an unstructured decision space that follows four decision making phases of intelligence, design, choice, and review. Human decision makers articulate evolving criteria (C_t) through visualization interfaces that support several functions across these phases, including input visualization, data exploration, information synthesis, explainable AI, and choice support. The AI layer models these criteria as interpretable representations (L_t) and relates them through probabilistic and conceptual reasoning. Iterative feedback between C_t and L_t enables criteria refinement and the gradual emergence of decision structure. Together, visualization and AI instantiate the DARE principles.

A. Illustrative Use Cases

To make the abstract logic of the DARE framework more concrete and relatable, we introduce two ill-defined decision problems that serve as running examples throughout the paper.

- **Anti–Money Laundering (AML):** A compliance analyst must decide whether to escalate a client for suspected money laundering [112], [113]. The task involves incomplete evidence, shifting regulations, and reputational risk. What counts as "suspicious" changes as criminals adapt and auditors reinterpret past cases. The analyst starts with rough heuristics, such as transfer frequency or threshold amounts, and refines them through feedback and contextual reasoning.
- **Urban Heat Resilience (UHR):** A city planning team must decide how to allocate limited cooling measures, such as shade trees and water-spraying systems, before a heatwave [114], [115]. The decision balances competing goals: protecting vulnerable residents, conserving water, and minimizing disruption to traffic and commerce. Data are partial, expectations diverge, and success becomes evident only after implementation. Planners iteratively negotiate and adjust criteria for fairness, feasibility, and impact.

These domains differ in the nature of their unstructured elements. In AML, the decision space, data schema, and regulatory constraints are well defined, but the criteria for suspiciousness evolve with changing tactics and interpretations. Ground truth is delayed, trade-offs between compliance, fairness, and risk remain unresolved, and decision makers must continuously reinterpret patterns. The process is procedurally structured yet semantically fluid. In contrast, UHR decisions are unstructured at a deeper level: while data and the general goal of reducing heat stress are clear, the objectives, alternatives, and evaluation criteria are negotiated and often conflicting. Ill-definition thus extends beyond how to decide (criteria) to what to decide for (objectives) and among (alternatives), making the decision space open-ended and contextually contested.

B. DARE Core: Iterative Decision Criteria Formation

The DARE framework operates within an ill-defined *decision space* (Fig. 1), whose components, including objectives, criteria, alternatives, preferences, strategies, and constraints, are fluid. Within this space, provisional decision criteria serve as tangible anchors that connect evidence to evaluative meaning. The framework builds on the assumption that through continuous articulation, testing, and revision of these criteria, a decision structure can emerge.

This iterative process is supported by algorithmic aids that help scaffold emerging reasoning. Decision makers articulate provisional decision criteria $C_t = \{C_1, C_2, ..., C_M\}$ as verifiable rules that express their current understanding of the situation. Algorithms generate corresponding human-interpretable representations $L_t = \{L_1, L_2, ..., L_N\}$ that model how each criterion behaves relative to available evidence. The sets C_t and L_t need not have a one-to-one correspondence: several criteria may partially share the same representation when they draw on overlapping evidence, while a single criterion may produce multiple representations when its logic spans heterogeneous signals. Through probabilistic aggregation, the system estimates how these representations are supported by evidence, how they agree or contradict one another, and where coverage gaps remain within the decision space. The outcomes of this evaluation guide the next refinement cycle $C_t \rightarrow L_t \rightarrow C_{t+1}$, corresponding to the central human–AI feedback loop in Fig. 1.

For instance, in the AML use case, the compliance analyst can define provisional criteria such as "flag repeated small transfers within three days" and "flag transactions above €10,000." Algorithms test these rules on transactional and network evidence, revealing suspicious accounts, contradictions, and missing data for unlinked transfers. In UHR, planners can combine thermal readings, canopy coverage, demographic indicators, and citizen reports mentioning "heat stress." Algorithmic aggregation shows convergence in peripheral areas but conflict in central districts where social vulnerability is

high despite moderate heat. Interactive visualization exposes these inconsistencies and data gaps, guiding decision makers to adjust and extend criteria for the next iteration.

A well-functioning DARE AI-visualization system is characterized by the extent to which human-defined criteria can be expressed in its representations, tested against available evidence, and made transparent in their uncertainty when evidence is insufficient. Ideally, the algorithmic representation space should be expressive enough that every human-defined criterion can be captured within it, such that $C_t \subseteq f(L_t)$, where f denotes the mapping between algorithmic and human representations. The system must therefore accommodate diverse rule types, from boundary cases such as legally mandated criteria that cannot be empirically validated (e.g., AML regulations) to soft rules grounded in qualitative or experiential judgment (e.g., UHR planning, where fairness and community feedback lack direct datasets). The algorithmic layer may also propose new candidate criteria by detecting regularities in its evidence base and drawing on knowledge from prior decisions. These algorithmically generated propositions initially belong to L_t and enter the next iteration only if decision makers choose to adopt or refine them, becoming part of C_{t+1}. In AML, these may include recurrent transaction motifs across historical cases, such as sequences of cross-border transfers followed by rapid withdrawals. In UHR, algorithms can surface patterns from earlier planning outcomes, such as districts where cooling measures underperformed despite high vegetation. Decision makers, supported by interactive visualization, evaluate propositions, determining which to adopt, refine, or discard while maintaining transparency across evidence, uncertainty, and decision logic.

While individual rules provide testable anchors, decision structure cannot stabilize through isolated criteria alone. Each rule captures only a fragment of meaning, and without integration, their collective logic remains disjointed. To gain coherence, related criteria must be interpreted in relation to one another and organized into higher-level themes that convey why specific rules matter. The algorithmic layer supports this consolidation by grouping compatible rules into conceptual clusters that summarize shared patterns or causal explanations. Decision makers can inspect these clusters [116], trace which rules contribute to them, and refine or merge criteria accordingly. In AML, rules concerning repeated small transfers and near-threshold amounts may become the *fund structuring* concept. Recognizing this theme, the analyst reinterprets one of the earlier criteria to focus not on transaction size but on temporal regularity, improving its precision and explanatory reach. In UHR, planners may form a *socially amplified heat vulnerability* concept because of rules linking temperature, canopy cover, and isolation repeatedly. Hence, they may introduce a new rule on accessibility to cooling centers or adjust weightings to prioritize densely populated districts. Through such conceptual feedback, the system helps transform dispersed heuristics into an interpretable and evolving structure of decision knowledge, extending the criterion-refinement loop shown in Fig. 1 from individual rules to higher-level concepts.

In a mature state, the DARE system enables explicit comparison [117] and choice [23], as reflected in the choice-support component of Fig. 1. In AML, once concepts such as *fund structuring* or *rapid withdrawal pattern* have stabilized, decision makers can classify each case as *fraudulent* or *non-fraudulent* and justify their choices through the supporting concepts. In UHR, likewise, planners can specify intervention options, e.g., *increase canopy cover* or *expand social cooling centers*, and deliberate among them using the framework. In both domains, the transition from individual rules to structured concepts provides the coherence required for meaningful comparison and informed decision making.

The iterative formation of criteria is designed to realize the DARE principles. ***D**eliberation* arises from repeated inspection and revision of rules, concepts, and their interrelations. ***A**gency* is preserved because decision makers author, edit, approve, and reject rules and concepts. ***R**esilience* is supported by probabilistic aggregation and the inclusion of soft evidence, enabling progress under uncertainty, missing data, and disagreement.

Over successive iterations, the human and algorithmic representations, C_t and L_t, evolve toward mutual intelligibility. The goal is not symmetry but complementarity: algorithms learn to express reasoning in forms humans can interpret, while decision makers refine assumptions in response to what these representations reveal. As human criteria become encoded within algorithmic structures and algorithmic patterns gain human meaning, the two systems approach a shared language of reasoning. The DARE framework can thus be viewed as a form of hybrid intelligence [118], where human and computational reasoning develop together to extend rather than replace human decision making. This alignment limits the algorithmic search space to interpretable representations, favoring coherence, accountability, and shared understanding, qualities essential to decision making under ambiguity. It also reframes bias as an inherent part of reasoning. Both humans and algorithms rely on heuristics that can mislead or guide depending on how they are surfaced and examined [119], [120]. The DARE framework treats these heuristics not as biases or truths but as evolving hypotheses to articulate, test, and refine, turning bias from a hidden liability into an object of joint reflection across data, algorithms, and human judgment.

