

ORIGINAL RESEARCH

Open access

From Prediction to Intervention: A Causal Explainability Framework for Counterfactual Reasoning, Mechanism Discovery, Confounding Control, and Decision Validity

Hassan Al-Rashid¹, Julia Schmidt^{2*}, Peter O'Connor³

Abstract

Explainable artificial intelligence increasingly supports decisions whose consequences extend beyond prediction, yet many explanations remain grounded in associations, feature contributions, saliency, or local perturbations. Such outputs may clarify model behaviour without establishing what would happen under intervention, why an outcome arose through a particular mechanism, or whether a proposed change is feasible and decision-valid. This article develops a conceptual causal explainability framework for moving from prediction-oriented explanation toward intervention-aware reasoning. The framework begins by separating predictive, associative, and model-behaviour claims from causal questions. It then connects explicit causal assumptions with causal structures, mechanistic reasoning, interventional explanation, counterfactual reasoning, confounding assessment, identifiability, sensitivity, external validity, uncertainty communication, and human interpretation. Particular attention is given to the distinction between model counterfactuals and real-world recourse, the risks created by unmeasured confounding, selection and collider bias, and the need to define decision boundaries before explanations inform action. Human users are treated as active interpreters whose causal literacy, expertise, cognitive constraints, and susceptibility to false causal confidence affect explanation usefulness. The main contribution is an integrated framework that organizes the evidential transition from prediction to intervention while preserving explicit no-guarantee boundaries concerning causal truth, intervention validity, recourse, identifiability, human understanding, decision validity, and deployment readiness. The framework is intended to support disciplined conceptual development and evaluation planning, not to prescribe a universal causal model, intervention procedure, or operational decision system.

Keywords Explainable AI, Causal explainability, Counterfactual reasoning, Causal inference, Confounding, Identifiability

*Correspondence:

Julia Schmidt

julia.schmidt@unisg.ch

¹ Department of Causal AI and Decision Science, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

² Department of Computer Science, University of St. Gallen, St. Gallen, Switzerland

³ Department of Information Systems, University College Dublin, Dublin, Ireland

Introduction

Explainable artificial intelligence has developed into a broad field concerned with transparency, interpretability, accountability, and responsible use. Its methods range from intrinsically interpretable models to post-hoc feature attributions, examples, rules, visualisations, and local approximations. This diversity is valuable, but it also creates a recurrent ambiguity: an explanation of how a model produced a prediction may be interpreted as an explanation of why

the represented event occurred. A causal explainability agenda must therefore begin by distinguishing model-relative explanation from claims about the data-generating process [1].

That distinction is also human-centred. People rarely seek exhaustive descriptions of every factor associated with an output; they commonly ask selective and contrastive questions such as why one outcome occurred rather than another, or what would have changed it. The usefulness of an explanation consequently depends on the contrast being addressed, the user's knowledge, and the social setting in which the explanation is communicated. Causal explainability must preserve these interpretive features without allowing intuitive narratives to outrun evidential support [2].

Machine-learning systems are usually optimised to exploit stable statistical regularities in observed data. Even highly accurate representations may fail when environments change, interventions alter the generating process, or relationships that were predictive cease to be invariant. Moving from prediction toward intervention therefore requires assumptions about causal variables, mechanisms, and contextual stability that are not supplied by predictive performance alone. Causal representation and mechanism-oriented reasoning provide a stronger conceptual basis, but they remain dependent on explicit assumptions and validation [3].

This article proposes a causal explainability framework that organises this transition without claiming that causal explanations reveal true causes automatically. It links the review of correlational explanations to causal-question formulation, assumption specification, mechanism mapping, intervention and counterfactual reasoning, confounding and identifiability assessment, uncertainty communication, human interpretation, and decision-validity boundaries. Counterfactuals and causability are treated as connected but non-equivalent: a generated alternative may support contrastive understanding while still lacking causal, feasible, or consequential validity [4].

Limitations of Correlational Explanation

Predictive explanations can show which variables, patterns, or regions of an input influence a model's output. Local surrogates, rule approximations, examples, global feature rankings, and perturbation methods can therefore make black-box behaviour more inspectable. Their explanatory object, however, is normally the trained model rather than the world that generated the data. Unless causal assumptions are added and justified, these methods cannot establish direction, mechanism, manipulability, or the effect of changing a feature in practice [5].