C. Decision Making Phases

The DARE framework follows Simon's four-phase model of decision making, also summarized in Fig. 1, (*intelligence*, *design*, *choice*, and *review*), understood as an iterative process that can move back and forth across phases [58], [61]. In the *intelligence* phase, decision makers make sense of the situation by gathering and interpreting information, framing the problem, and situating it in its context. In the *design* phase, they develop and refine possible courses of action, strategies, or solutions. The *choice* phase involves comparing, prioritizing, and committing to one or more of these candidates using decision criteria that may be fixed in structured problems (e.g., in MCDM) or that may coevolve with understanding in ill-defined problems, sometimes emerging even before the candidates are fully formulated [58], [61]. In the DARE framework, decision criteria act as provisional anchors that guide reasoning across phases. The *review* phase assesses

outcomes and process, draws lessons, and feeds them into the next cycle. For AI components, review corresponds to updating data inputs, retraining internal representations, and recalibrating evaluation functions.

This articulation is intended to cover decisions at many scales and stakes, from habitual parameter tuning and micro strategies to high-stakes public policy [27]. Across domains, the prominence of each phase depends on the structure of the problem. In AML, alternatives are predefined (flag or not flag), so the process cycles mainly between gathering evidence (intelligence), refining evaluative criteria (choice), and minor adjustments of decision rules that substitute for full design, with confirmed cases guiding future heuristics (review). In UHR, design is central: planners construct and compare interventions, such as shade trees and water-spraying systems, before deliberating on fairness and feasibility. In all cases, criterion formation can begin even before alternatives exist, resembling agreeing on what "livability" means before deciding where to place the next cooling intervention.

We envision a new generation of DARE systems in which AI and visualization support decision makers across all decision making phases (e.g., by organizing data, generating scenarios, or simulating courses of action [23]). Such support, however, depends on domain-specific needs and cannot be defined in a fully general, decision-agnostic way for ill-defined problems. This paper therefore centers on the *choice* phase, where evaluative criteria C_t are iteratively formed, tested, and refined through interaction with AI representations L_t, corresponding to the central interaction layer emphasized in Fig. 1. This criterion-formation loop provides a universal anchor for human–AI collaboration, while other forms of assistance in intelligence, design, or review become meaningful only once the decision space gains clearer structure.

D. DARE Support in Choice Phase

1) Input Visualization Support: Input visualization, highlighted as one of the central support functions in Fig. 1, refers to visual representations designed to collect or modify data rather than to encode pre-existing datasets [19]. It reverses the classical Card and Mackinlay [121] information visualization pipeline, which transforms data into visual form for analysis, by starting from human interaction: people express meaning through visual structures that are then captured as data. This paradigm acknowledges that data are malleable, incomplete, and context-dependent, and that visualization can serve as an active medium for sense-making and negotiation. Prior research has applied input visualizations in participatory, civic, educational, and personal-informatics contexts, where people contribute entries, annotations, or tangible marks to shared representations [19]. The DARE framework extends this logic to decision making. Here, input visualizations act as interfaces for articulating and revising evaluative knowledge rather than for entering factual records. They allow decision makers to express provisional decision criteria C_t, encode alternative hypotheses, and gradually transform intuitive reasoning into structured information that algorithms can interpret.

Beyond interpreting or comparing information, visualization can also collect decision-relevant knowledge. Input visualizations enable the DARE framework to capture early, uncertain reasoning directly through interaction, translating intuition into structured and revisable form. In AML, decision makers often do not know in advance which behaviors to flag as suspicious. An analyst may sketch a suspected money-flow path directly on a network canvas, annotate nodes with qualitative judgments like "unusual intermediary," and have the system search for similar patterns. In UHR, planners might adjust map-based sliders representing social or environmental priorities, effectively encoding evolving definitions of "vulnerability" or "livability." These actions externalize tacit reasoning that precedes formal criteria, providing interpretable inputs before any rule or dataset is fixed.

Input visualizations in DARE support a wide spectrum of expression. In the early stages of understanding a problem, decision makers may rely on qualitative, open-ended forms such as sketching, tagging, or spatial grouping, which make reasoning visible without forcing premature precision. As structure emerges, these interfaces can support more formalized expressions of logic and preference, such as conditional filters, relational mappings, or rule-based configurations akin to WS labeling functions (see Sec. V-A). Thereafter, a visual form may encode a probabilistic or logical query, for instance "flag districts with high temperature and social vulnerability," or a concept-level relation such as linking heat exposure, canopy cover, and isolation into the *environmental vulnerability* theme (see Sec. V-B). Through these mechanisms, input visualization becomes the visual entry point of the iterative loop $C_t \rightarrow L_t \rightarrow C_{t+1}$, where humans specify, refine, and generalize decision logic in ways that remain observable and editable.

In the AML case, analysts build queries by linking account nodes, setting temporal constraints, or adjusting transaction thresholds. Each action becomes an input visualization that reflects their current view of suspicious behavior. In the UHR case, planners overlay demographic and thermal layers, sketch correlations, or weight map regions. These interactions encode domain reasoning and feed it to the AI as structured evidence. Input visualizations therefore function as a dynamic interface between tacit expertise and algorithmic representation, capable of expressing everything from exploratory hypotheses to precise conditional statements.

By linking these visual inputs to the labeling and concept representations L_t, the system learns to interpret partial signals as provisional hypotheses. Over iterations, the sketches, annotations, and adjustments evolve into explicit, testable rules within the WS–CBM pipeline (see Sec. V). This process forms an interface between human judgment and algorithmic learning, turning vague intuitions into structured inputs that stay interpretable and traceable across decision cycles.

By broadening the input visualization paradigm from participatory data collection to decision-criteria articulation, DARE accommodates both implicit and explicit forms of reasoning. Visual interfaces no longer only elicit opinions or measurements but help formalize decision logic in ways that remain human-interpretable. This integration supports the framework's focus on iterative criterion formation, providing the expressive means through which deliberation, agency, and resilience can develop before analytic or predictive modeling takes place.

2) XAI Visualization Support: Explainability in the DARE framework, represented as XAI support in Fig. 1, differs from conventional post-hoc explanations that describe model behavior and outputs after training. Instead, it emerges from the continuous visibility of reasoning as humans and algorithms coevolve their understanding of the decision space. Visualization serves as the medium that externalizes this reasoning process, connecting the algorithmic representations L_t with the human criteria C_t in ways that are transparent, revisable, and interpretable. Thus, explainability becomes a property of the interaction, not of a single model or output.

To situate this view within the broader XAI landscape, traditional approaches can be characterized by three dimensions: the *explanation level* (local vs. global), the *implementation level* (intrinsic vs. post-hoc), and *model dependency* (specific vs. agnostic) [122], [123]. Methods such as LIME and SHAP provide local, post-hoc, model-agnostic explanations by approximating predictions through feature attributions. In contrast, intrinsic approaches such as rule lists and self-explaining neural networks embed interpretability into the model itself. VA systems such as RuleMatrix, explAIner, and MELODY, extend these methods by combining algorithmic explanations with interactive exploration [122], [124], [125]. Building on this, the DARE framework reframes explainability as an ongoing dialogue [13] between human and machine reasoning rather than a one-time translation of model behavior.

Visualization plays three complementary roles in this process. First, it *renders model reasoning interpretable*. Internal representations L_t can be visualized as hypotheses that show how individual heuristics or labeling functions combine into higher-level structures. For example, analysts can inspect which weak signals dominate probabilistic aggregation and where conflicts occur. These visualizations extend local and global explanation concepts by revealing not only what drives a single prediction but also how reasoning patterns evolve across iterations.

Second, visualization *supports diagnostic and corrective control*. Instead of producing static explanations such as feature-importance bars or saliency maps, DARE visualizations allow users to interrogate and modify reasoning itself. By adjusting or disabling heuristics, users can reshape the algorithm's logic and observe the effects on predictions or concept structures. In AML, an analyst can examine clusters of flagged accounts, identify that "foreign shell recipients" are over-weighted, and reweight this rule to recalibrate the model. This interactivity transforms explanation into reasoning and correction.

Third, visualization *integrates multiple explanation levels*. Local views reveal how individual criteria affect single cases, while global views summarize the conceptual landscape of reasoning across cases. Provenance traces can show how rules and concepts evolve from C_t to C_{t+1}, linking each decision to its human and algorithmic lineage [126]. Uncertainty views can differentiate between epistemic gaps (missing data) and conceptual disagreement (conflicting criteria) [127], [128]. Counterfactual exploration tools can visualize how modifying a criterion, such as redefining the "vulnerability" of minority populations in the UHR scenario, propagates through priorities and trade-offs [129]. These multi-level representations unify local interpretability with systemic transparency.

Collectively, these mechanisms move beyond feature attribution toward *criterion- and concept-level explanation*, where visualization reveals how decision logic is structured, debated, and refined together with the conceptual relations that give it meaning. Explainability in this sense concerns both the rules through which criteria operate and the interpretive structures that organize them. This view aligns with trends in concept-based XAI [130], [131], which emphasize interpretable conceptual reasoning rather than numerical attribution. Within the DARE framework, visualization does not simply explain what the system produces but makes visible how human and algorithmic reasoning co-construct both evaluative criteria and the concepts that connect them, showing explanation as a living process of shared understanding and communication.

Designing XAI visualizations for ill-defined decisions requires a careful balance between interpretive depth and cognitive accessibility. The same interfaces that expose reasoning must protect users from overload. When people navigate evolving human decision criteria or AI-suggested refinements, the goal is not to simplify complexity away but to make it navigable. Progressive-disclosure techniques can let explanations unfold in layers, from intuitive summaries to detailed traces of logic as users demand more control. Mixed-initiative interaction allows the system to highlight emerging inconsistencies or untested assumptions, inviting reflection rather than prescribing answers. Collaborative review spaces can transform explanation into a shared process: analysts, domain experts, and other stakeholders can each inspect how reasoning has evolved and propose adjustments. Through these mechanisms, XAI visualization in DARE helps sustain ***D****eliberation* by turning reasoning into a visible and discussable object, reinforces ***A****gency* by giving users the ability to shape explanations, and supports ***R****esilience* by revealing drift, disagreement, and uncertainty before they accumulate into failure.

3) MCDM-style Visualization Support: As decision criteria become more stable and alternatives take shape, reasoning shifts from exploration to comparison, corresponding to the choice-support function summarized in Fig. 1. MCDM provides useful vocabulary for this stage: alternatives, criteria, and preference relations connected through a decision model [23], [67]–[70]. In ill-defined settings, however, these components remain partial and fluid [28], [84]. The goal is not to compute an optimal outcome but to help decision makers compare, rank, and refine emerging options even when the structure is incomplete.

Visualization in DARE supports such semi-structured comparison. Interactive ranking and weighting techniques let decision makers include qualitative alongside quantitative criteria [84] and express preferences that are descriptive, ambiguous, or evolving [28]. Users can add or adjust criteria on the fly, view how provisional weights change alternative order, or note which comparisons are uncertain. This flexibility allows evaluation to proceed before data or models are finalized.

When only fragments of the decision space exist, the system can represent hypothetical or idealized alternatives (e.g., an analyst's notion of a transaction pattern). These serve as reference points against which new candidates can be contrasted. Visualization shows how far each alternative lies from such aspirational benchmarks, guiding the refinement of criteria and

search for new options.

As alternatives and criteria multiply, DARE integrates recommender-style assistance [71]–[73]. These mechanisms do not prescribe choices but highlight promising comparisons, trade-offs, or underexplored regions of the decision space. Suggestions remain transparent: users see the rationale for each recommendation and may accept, adapt, or reject it. By combining MCDM reasoning with transparent, suggestive guidance, DARE turns the choice phase into a navigable space. Visualization helps decision makers experiment with evolving criteria, mix qualitative and quantitative judgments, and test hypothetical directions before deciding. MCDM-style support in DARE links criterion formation to actionable deliberation while preserving interpretability and agency under uncertainty.

E. DARE Support across the Decision Process

While choice constitutes the conceptual core of DARE, decision making unfolds across additional phases that prepare, contextualize, and refine it, as schematically displayed in Fig. 1. The phases, *intelligence*, *design*, and *review* provide complementary functions that sustain the iterative criterion-formation loop. Cross-cutting mechanisms for uncertainty representation and collaborative deliberation ensure that all phases remain interpretable and participatory.

1) Intelligence Phase: Decision-Oriented Exploration: Exploratory visualization in DARE, corresponding to the data-exploration function in Fig. 1, differs from analytic or sensemaking exploration aimed at fully understanding a dataset. Its purpose is to surface patterns, irregularities, or constraints that reshape how a decision is framed. Decision-oriented exploration focuses on relationships that redefine what is desirable or feasible, rather than describing everything that exists. The outcome is not complete knowledge but a clearer sense of what to consider next.

Such exploration must also accommodate qualitative and contextual information. Decision makers may work with mixed data – text, spatial traces, notes, or sketches – and use visual interaction to externalize early interpretations. These traces act as provisional evidence that can later inform criteria or alternatives. Input visualization mechanisms capture this reasoning within the exploration space, enabling hypotheses, annotations, and conceptual cues to feed forward into later phases of the decision process.

In AML, timeline and network views can expose repeating €9,900 transfers or cycles among shell entities, suggesting thresholds worth scrutiny. The AI layer highlights weak correlations – shared beneficiaries, new accounts – as provisional hypotheses. In UHR, heat maps and demographic overlays can reveal districts where temperature and social isolation intersect, prompting planners to revisit definitions of vulnerability.

Decision-oriented exploration in DARE is thus a qualitative and quantitative bridge between perception and commitment. It preserves ***A****gency* by keeping humans in charge of where and how to look, while AI contributes by expanding what can be perceived. The result is a focused exploratory stage that ensures subsequent design and choice phases are grounded in situational awareness rather than exhaustive analysis.

2) Design Phase: Information Synthesis and Simulation: This is the DARE phase in which decision makers start shaping the decision space itself, corresponding to the information-synthesis function highlighted in Fig. 1. After initial exploration, they must translate partial insights into structured yet provisional courses of action. Design is thus an act of construction rather than discovery: it combines fragments of evidence, constraints, and intentions into hypothetical configurations that can later be compared or refined.

Visualization supports this work by assembling and transforming possibilities. Decision makers can create composite alternatives by combining quantitative with qualitative factors, sketching interventions, or remixing past solutions into new constellations. AI assists by proposing plausible completions or analogies, expanding what humans can imagine without dictating outcomes. Together, they form a sandbox for reasoning where patterns and narratives can be recombined to express "what could be" before committing to "what should be."

Simulation extends this sandbox into dynamic feedback. By adjusting parameters or assumptions, users can observe the directional effects of their ideas, revealing tensions and trade-offs without claiming predictive precision. In AML, analysts may simulate how different risk-scoring policies would change the balance between missed and false alerts. In UHR, planners can sketch hypothetical combinations of shading, vegetation, and social programs to see how they alter cost or equity. Each exploration makes the evolving structure of possibilities explicit, bridging intuition and evaluation.

Through such constructive and reflective use of visualization, the design phase helps decision makers give form to alternatives not present in the data. It turns synthesis and simulation into tools for imagination, grounded enough to inform comparison yet open enough to preserve deliberation and learning.

3) Review Phase: Reflection, Feedback, and Learning: The review phase in DARE, linked to the feedback dynamics shown in Fig. 1, transforms past decisions into data for future reasoning. It closes the loop between human judgment, algorithmic adaptation, and social accountability. Technically, this phase acts as a dynamic repository where decision outcomes, criteria revisions, and contextual notes are captured as new inputs via visualization interfaces. These input visualizations allow decision makers to document how criteria were applied, why specific options were chosen, and what consequences followed. The resulting records form a growing evidence base that supports model retraining and institutional memory.

For the AI layer, review data serve as delayed supervision (Fig. 1, AI). Confirmed outcomes, revised rules, or retrospective justifications can be integrated as weak (probabilistic labels Y_P) or strong labels (ground truth labels for various samples $Y_1, Y_2, ..., Y_S$), recalibrating models and enriching conceptual representations. For humans, visualization of this evolving archive enables reflection, audit, and cross-case comparison. Analysts can inspect which criteria led to reliable outcomes, detect drift or bias, and trace how definitions of success have shifted over time. For affected communities, these same visual summaries provide transparency and grounds for explanation, showing how reasoning evolved and where accountability lies.

By integrating outcome feedback, justification, and retro-

spective annotation into a single interactive layer, the review phase ensures that learning in DARE is both computational and ethical. It turns visualization into an input channel for collective memory, where human and algorithmic reasoning continue to co-evolve through evidence, critique, and reflection.

4) Uncertainty Representation: In ill-defined decision making, uncertainty is not an exception but a structural feature of reasoning. Data are incomplete, criteria provisional, and objectives may evolve mid-process. Visualization in DARE therefore treats uncertainty not as a residual to be minimized but as a property to be understood and acted upon. Its purpose is to make uncertainty accessible without overwhelming attention or inducing false confidence. Visual cues can distinguish data sparsity, model disagreement, and conceptual ambiguity, helping users identify which type of uncertainty matters in the current phase. In practice, this means showing uncertainty as a gradient of commitment: what is known enough to act on, what remains open to revision, and what requires further evidence. Such representations let decision makers treat uncertainty as a guide for deliberation rather than a barrier to decision. They can see where to seek more information, where to proceed cautiously, and where disagreement is acceptable. In this way, DARE positions uncertainty visualization as a mechanism of ***R**esilience*, keeping reasoning transparent under ambiguity and sustaining progress without pretending to eliminate doubt.