The gap becomes especially important in high-stakes settings. A post-hoc account may appear coherent while remaining incomplete, unstable, or disconnected from the model's actual decision process. It can also divert scrutiny from whether an opaque model should be used at all. Interpretability and causal explainability are not identical, but both require a no-sufficiency boundary: a plausible explanation is not by itself evidence that a decision is valid, fair, safe, or appropriate [6].

Saliency, attention, and feature importance can intensify false causal confidence because they visually or numerically foreground selected inputs. A highlighted region or highly ranked feature may be predictive, correlated with a latent cause, or functioning as a proxy; none of these possibilities establishes that intervening on it will alter the outcome. In clinical decision support, this gap can make post-hoc explanations appear actionable even when they do not demonstrate causal relevance, external validity, or benefit to the affected decision [7].

More elaborate attribution does not remove this boundary. Propagating Shapley values through linked models can improve traceability across a predictive pipeline and show how intermediate outputs contribute to a final prediction. Yet the resulting decomposition remains tied to model dependencies and background-data choices. It should not be interpreted as a discovered causal pathway, an identified intervention effect, or proof that changing the attributed variable will produce the predicted consequence. Correlational explanation can support inspection, but not a prediction-to-intervention guarantee [8].

Causal explainability begins with a well-formed causal question. It must specify the putative cause or action, comparator, outcome, population, time horizon, and decision context. This differs from asking which variables predict an outcome or which features influence a model. Causal learning further separates structure discovery, effect estimation, counterfactual inference, and validation; combining these tasks under a single explanation label can conceal distinct assumptions and failure modes [9].

Mechanism discovery is likewise different from mechanism verification. Data-driven discovery methods may generate candidate directions or equivalence classes, whereas causal inference asks whether a defined effect can be recovered under stated assumptions. Tool-supported workflows can improve transparency, but method outputs remain sensitive to sampling, measurement, temporal information, prior restrictions, and model class. A framework should therefore document whether a relation is assumed, learned, externally supported, contested, or not identifiable rather than presenting every edge as established [11].

Mediators, moderators, and causal pathways help clarify how an effect may operate and for whom it may differ, but they must be linked to the estimand and temporal process under examination. Adjustment for a mediator may remove part of the effect of interest, while an unrecognised moderator may make a population-average explanation misleading for a subgroup. **Table 1** summarises why correlational explanations require causal structures, causal assumptions, and mechanistic reasoning before they can support stronger causal or intervention-oriented claims.

Explanation or causal requirement	Core explanation question	What it can support	What it cannot establish	Failure or overclaiming risk	Validation or assumption need	Decision-validity relevance	Framework role
Prediction	What outcome is estimated?	Expected model output	Effect of acting	Forecast treated as prescription	Target and context validation	High	Starting point
Association or correlation	Which variables co-vary?	Statistical dependence	Direction or mechanism	Spurious causal wording	Structural and temporal assumptions	High	Non-causal evidence

Feature importance	Which input influences output?	Model-relative contribution	Real-world causal leverage	Importance equated with cause	Causal-role assessment	High	Behaviour inspection
Saliency or attention	Where is model sensitivity focused?	Local or internal emphasis	Faithful mechanism or effect	Persuasive visual overclaim	Stability and faithfulness tests	Moderate to high	Scope-limited explanation
Post-hoc or local explanation	Why did the model respond locally?	Approximate decision logic	Data-generating process	Rationalisation	Fidelity and robustness review	High	Correlational review
Global ranking	What matters across cases?	Aggregate dependence	Individual effect	Population-to-person transfer	Heterogeneity assessment	High	Global boundary
Proxy-variable review	Is a feature standing for another construct?	Possible indirect dependence	Legitimate intervention target	Proxy treated as mechanism	Construct and pathway validation	High	Measurement checkpoint
Causal assumptions	What conditions enable the claim?	Transparent claim basis	Truth of assumptions	Hidden dependence	Explicit assumption register	Critical	Mandatory input
Causal graph or structural model	Which pathways are hypothesised?	Organised causal reasoning	Verified true structure	Model reification	Domain and empirical review	Critical	Structure layer
Causal pathway	How may an effect operate?	Process hypothesis	Proven mechanism	Plausible story treated as fact	Temporal and mechanistic evidence	High	Mechanism mapping
Mechanism discovery	What structures are data-compatible?	Candidate mechanisms	Unique causal truth	Equivalence ignored	External or interventional validation	High	Hypothesis generation
Mediator	What transmits an effect?	Indirect-path reasoning	Valid mediation without assumptions	Overadjustment	Estimand-specific analysis	High	Pathway annotation
Moderator	When or for whom does an effect vary?	Heterogeneity reasoning	Stable subgroup effect	Fragile subgroup claim	Adequate data and validation	High	Boundary refinement
Confounder	What common cause biases the relation?	Bias-path identification	Complete bias removal	Omission or misclassification	Defensible causal structure	Critical	Validity gate