5) Collaborative Deliberation: Many ill-defined decisions are collective, involving experts, institutions, and affected populations whose perspectives differ in values, expertise, and priorities. DARE treats collaboration not as consensus building but as the coordinated articulation of these viewpoints. Visualization provides the medium for this process: it externalizes individual reasoning so that differences in criteria, evidence, or interpretation become visible and discussable. In early phases, shared exploratory views support collective sensemaking and reveal how participants frame the problem. During design and choice, comparative visualizations show how proposed criteria or alternatives affect outcomes for different stakeholders. In review, provenance views trace who contributed which rules or justifications, supporting accountability and dialogue. By allowing participants to see both their own reasoning and that of others, visualization enables co-deliberation as an ongoing process rather than a single meeting. It gives teams and the public a shared representational ground where disagreement can be examined without erasing it. Hence, DARE extends deliberation, agency, resilience, and empathy from individual reasoning to collective practice, keeping decision making transparent, participatory, and socially grounded.

F. Extending and Consolidating Decision Knowledge

1) Beyond Fixed Data: Externalizing Decision Knowledge: In ill-defined decisions, datasets are often incomplete, unstable, or unable to capture experiential or contextual knowledge. Decision makers routinely reason with information that lies outside formal data representation, such as field observations, undocumented heuristics, or intuitions about data reliability. To remain valid in these settings, the DARE framework moves beyond fixed datasets and supports the externalization of decision knowledge that is implicit, distributed, or qualitative.

Visualization research provides precedents for this extension. McCurdy *et al.* [132] formalize how experts externalize *implicit error* – qualitative awareness of discrepancies between data and reality – through annotation and contextual description. Their framework shows how visualization can capture and contextualize such knowledge, integrating it with quantitative data for validation. Similarly, Lin *et al.* [133] show that analysts often articulate *data hunches*, or personal insights absent from the data, and that systems can use these as structured, revisable inputs to collective reasoning.

Within DARE, these mechanisms enrich the iterative loop between human criteria C_t and algorithmic representations L_t. A decision maker may:

1) express qualitative or experiential knowledge as provisional criteria that highlight missing or uncertain relations (e.g., "flag when social vulnerability is likely under-reported");
2) attach annotations or "data hunches" to visual elements, which serve as soft evidence with adjustable confidence;
3) connect such statements to existing conceptual structures, bridging experiential and computational knowledge.

Through these channels, information outside the dataset becomes part of the deliberative loop. Algorithmic reasoning can absorb it probabilistically or conceptually, while visualization keeps its provenance and uncertainty visible. This extension allows the DARE framework to bridge structured data, qualitative insights, and expert judgment. It reinforces ***D**eliberation* by making tacit understanding explicit and discussable, and ***R**esilience* by enabling decision logic to adapt as data coverage or interpretation evolves.

2) Toward Conceptual Modeling of Decision Knowledge: Repeated use of the DARE framework gradually transforms episodic reasoning into explicit and inspectable representations of decision knowledge, extending the iterative dynamics shown in Fig. 1 into longer-term conceptual accumulation. As labeling functions and concept relations evolve, they can be consolidated into structured models, concept graphs, ontologies, or other interpretable semantic layers that document how human criteria, alternatives, and justifications change over time. CBM like Large Concept Model (LCM) [18] provides the substrate for such consolidation, encoding the co-evolution of human criteria C_t and AI representations L_t within an editable conceptual space (see Sec. V for further details).

The long-term outcome is an evolving model of domain-specific decision logic aided by visualization that reveals the emerging reasoning structure. Temporal concept maps can show how notions such as "suspicious pattern" in AML or "urban vulnerability" in UHR shift across iterations, which heuristics drive those shifts, and how model predictions adapt. These visualizations serve as living records of deliberation, enabling experts to trace, critique, and reuse reasoning across contexts.

Conceptual modeling completes the DARE cycle by turning cumulative human–AI interactions into durable knowledge. It operationalizes ***A**gency* through editable representations, ***D**eliberation* through traceable reasoning structures, and potentially ***E**mpathy* through awareness of how decisions affect people and communities. The goal is not merely to improve prediction accuracy but to build a shared, interpretable foundation for AI systems grounded in evolving human judgment.

V. DARE Framework Instantiation: Weak Supervision & Concept-Based Modeling

We next show how DARE can be realized through WS and CBM, though it is generalizable to other approaches (e.g., neurosymbolic or human-centric agentic AI).

A. Weak Supervision for Human Decision Criteria

WS [17], [108] provides a way to train models when labeled data are scarce, expensive, or conceptually ambiguous. Instead of manually annotating large datasets, users encode fragments of their reasoning as small programs called *labeling functions* (Fig. 1, white box). Each labeling function represents a partial interpretation of what a correct decision might look like, using diverse signals such as heuristic rules, keyword or regular-expression matchers, semantic or dependency patterns, metadata filters, knowledge-base lookups, or other model outputs. This flexibility makes WS modality-agnostic: the same logic can integrate text, numbers, images, graphs, or multimodal features within one reasoning process. Such heterogeneity parallels how decision makers form and test criteria in ill-defined problems: they draw on mixed cues, some quantitative, some experiential, and some intuitive. Visualization can help expose and mediate this diversity by showing how inputs interact, where reasoning overlaps or diverges, and which sources dominate the outcomes. These heterogeneous inputs can later be organized into higher-level decision concepts (discussed in Sec. V-B).

Because labeling functions often overlap, contradict, or capture only parts of the underlying logic, a *label model* aggregates their outputs into probabilistic labels. Conceptually, the process resembles a deliberative committee in which each labeling function expresses one voice, sometimes confident, sometimes uncertain. These voices may come from a single decision maker experimenting with multiple heuristics or from several people contributing domain knowledge at different levels of expertise, such as trainees and senior analysts. The label model, acting as a *generative model* (Fig. 1, black box), learns how accurate and interdependent these voices are, estimating which members of the committee tend to agree for the right reasons. Instead of relying on a single definitive "ground truth," which in real decision making is rarely available, the label model constructs a probabilistic approximation of consensus, where uncertainty is treated as evidence rather than noise. This step transforms individual heuristics into a collective judgment about the data, expressed as *probabilistic labels* Y_P.

These probabilistic labels train a *downstream model* that learns to predict the target label from data. In statistical terms, it acts as a *discriminative model* (Fig. 1, black box) that estimates the conditional probability $P(Y|X)$, where X represents input features and Y the aggregated probabilistic label. This component operationalizes collective reasoning, generalizing it to unseen cases. In the committee metaphor, if the label model captures the discussion leading to a shared judgment, the discriminative model learns to recognize the patterns that would lead the committee to similar conclusions.

The generative and discriminative models thus play complementary roles. The label model focuses on reasoning quality, estimating which labeling functions are reliable and how their outputs correlate, while the discriminative model learns to apply that collective reasoning to raw data. Their interaction resembles an iterative peer-review cycle: the label model aggregates and critiques individual arguments, the discriminative model tests how well these arguments generalize, and visualization reveals discrepancies between them. Users can then refine labeling functions, adjust assumptions, or introduce new rules based on contextual insight. The updated functions feed back into the label model, producing a revised probabilistic consensus. This iterative loop transforms WS from a one-off labeling method into a dynamic reasoning environment in which human and machine knowledge develop jointly.

WS was originally designed for human-authored heuristics [17], [108], but it can also incorporate hypotheses generated by computational systems. For example, a *transformer-based architecture* (Fig. 1, black box) like recent LLMs can propose labeling functions that enter the aggregation process [134], [135]. These AI suggestions extend the deliberative "committee" with new members offering other perspectives, subject to human review, weighting, and correction. Here, humans provide principled knowledge while AI expands the hypothesis space, and both remain accountable to transparent combination rules. This iterative process of defining, combining, and revising labeling functions instantiates the loop $C_t \rightarrow L_t \rightarrow C_{t+1}$. Decision criteria C_t are expressed as labeling functions, the label model aggregates labels L_t, and user reflection updates the criteria to C_{t+1}. This approach supports ***D**eliberation* by making the evolution of decision logic explicit, ***A**gency* by leveraging human reasoning, and ***R**esilience* by treating uncertainty as an intrinsic part of productive reasoning.

B. Concept-based Modeling for Decision Modeling

While WS structures how decision rules are expressed, it does not capture their meaning. CBMs such as CB-LLMs [107], LCMs [18], and Concept-Based Reasoning Models (CRMs) [136] fill this gap by introducing a semantic reasoning layer that operates on interpretable *concepts* instead of token-level patterns (surface regularities in word or symbol sequences without explicit meaning). A concept is an atomic idea, often a sentence or short statement, encoded within a shared representation space like SONAR [18], a multilingual embedding framework for meaning alignment across languages and modalities. This enables reasoning over textual, numerical, and spoken information within a unified semantic structure.

LLMs and other popular AI models capture associations between words and contexts but often represent meaning only implicitly. CBMs extend this capability by introducing explicit, reusable concept representations. They learn how concepts relate through similarity, hierarchy, or causal dependency, producing interpretable links that humans can examine and edit. In practice, a CBM can associate the pattern "multiple near-threshold transfers" with higher-level notions such as *structuring to avoid regulation* or *emerging financial risk*. These associations transform surface heuristics into conceptual explanations and connect probabilistic patterns to meaningful categories. Because their concept relations are explicit, CBMs are also better suited to value-laden domains, where reasoning must remain inspectable and accountable.

Within the WS pipeline, CBMs provide semantic grounding for labeling functions. Embeddings learned by the discriminative model can be organized by the CBM into clusters of related concepts. Users can inspect these clusters, assign domain-relevant names, and reuse them as new or refined labeling functions L_T. These L_T are primarily generated by the CBM based on its inferred concept structure. In this way, CBMs translate statistical representations into editable conceptual structures, making the iterative refinement of criteria more transparent and interpretable. Over time, such concept networks could evolve into representations of decision logic itself, revealing how criteria, hypotheses, and trade-offs interact in ill-defined decisions (see Y_T in Fig. 1, black box).

CBMs primarily reinforce ***D**eliberation* and ***A**gency* by enabling reasoning with interpretable concepts rather than opaque model features and by keeping learned representations open to inspection and modification, respectively. By stabilizing and tracing meanings across heterogeneous data, CBMs support ***R**esilience* by maintaining coherence in evolving or contested decision contexts. They also complement WS by shifting the focus from defining probabilistic rules to modeling the conceptual structure of decision knowledge.

C. WS–CBM XAI Pipeline

This section describes how WS and CBMs interact within a unified pipeline. The goal of this integration is to align human reasoning – criteria, alternatives, and conceptual framing – with machine representations so both can evolve toward a shared and transparent understanding of the decision space.

WS captures fragments of human decision making as probabilistic rules, while CBMs provide the semantic structure linking those rules to explicit meanings. The process begins when a decision maker formulates heuristics as natural-language prompts or structured conditions, which are translated into *labeling functions*. A *generative model* aggregates these signals into probabilistic annotations, and a *discriminative model* generalizes from them to infer broader patterns. The resulting latent representations are passed to the *transformer model* (i.e., the CBM), which clusters them into interpretable concepts and reveals thematic relations or inconsistencies among labeling functions. Visualization mediates this interaction by showing how probabilistic and semantic layers intersect and where human rules overlap, conflict, or leave gaps.

Through this coupling, the system supports continuous refinement. Decision makers inspect how rules group within conceptual clusters, trace ambiguous regions, and modify or merge labeling functions. The CBM, informed by these updates, can suggest new candidate concepts or rephrased heuristics that extend coverage or improve consistency. When conflicts persist, visual summaries help users revise or retire labeling functions keeping uncertainty and disagreement traceable. Authoritative ground-truth evidence, when available, can recalibrate both the WS model and the CBM by reinforcing validated criteria and down-weighting contradictory rules. These mechanisms keep the reasoning process auditable and empirically grounded. Each iteration thus updates both the probabilistic labeling model and the semantic organization of decision knowledge.

The integrated pipeline aligns human and machine reasoning in complementary ways. WS maintains explicit, testable rules that mirror human judgment, while CBMs preserve the semantic coherence of those rules. Visualization acts as a boundary object that exposes both probabilistic uncertainty and conceptual relationships. Together, these components realize ***D**eliberation* and ***A**gency* by keeping decision reasoning visible, editable, and semantically grounded. They also support ***R**esilience* by stabilizing shared concept meanings as data, domains, and stakeholder values evolve. Once the decision structure has stabilized, the process transitions naturally to the choice phase, supported by the MCDM approaches discussed in Sec. IV-D3.

D. Preliminary Testing of the WS–CBM Pipeline

The DARE framework is designed to reason with ambiguity, context dependence, and value-laden character of real decision making. Such situations rarely lend themselves to fixed benchmarks or standardized evaluations, since what counts as "ground truth" is itself under discussion. Nevertheless, it is important to show that the underlying mechanisms of the WS–CBM pipeline are computationally realizable and can support iterative, human-centered reasoning. This section therefore offers a modest illustration of systems embodying elements of DARE's logic, along with a small proof of concept in the AML domain. Although the examples rely on LLMs for implementation, the focus lies not on the models themselves but on how their prompting and refinement capabilities can operationalize the WS–CBM process of articulating and revising decision knowledge, with concepts serving as an added layer for humans to organize and interpret that knowledge, as in CB-LLM [107].

Lin *et al.* [137] developed an interactive system that integrates WS with LLMs to help users build and refine labeling functions for text classification. Their system connects Snorkel with GPT-4o for prompt-based labeling function generation and BERT for validation, while visualizing coverage, conflict, and agreement among rules. Users iteratively modify prompts in response to disagreement matrices and keyword highlights, clarifying vague or overlapping heuristics. For instance, a user who begins with "Does the text contain a request for social media engagement?" refines it into "Does the text ask you to follow someone on social media?", improving precision and interpretability. Although the task is bounded (phishing email detection), the study demonstrates several principles central to DARE: that human heuristics can be expressed, revised, and merged through interactive visual workflows; that probabilistic aggregation can mediate disagreement among rules; and that prompt-driven iteration supports deliberation between human and machine reasoning. This work serves as an empirical precedent for the WS component of DARE, showing that iterative co-supervision is feasible with current tools.

To explore feasibility in a decision making domain, we applied a simplified version of this logic to the AML case described earlier. Using the IBM HI-Small_Trans dataset [113], which contains roughly five million transactions with 0.1% labeled as laundering, a decision maker defined an initial heuristic: "Label transactions forming a circular flow of funds that begin and end in the same account." This natural-language

prompt was translated into a labeling function using GPT-4o, which detected simple cycles within a transaction graph and flagged 1.92% of cases on a 20% development subset. The LLM layer then proposed semantically related prompts such as "flag transactions above €50,000 with identical sending and receiving currencies," which the decision maker reviewed and refined; such prompts can be further organized around interpretable AML concepts such as structuring/smurfing, fan-in patterns, or circular flows. While preliminary and small-scale, this exercise illustrates how heuristic formulation and probabilistic aggregation can be combined with conceptual organization to structure reasoning in a real decision context.

Although a framework such as DARE cannot be "tested" in isolation, since its ***E**mpathy* principle implies that it must be experienced through participatory use engaging decision makers, domain contexts, and the people affected by their decisions, these examples show that its core operations are technically attainable. The interactive workflow of Lin *et al.* demonstrated how labeling functions can be generated, inspected, and refined through iterative human–AI exchange, providing a concrete basis for criterion formation and probabilistic reconciliation. The AML exercise extended this logic to a setting where data are incomplete and only partially labeled, showing that heuristic prompts can still organize model behavior and help reveal how conceptual structure emerges through refinement. Together, these cases indicate that the computational components required for the DARE framework, including iterative criterion formation, probabilistic aggregation, and semantic organization, can already be realized with current tools and methods. At the same time, they highlight a deeper challenge central to real-world decision making: decision making rarely begins or ends within a well-defined dataset, but extends to forms of understanding and evidence that are qualitative, contextual, or only partially captured in data (see Sec. IV-F1).

VI. DARE Challenges

While our framework presents a novel and promising approach to supporting ill-defined decision making through visually-assisted and CBM-enhanced WS, several technical considerations arise that we see as future research directions.

Dependence on Foundation Models & Prompt Complexity: A fundamental component of the framework is its reliance on large pre-trained models, particularly LLMs and, more recently, LCMs, to generate labeling functions. While LLMs effectively generate fluent prompts and candidate rules, they remain prone to hallucinations, overgeneralization, and logical inconsistencies, especially when faced with complex or nuanced instructions. LCMs offer improved reasoning and abstraction capabilities that more closely align with human conceptual thinking, yet they remain a new technology that requires further testing. Encouraging users to frame their heuristics in minimal, well-scoped terms is often more productive than attempting to encode intricate domain logic all at once. Empirical observations suggest that simpler, more focused prompts often yield more accurate and interpretable labeling functions than complex, multi-condition rules, reflecting the DARE framework's hybrid design that combines foundation-model capabilities with human-interpretable concepts [107].