Domain knowledge	Which relations are plausible?	Contextual constraints	Empirical proof	Authority bias	Diverse expert review	High	Validation resource
No causal-truth boundary	Has causation been proved?	Qualified hypothesis status	Universal causal truth	Framework presented as proof	Continuing validation	Critical	Hard boundary

Interventional and Counterfactual Explanations

An interventional explanation addresses what would happen if an action were imposed, not merely observed. This requires a defined action, comparator, outcome, and population, together with assumptions linking the intervention to the data-generating process. Treatment-effect, action-effect, or policy-intervention language is appropriate only where the corresponding estimand and identification conditions are specified. Predictive accuracy, even when high, does not substitute for counterfactual validation or establish that a recommended action will improve outcomes [13].

Counterfactual explanations often answer a narrower model-level question: what change to an input would alter the model's output? This can provide a useful contrastive explanation, especially for an affected person seeking to understand an adverse decision. It becomes recourse only when the proposed change is controllable, feasible, causally connected to the desired outcome, and likely to remain effective after the system responds. A contrastive example and a consequential recommendation must therefore be separated explicitly [14].

Counterfactual-generation methods operationalise competing criteria such as target validity, proximity, sparsity, diversity, and plausibility. A nearby alternative may be mathematically convenient yet socially costly, temporally impossible, or inconsistent with the joint distribution of features. Conversely, a plausible data point may still violate causal dependencies. These trade-offs mean that no single distance function or benchmark criterion can establish counterfactual quality across users, tasks, and decision contexts [15].

Feasibility constraints should include immutable characteristics, temporal ordering, resource access, institutional rules, indirect effects, and the possibility that a decision system will change after users act. Recourse assessment must also ask whether altering one variable causes downstream changes in others and whether the resulting state remains acceptable. These requirements extend beyond prediction flipping and require human, domain, and causal validation. Accordingly, the framework treats recourse as a separately evaluated claim rather than an automatic property of counterfactual explanation [16].

Finally, counterfactual reasoning should not be equated with causal understanding. A model may support internally coherent what-if statements while misrepresenting the real process, omitting relevant causes, or relying on an unsuitable contrast. Interventional and counterfactual explanations can strengthen causal explainability when their assumptions, feasibility constraints, and uncertainty are explicit, but they retain a no-intervention-validity guarantee and a no-recourse guarantee. Their role is to structure disciplined questions and candidate interpretations, not to prove causation or prescribe action [17].

Confounding, Identifiability, and Decision Validity

Causal explanations derived from observational data require an explicit account of confounding. A confounder is a common cause of the proposed action and outcome, and its role should be determined through substantive causal reasoning rather than correlation, statistical significance, or automated variable selection alone. Directed acyclic graphs can clarify assumed pathways, mediators, colliders, and possible adjustment variables, but their usefulness depends on the completeness and plausibility of the encoded structure [18].

Adjustment can also introduce bias. Conditioning on a collider may open a non-causal path, while controlling for a mediator or post-treatment variable may remove or distort part of the effect being examined. Selection mechanisms can similarly create associations that do not represent the target population or intervention context. A causal explanation should therefore state why variables were included, excluded, or treated as intermediates rather than implying that larger adjustment sets necessarily produce more credible causal claims [19].

Unmeasured confounding remains possible even after a defensible adjustment strategy has been specified. Sensitivity analysis can examine how strongly an omitted variable would need to relate to the proposed cause and outcome to alter a conclusion. Such analysis does not prove that hidden bias is absent, identify the omitted cause, or repair an incorrect causal structure. Its value lies in communicating the dependence of the explanation on assumptions and showing where causal confidence should remain limited [20].

External validity creates an additional boundary. An effect or causal explanation developed in one population, institution, period, or environment may not transfer when baseline risks, mechanisms, intervention access, or effect modifiers differ. Transportability consequently requires an explicit target context and evidence about relevant population differences. Decision usefulness cannot be inferred from internal identification alone when the intended users or circumstances lie outside the evidential setting [21].