Expressiveness & Scope of Labeling Functions: While DARE lets domain experts encode knowledge via labeling functions, not all decision logic can be captured with rule-based heuristics. Real-world judgments often rely on implicit understanding, emotion, institutional memory, or subjective prioritization, which are hard to translate into formal or semi-formal forms, even with LCM assistance. As such, we may oversimplify complex reasoning or exclude tacit factors. Future work could explore alternative representations, such as fuzzy logic, hybrid statistical-symbolic reasoning, or user-tuned embeddings that better capture nuance and ambiguity.

Challenges in Human–AI Alignment: A central goal of the framework is to iteratively align human decision criteria with model-generated labeling functions. However, this alignment process is vulnerable to drift, that is, users may unknowingly accept flawed or misleading AI-generated suggestions, or fail to detect when a labeling function diverges from their original intent. Although the framework includes visual supports for traceability and conflict detection, maintaining true alignment requires both technical scaffolding and sustained cognitive engagement from users. Future extensions might include automatic surfacing of critical disagreements, guided walk-throughs for refining problematic functions, and even conversational agents that question the user's assumptions in a reflective loop.

Evaluation Constraints & Feedback Quality: Without ground truth, evaluating labeling-function quality and downstream predictions is inherently difficult. Metrics such as coverage or conflict rate only partially reflect usefulness or correctness. Additionally, feedback signals are often weak or ambiguous; for instance, a low-confidence prediction might reflect label noise, edge-case behavior, or misalignment between human intent and model learning. Incorporating uncertainty quantification (e.g., via conformal prediction) and supporting comparative visualizations of competing explanations could help users better interpret these signals and improve decision outcomes.

Mitigating Cold Start Challenges: DARE should further address the cold start problem, particularly in domains where ground truth labels are scarce. In such settings (e.g., financial fraud or policy evaluation), it is difficult to bootstrap reliable labeling functions or evaluate model outputs early on. This scarcity hinders model training and user trust. Strategies like exemplar-based seeding, adopting visually assisted function refinement, and integrating high-confidence anchor cases (e.g., legally confirmed frauds) can help mitigate these challenges.

Scalability & Domain Generalization: Although the conceptual architecture is domain-independent, practical implementation and user interaction design may need significant adaptation across domains. The AML use case shows feasibility, but settings such as clinical triage, policy analysis, or educational diagnostics would require rethinking input mechanisms, model architectures, and user interfaces. Moreover, in multi-stakeholder environments, disagreements about criteria or definitions (e.g., fairness, risk, relevance) introduce additional complexity. Future iterations should explore how the framework

can support collaborative decision making, surfacing not only individual heuristics but also the tensions between them.

Usability for Non-technical Users: Despite its human-centered orientation, the framework presupposes a degree of fluency with VA and AI-supported reasoning. For non-technical users, even high-level interactions with labeling functions may feel abstract or unintuitive. This needs empirical testing in a DARE VA system designed following the principles presented in this framework paper. For example, the VA system could integrate pre-built templates for common tasks, provide contextual tooltips and examples, or track interaction histories to offer undo/redo support and model transparency over time.

Empathy & Domain-Grounded Decision Making: Within the DARE framework, the realization of ***Empathy*** differs from the other principles in that it cannot be adequately grounded through abstract criteria, technical system design, or model architecture alone. Decision making in high-stakes domains depends on domain-grounded knowledge that is not fully captured in data or formal representations: clinicians know which signals are missing or unreliable in medical records, policy makers interpret indicators within institutional and societal contexts, and urban planners understand which local conditions, infrastructure constraints, or lived vulnerabilities are absent from spatial datasets. At the same time, decisions often affect individuals and communities whose experiences, constraints, and vulnerabilities remain outside both datasets and expert reasoning. As a result, both domain expertise and the perspectives of affected populations must be engaged empirically to ensure that decision criteria are meaningful and aligned with real-world implications. Within DARE, several mechanisms can support such incorporation without guaranteeing it: the collaborative deliberation layer can surface differences in assumptions, expertise, and values across stakeholders; the review phase can incorporate feedback on consequences, critique, and newly surfaced perspectives; the iterative criterion-formation loop can revise decision logic in response to empirically grounded insights; and the synthesis of criteria into higher-level concepts can help reconcile heterogeneous perspectives into shared, interpretable representations. Future work should therefore examine how DARE can be coupled with participatory and longitudinal methods, including stakeholder elicitation, qualitative feedback integration, and impact-sensitive evaluation, so that empathy is grounded in situated evidence about what the data mean, whose perspectives are represented, and how decisions affect different populations.

VII. Conclusion

We introduced DARE, an explainable AI–visualization framework for ill-defined decision making that emphasizes ***D**eliberation*, ***A**gency*, ***R**esilience*, and ***E**mpathy* as guiding design principles. By integrating weak supervision with concept-based modeling, DARE bridges statistical learning and human reasoning, enabling decision makers to iteratively articulate, refine, and align decision criteria with AI behavior. Through visualization, users can inspect and adjust data- and concept-level reasoning, transforming AI from a static predictor into an adaptive collaborator. While future work will further evaluate empirical usability, scalability, and alignment across domains, DARE contributes a conceptual and technical foundation for developing transparent, revisable, and participatory decision-support systems grounded in real-world complexity.

We focused on instantiating the DARE framework through the methodological synthesis of weak supervision and concept-based modeling, but DARE reflects a broader shift from full automation to systems that extend human reasoning. DARE can inform frameworks, systems, and tools across AI paradigms, orienting their design around the principles of ***D**eliberation*, ***A**gency*, ***R**esilience*, and ***E**mpathy*. DARE can extend beyond decision making to tasks such as sensemaking or hypothesis generation, among others, where understanding evolves through interaction rather than prescription. ***Empathy*** within DARE requires engagement with empirical contexts and participants. Any system aspiring to embody DARE must be grounded in empirical investigation, attentive to the perspectives of both decision makers and those affected by their decisions, as well as to the social and domain conditions shaping its application. Beyond technical design, DARE invites reflection on how intelligence, human or artificial, can ethically uphold the slow and imperfect work of shared reasoning in an uncertain world.

VIII. Acknowledgments

We thank C. Fountos for helping improve the aesthetic quality of Fig. 1 and our reviewers for their valuable feedback. The work has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101206814.

References

[1] B. Shneiderman, *Human-centered AI*. Oxford University Press, 2022.
[2] European Commission, "Regulation of the European Parliament and of the council laying down harmonised rules on AI (AI Act)," 2021.
[3] K. Kucher, E. Zohrevandi, and C. A. L. Westin, "Towards VA for explainable AI in industrial applications," *Analytics*, 2025.
[4] E. Dimara and J. Stasko, "A critical reflection on visualization research: Where do decision making tasks hide?" *IEEE TVCG*, 2021.
[5] Y. Jiang, S. Fan, Y. Zhu, L. Wang, K. Ye, J. Zhou, L. Zhang, Z. Wang, L. Wu, and P.-L. P. Rau, "A human-centered algorithmic management framework: A literature review," in *HCI Int.* Springer, 2024.
[6] E. Dimara, H. Zhang, M. Tory, and S. Franconeri, "The unmet data visualization needs of decision makers within organizations," *IEEE TVCG*, 2021.
[7] F. Emmert-Streib and M. Dehmer, "Taxonomy of ML paradigms: A data-centric perspective," *WIREs Data Min. Knowl. Discov.*, 2022.
[8] D. Collaris and J. J. van Wijk, "StrategyAtlas: Strategy analysis for ML interpretability," *IEEE TVCG*, 2023.
[9] J. Liu, H. Chen, J. Shen, and K.-K. R. Choo, "FairCompass: Operationalizing fairness in ML," *IEEE TAI*, 2025.
[10] S. Pajer, M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Möller, and H. Piringer, "WeightLifter: Visual weight space exploration for MCDM," *IEEE TVCG*, 2017.
[11] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "LineUp: Visual analysis of multi-attribute rankings," *IEEE TVCG*, 2013.
[12] D. Weng, H. Zhu, J. Bao, Y. Zheng, and Y. Wu, "HomeFinder Revisited: Finding ideal homes with reachability-centric MCDM," in *CHI*. ACM, 2018.
[13] C. Gomez, S. M. Cho, S. Ke, C.-M. Huang, and M. Unberath, "Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review," *Front. Comput. Sci.*, 2025.
[14] M. Vaccaro, A. Almaatouq, and T. Malone, "When combinations of humans and AI are useful: A systematic review and meta-analysis," *Nat. Hum. Behav.*, 2024.

[15] S. Bankins, A. C. Ocampo, M. Marrone, and S. L. Restubog, "A multilevel review of AI in organizations: Implications for organizational behavior research and practice," *J. Organ. Behav.*, 2024.

[16] S. Alon-Barkat and M. Busuioc, "Human–AI interactions in public sector decision making: Automation bias and selective adherence to algorithmic advice," *J. Public Adm. Res. Theory*, 2023.

[17] J. Zhang, C.-Y. Hsieh, Y. Yu, C. Zhang, and A. Ratner, "A survey on programmatic WS," arXiv, 2022.

[18] LCM Team *et al.*, "Large Concept Models: Language modeling in a sentence representation space," arXiv, 2024.

[19] N. Bressa, J. Louis, W. Willett, and S. Huron, "Input visualization: Collecting and modifying data with visual representations," in *CHI*. ACM, 2024.

[20] C. D. Brumar, S. Molnar, G. Appleby, K. Potter, and R. Chang, "A typology of decision-making tasks for visualization," *IEEE TVCG*, 2025.

[21] T. Alves, T. Delgado, J. Henriques-Calado, D. Gonçalves, and S. Gama, "Exploring the role of conscientiousness on visualization-supported decision-making," *C&G*, 2023.

[22] B. Oral, P. Dragicevic, A. Telea, and E. Dimara, "Decoupling judgment and decision making: A tale of two tails," *IEEE TVCG*, 2024.

[23] B. Oral, R. Chawla, M. Wijkstra, N. Mahyar, and E. Dimara, "From information to choice: A critical inquiry into visualization tools for decision making," *IEEE TVCG*, 2024.

[24] L. Cibulski, H. Mitterhofer, T. May, and J. Kohlhammer, "PAVED: Pareto front visualization for engineering design," *CGF*, 2020.

[25] S. Afzal, R. Maciejewski, and D. S. Ebert, "VA decision support environment for epidemic modeling and response evaluation," in *VAST*. IEEE, 2011.

[26] G. Carenini and J. Loyd, "ValueCharts: Analyzing linear models expressing preferences and evaluations," in *AVI*. ACM, 2004.

[27] L. Cibulski and S. Bruckner, "Towards understanding decision problems as a goal of visualization design," in *IEEE Vis*, 2025.

[28] B. Oral, A. Chatzimparmpas, Y. Zhang, L. van Dijk, R. Võeras, and E. Dimara, "Exploring visualization support for early-career decision making," *Information Visualization*, 2026.

[29] M. Sugihara, S. Takakai, K. Takamatsu, and K. Misue, "Contribution of data visualization to decision-making: A classification of data visualization research based on the characteristics of decision problems," in *PacificVis*. IEEE, 2025.

[30] Y. Ahn and Y.-R. Lin, "FairSight: VA for fairness in decision making," *IEEE TVCG*, 2020.

[31] G. Andrienko, N. Andrienko, and U. Bartling, "VA approach to user-controlled evacuation scheduling," in *VAST*. IEEE, 2007.

[32] J. Müller, M. Cypko, A. Oeser, M. Stoehr, V. Zebralla, S. Schreiber, S. Wiegand, A. Dietz, and S. Oeltze-Jafra, "Visual assistance in clinical decision support," in *EuroVis*. Eurographics, 2021.

[33] S. Guo, F. Du, S. Malik, E. Koh, S. Kim, Z. Liu, D. Kim, H. Zha, and N. Cao, "Visualizing uncertainty and alternatives in event sequence predictions," in *CHI*. ACM, 2019.

[34] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert, "Podium: Ranking data using mixed-initiative VA," *IEEE TVCG*, 2018.

[35] OpenAI, "Gpt-4 technical report," arXiv, 2024.

[36] DeepSeek-AI, "Deepseek-R1: Incentivizing reasoning capability in LLMs via RL," arXiv, 2025.

[37] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," arXiv, 2021.

[38] Qwen Team, "Qwen2.5-vl," arXiv, 2025.

[39] M. Angelini, G. Blasilli, S. Lenti, and G. Santucci, "A VA conceptual framework for explorable and steerable partial dependence analysis," *IEEE TVCG*, 2024.

[40] S. Laguna, J. N. Heidenreich, J. Sun, N. Cetin, I. Al-Hazwani, U. Schlegel, F. Cheng, and M. El-Assady, "ExpLIMEable: An exploratory framework for LIME," in *XAI in Action*, 2023.

[41] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, 2001.

[42] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the predictions of any classifier," in *KDD*. ACM, 2016.

[43] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, "explAIner: A VA framework for interactive and explainable ML," *IEEE TVCG*, 2020.

[44] D. Sacha, M. Kraus, D. A. Keim, and M. Chen, "VIS4ML: An ontology for VA assisted ML," *IEEE TVCG*, 2019.

[45] H. T. T. Nguyen, L. P. T. Nguyen, and H. Cao, "XEdgeAI: A human-centered industrial inspection framework with data-centric explainable edge AI approach," *Inf. Fusion*, 2025.

[46] S. Gathani, Z. Liu, P. J. Haas, and C. Demiralp, "What-if analysis for business professionals: Current practices and future opportunities," in *CHI*. ACM, 2025.

[47] D. Sacha *et al.*, "What you see is what you can change: Human-centered ML by interactive visualization," *Neurocomput.*, 2017.

[48] M. El-Assady, V. Gold, A. Hautli-Janisz, W. Jentner, M. Butt, K. Holzinger, and D. Keim, "VisArgue: A visual text analytics framework for the study of deliberative communication," in *PolText*, 2016.

[49] F. Sperrle, D. Ceneda, and M. El-Assady, "Lotse: A practical framework for guidance in VA," *IEEE TVCG*, 2022.

[50] S. Monadjemi, M. Guo, D. Gotz, R. Garnett, and A. Ottley, "Human–computer collaboration for VA: An agent-based framework," *CGF*, vol. 42, no. 3, pp. 199–210, 2023.

[51] A. Wong *et al.*, "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients," *JAMA Int. Med.*, vol. 181, no. 8, pp. 1065–1070, 08 2021.

[52] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," in *Ethics of Data and Anal.* Auerbach Public., 2022.

[53] F. L. Dennig, M. Miller, D. A. Keim, and M. El-Assady, "FS/DS: A theoretical framework for the dual analysis of feature space and data space," *IEEE TVCG*, 2024.

[54] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins, "Progressive learning of topic modeling parameters: A VA framework," *IEEE TVCG*, 2018.

[55] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for VA," *IEEE TVCG*, 2014.

[56] J. Kandel, J. Liu, A. Z. Wang, C. Tseng, and D. Szafir, "Graphical perception of icon arrays versus bar charts for value comparisons in health risk communication," arXiv, 2025.

[57] K. E. Weick, *Sensemaking in Organizations*. Sage Publications, 2010.

[58] H. A. Simon, *The New Science of Management Decision*. Harper & Brothers, 1960.

[59] S. French, J. Maule, and N. Papamichail, *Decision Behaviour, Analysis and Support*. Cambridge University Press, 2009.

[60] B. Fischhoff and P. Slovic, *Decisions: Studying and Supporting People Facing Hard Choices*. The MIT Press, 2025.

[61] H. Mintzberg, D. Raisinghani, and A. Theoret, "The structure of unstructured decision processes," *Adm. Sci. Q.*, 1976.

[62] T. Munzner, *Visualization Analysis and Design*. A K Peters / CRC Press, 2014.

[63] IEEE VIS, "VIS 2025 area model for paper submissions," https://ieeevis.org/year/2025/info/call-participation/area-model, 2025.

[64] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, 2013.

[65] R. H. Sprague, "A framework for the development of decision support systems," *MIS Q.*, 1980.

[66] S. Gupta, S. Modgil, S. Bhattacharyya, and I. Bose, "Ai for decision support systems in the field of operations research: Review and future scope of research," *Annals of Operations Research*, vol. 308, no. 1, pp. 215–274, 2022.

[67] B. Shavazipour, M. López-Ibáñez, and K. Miettinen, "Visualizations for decision support in scenario-based multiobject. optimiz." *Inf. Sci.*, 2021.

[68] P. A. Alvarez, A. Ishizaka, and L. Martínez, "Multiple-criteria decision-making sorting methods: A survey," *Expert Syst. Appl.*, 2021.

[69] K. Miettinen, "Survey of methods to visualize alternatives in multiple criteria decision making problems," *OR Spectr.*, 2014.

[70] M. Aruldoss, T. M. Lakshmi, and V. P. Venkatesan, "A survey on MCDM methods and its applications," *Am. J. Inf. Syst.*, 2013.

[71] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Comput. Surv.*, 2020.

[72] Z. Zhou, W. Wang, M. Guo, Y. Wang, and D. Gotz, "A design space for surfacing content recommendations in visual analytic platforms," *IEEE TVCG*, 2023.

[73] M. A. Chatti, M. Guesmi, and A. Muslim, "Visualization for recommendation explainability: A survey and new perspectives," *ACM TiiS*, 2024.

[74] L. J. Savage, *The Foundations of Statistics*. Dover Publications, 2012.