The movement from prediction to intervention therefore requires more than an accurate predictive model. It requires a defined decision objective, causal assumptions, mechanism reasoning, confounding assessment, identifiability analysis, uncertainty communication, and explicit limits on when an explanation should guide action. Prescriptive analytics likewise depends on action spaces, constraints, objectives, and consequences rather than prediction alone [22]. **Figure 1** illustrates the prediction-to-intervention logic of causal explainability, showing how correlational explanation must pass through causal assumptions, mechanism reasoning, confounding assessment, identifiability, uncertainty, and decision-boundary checks before supporting intervention-oriented interpretation.

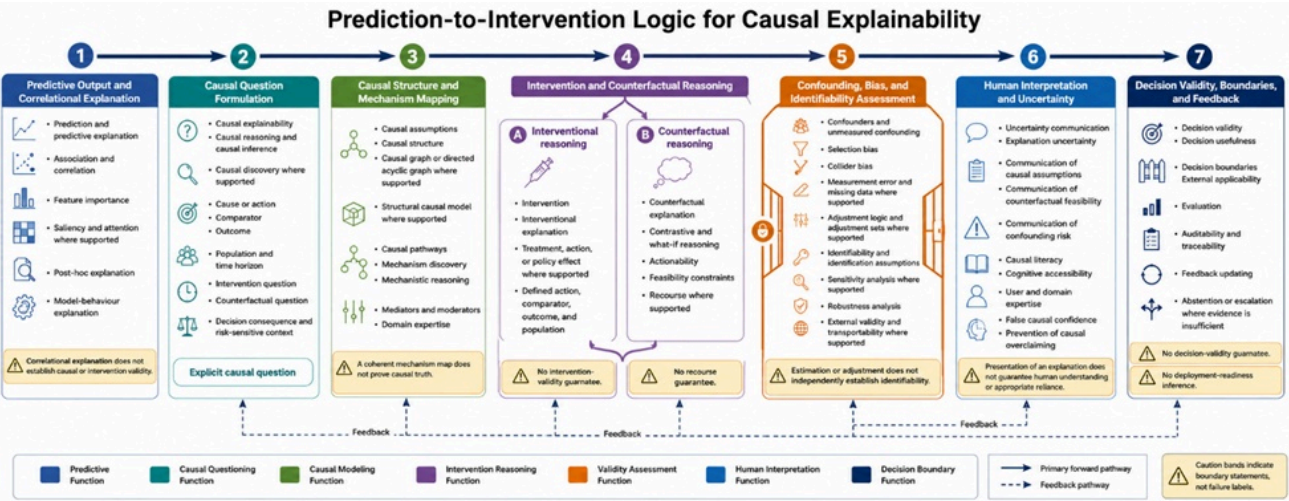


Figure 1. Prediction-to-Intervention Logic for Causal Explainability

Alt-text description.

A conceptual pathway showing prediction and correlational explanation moving through causal assumptions, mechanisms, interventions, counterfactuals, confounding, identifiability, uncertainty, human interpretation, and decision boundaries.

Human Interpretation of Causal Claims

Causal explanations are interpreted by users with different objectives, expertise, responsibilities, and opportunities to contest decisions. A data scientist may need assumptions and sensitivity details, an operator may need escalation conditions, and an affected person may need an accessible account of why an outcome occurred and whether a proposed alternative is feasible. Explanation design and evaluation should therefore begin with stakeholder goals rather than assuming that one representation can satisfy every user [23].

Cognitive accessibility should not be confused with evidential completeness. Graphs, narratives, contrasts, and simplified causal language may improve readability while hiding identification assumptions, omitted variables, or uncertainty. Evaluation should consequently assess fidelity, stability, completeness, usability, and task relevance rather than relying on visual plausibility or anecdotal satisfaction. Existing evaluation research shows that explanation quality is multidimensional and cannot be represented by one universal metric [24].

Human-centred evaluation must also examine reliance. Explanations can increase confidence without improving judgement, and their effects depend on task difficulty, user knowledge, model accuracy, and the information available before the explanation is presented. Appropriate reliance means accepting useful assistance while detecting limitations and resisting erroneous recommendations. It is therefore different from trust, satisfaction, perceived transparency, or stated preference [25].

Counterfactual explanations illustrate this dependency clearly. Combining counterfactual, contrastive, or normative information may affect comprehension and decision performance differently across tasks. Additional explanation can impose cognitive burden, direct attention toward irrelevant changes, or produce an impression of causal certainty unsupported by the evidence. Human-performance testing is therefore needed before claiming that counterfactual explanations improve understanding or decisions [26].