[75] B. Fischhoff and S. B. Broomell, "Judgment and Decision Making," *Annual Review of Psychology*, vol. 71, no. 1, pp. 331–355, Jan. 2020.

[76] M. Friedman and L. J. Savage, "The utility analysis of choices involving risk," *J. Polit. Econ.*, 1948.

[77] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, 1974.

[78] G. Gigerenzer and W. Gaissmaier, "Heuristic decision making," *Annu. Rev. Psychol.*, 2011.

[79] K. M. Eisenhardt and M. J. Zbaracki, "Strategic decision making," *Strateg. Manag. J.*, 1992.

[80] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior (60th Anniversary ed.)*. Princeton University Press, 2007.
[81] G. Klein, "Naturalistic decision making," *Hum. Factors*, 2008.
[82] A. Chatzimparmpas and E. Dimara, "Aiding humans in financial fraud decision making: Toward an XAI-visualization framework," in *IEEE VIS Posters*, 2024.
[83] L. W. Ge, M. Easterday, M. Kay, E. Dimara, P. Cheng, and S. L. Franconeri, "V-FRAMER: Visualization framework for mitigating reasoning errors in public policy," in *CHI*. ACM, 2024.
[84] B. Oral, R. Võeras, and E. Dimara, "Iterative quantification of categorical criteria for enhanced job seeking," in *IEEE VIS Posters*, 2024.
[85] M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, 2016.
[86] E. A. Huerta *et al.*, "FAIR for AI: An interdisciplinary and international community building perspective," *Sci. Data*, 2023.
[87] M. Correll, "Ethical dimensions of visualization research," in *CHI*. ACM, 2019.
[88] E. Dimara and C. Perin, "What is interaction for data visualization?" *IEEE TVCG*, 2020.
[89] Πλάτων, *Πολιτεία, 380 π.Χ.* Αθήνα: Κάκτος, 2022.
[90] J. Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, 1999.
[91] J. S. Dryzek, *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford University Press, 2009.
[92] N. Boukhelifa, M.-E. Perrin, S. Huron, and J. Eagan, "How data workers cope with uncertainty: A task characterisation study," in *CHI*. ACM, 2017.
[93] C. S. Holling, "Resilience and stability of ecological systems," *Annu. Rev. Ecol. Syst.*, 1973.
[94] R. Lempert, S. Popper, and S. Bankes, *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*. RAND Corporation, 2003.
[95] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, "A survey of RL from human feedback," arXiv, 2023.
[96] D. Li, Z. Wang, Y. Chen, R. Jiang, W. Ding, and M. Okumura, "A survey on deep AL: Recent advances and new frontiers," *IEEE TNNLS*, 2025.
[97] C. O. Retzlaff *et al.*, "HITL RL: A survey and position on requirements, challenges, and opportunities," *J. Artif. Intell. Res.*, 2024.
[98] E. E. Absalom, Y.-S. Ho, O. S. Egwuche, O. S. Ekundayo, A. V. D. Merwe, A. K. Saha, and J. Pal, "Classical ML: Seventy years of algorithmic learning evolution," arXiv, 2024.
[99] H. Lee, S. Phatale, H. Mansoor, K. R. Lu, T. Mesnard, J. Ferret, C. Bishop, E. Hall, V. Carbune, and A. Rastogi, "RLAIF: Scaling RL from human feedback with AI feedback," OpenReview preprint, 2024.
[100] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *NeurIPS*, 2022.
[101] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.
[102] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.
[103] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.
[104] J. Huang *et al.*, "Foundation models and intelligent decision-making: Progress, challenges, and perspectives," *The Innovation*, 2025.
[105] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "LLMs: A survey," arXiv, 2024.
[106] M. Xu *et al.*, "A survey of resource-efficient LLMs and multimodal foundation models," arXiv, 2024.
[107] C.-E. Sun, T. Oikarinen, B. Ustun, and T.-W. Weng, "Concept bottleneck LLMs," in *ICLR*, 2025.
[108] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with WS," in *VLDB*, 2017.
[109] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *ICML*. PMLR, 2020.
[110] V. Choudhary, A. Marchetti, Y. R. Shrestha, and P. Puranam, "Human–AI ensembles: When can they work?" *J. Manage.*, 2025.
[111] Y. R. Shrestha, S. M. Ben-Menahem, and G. Von Krogh, "Organizational decision-making structures in the age of AI," *Calif. Manage. Rev.*, 2019.
[112] S. Gao and D. Xu, "Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering," *Expert Syst. Appl.*, 2009.
[113] E. Altman, J. Blanuša, L. V. Niederhäusern, B. Egressy, A. Anghel, and K. Atasu, "Realistic synthetic financial transactions for anti-money laundering models," in *NeurIPS Datasets and Benchmarks*, 2023.
[114] N. H. Wong, C. L. Tan, D. D. Kolokotsa, and H. Takebayashi, "Greenery as a mitigation and adaptation strategy to urban heat," *Nat. Rev. Earth Environ.*, 2021.
[115] A. Middel, S. AlKhaled, F. A. Schneider, B. Hagen, and P. Coseo, "50 grades of shade," *Bull. Am. Meteorol. Soc.*, 2021.
[116] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers, "The state-of-the-art of set visualization," *CGF*, 2016.
[117] M. Gleicher, "Considerations for visualizing comparison," *IEEE TVCG*, 2018.
[118] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister, "Hybrid intelligence," *Bus. Inf. Syst. Eng.*, 2019.
[119] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE TVCG*, 2020.
[120] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in ML," *ACM Comput. Surv.*, 2022.
[121] S. K. Card and J. D. Mackinlay, "The structure of the information visualization design space," in *InfoVis*, 1997, pp. 92–99.
[122] A. Chatzimparmpas, K. Kucher, and A. Kerren, "Visualization for trust in ML revisited: The state of the field in 2023," *IEEE CG&A*, 2024.
[123] A. Abusitta, M. Q. Li, and B. C. M. Fung, "Survey on explainable AI: Techniques, challenges, and open issues," *Expert Syst. Appl.*, 2024.
[124] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The state of the art in enhancing trust in ML models with the use of visualizations," *CGF*, 2020.
[125] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren, "A SoS on the use of visualization for interpreting ML models," *Inf. Vis.*, 2020.
[126] C. North *et al.*, "Analytic provenance: Process+interaction+insight," in *CHI EA*. ACM, 2011, p. 33–36.
[127] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in va," *IEEE TVCG*, vol. 22, no. 1, pp. 240–249, 2016.
[128] M. Sun *et al.*, "Toward systematic considerations of missingness in VA," in *IEEE VIS*, 2022, pp. 110–114.
[129] F. Cheng, Y. Ming, and H. Qu, "DECE: Decision explorer with counterfactual explanations for ML models," *IEEE TVCG*, vol. 27, no. 02, pp. 1438–1447, Feb. 2021.
[130] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *ICML*, 2018.
[131] S. Zhang, H. Li, H. Qu, and Y. Wang, "AdaVis: Adaptive and explainable visualization recommendation for tabular data," *IEEE TVCG*, 2024.
[132] N. McCurdy, J. Gerdes, and M. Meyer, "A framework for externalizing implicit error using visualization," *IEEE TVCG*, 2019.
[133] H. Lin, D. Akbaba, M. Meyer, and A. Lex, "Data Hunches: Incorporating personal knowledge into visualizations," *IEEE TVCG*, 2023.
[134] R. Smith, J. A. Fries, B. Hancock, and S. H. Bach, "Language models in the loop: Incorp. prompting into WS," *ACM/IMS J. Data Sci.*, 2024.
[135] E. Hsu and K. Roberts, "Leveraging LLMs for knowledge-free WS in clinical natural language processing," *Sci. Rep.*, 2025.
[136] D. Debot, P. Barbiero, F. Giannini, G. Ciravegna, M. Diligenti, and G. Marra, "Interpretable concept-based memory reasoning," *arXiv*, 2024.
[137] Y. Lin, S. Wei, H. Zhang, D. Qu, and J. Bai, "An interactive visual enhancement for prompted programmatic WS in text classification," *CGF*, 2025.

Angelos Chatzimparmpas is Tenured Assistant Professor at Utrecht University. His fields of research are Information Visualization and Visual Analytics. His main research interests include visual exploration of the inner parts and the quality of machine learning (ML) models with a specific focus on making complex ML models better understandable and explainable, as well as providing reliable trust in the ML models and their results.

Evanthia Dimara is Tenured Assistant Professor at Utrecht University. Her fields of research are Information Visualization and Human-Computer Interaction. Her focus is on decision making – how to help people make unbiased and informed decisions alone or in groups. She is especially interested in the kinds of decisions for which the current decision-support systems, models, and people's heuristics tend to fail.