Causal literacy includes the ability to distinguish prediction, association, intervention, and counterfactual claims and to recognise that an identified estimate remains conditional on assumptions. Interfaces should communicate causal assumptions, feasible action constraints, confounding risk, uncertainty, and decision boundaries in forms appropriate to user expertise. Nevertheless, comprehension cannot be presumed from exposure to an explanation, and better comprehension does not guarantee better reliance or outcomes [23]. **Table 2** maps interventional explanations, counterfactual reasoning, confounding control, identifiability, decision validity, and human interpretation requirements for causal explainability.

Table 2. Interventional, Counterfactual, Confounding, Identifiability, Decision-Validity, and Human-Interpretation Requirements for Causal Explainability: What-If Reasoning, Recourse, Feasibility, Confounders, Selection Bias, Collider Bias, Adjustment, Sensitivity, External Validity, Causal Literacy, Uncertainty, and Decision Boundaries

Causal explainability requirement	Core causal or human-centred question	Required evidence or assumption	Causal or interpretation function	Potential output	Failure or residual risk	Evaluation or governance need	F
Intervention definition	What action is imposed?	Clear action, comparator, outcome, and population	Separates observation from manipulation	Qualified intervention question	Ambiguous or unrealistic action	Domain review	It e
Interventional explanation	What may follow from	Identified causal	Explains action–	Conditional causal	Prediction treated as effect	Empirical validation	It

	the action?	estimand	outcome relation	account			
Counterfactual reasoning	What would have occurred otherwise?	Factual case and causal assumptions	Supports why/why-not reasoning	Counterfactual statement	Alternative treated as observable fact	Assumption disclosure	C
Counterfactual validity	Does the alternative alter the target output?	Valid model and target definition	Tests model-level outcome change	Valid alternative under model	Model validity confused with causal validity	External review	C
Counterfactual plausibility	Is the alternative realistic?	Data and causal constraints	Filters incoherent alternatives	Plausibility-qualified case	Realistic-looking but impossible state	Domain validation	C
Actionability	Can the person alter the feature?	Control, timing, access, and resource evidence	Filters unusable changes	Actionable candidate	Immutable or inaccessible change	Stakeholder review	
Recourse	Will feasible action improve the outcome?	Causal response and system stability	Connects explanation with consequence	Candidate recourse pathway	Prediction flips without lasting benefit	Longitudinal evaluation	
Confounding	Are non-causal paths mixed with the target effect?	Causal structure and domain knowledge	Assesses exchangeability	Confounding statement	Residual bias	Transparent adjustment rationale	V
Selection bias	Does sample inclusion distort the relation?	Sampling-process analysis	Tests population relevance	Selection-risk statement	Non-representative effect	Population review	
Collider bias	Has conditioning opened a false pathway?	Graph-based control review	Prevents induced association	Collider warning	Adjustment creates bias	Independent audit	A
Identifiability	Can the target effect be recovered?	Identification assumptions and adequate data	Determines estimability	Identified-under-assumptions status	Estimation without identification	Methodological review	
Sensitivity analysis	Could hidden bias alter the	Plausible violation	Assesses robustness	Sensitivity profile	Narrow scenarios create	Assumption reporting	

	conclusion?	scenarios			reassurance		
External validity	Does the claim apply elsewhere?	Target-context comparison	Defines scope	Applicability statement	Overgeneralisation	Target-population validation	
Causal literacy	Can users distinguish causal claim types?	User-knowledge assessment	Supports correct interpretation	Literacy-adjusted explanation	Causal terminology overpersuades	Comprehension testing	p
Uncertainty communication	How fragile is the explanation?	Statistical and structural uncertainty	Calibrates confidence	Uncertainty profile	Caveats misunderstood	Interface evaluation	f c
Appropriate reliance	Does use improve judgement when warranted?	Behavioural task evidence	Evaluates human–AI interaction	Reliance assessment	Confidence rises without benefit	Controlled evaluation	
Decision validity	Does evidence support the contemplated decision?	Causal, utility, feasibility, and consequence evidence	Connects explanation with action	Bounded decision-support statement	Valid causal claim used for wrong decision	Multidisciplinary review	I
Decision boundary	When should the explanation not guide action?	Scope, uncertainty, and risk thresholds	Establishes abstention conditions	No-use or escalation rule	Boundary ignored	Salience and compliance monitoring	s

Proposed Causal Explainability Framework

The proposed framework begins with use-case and decision-context definition. It identifies the decision, affected parties, intended outcome, risk level, population, and evidence available. It then separates prediction from intervention and reviews existing model explanations as evidence about model behaviour rather than causation. This initial boundary reflects the distinction between technical explainability and human-oriented causability: an output can be interpretable without providing adequate causal understanding [27].

The second stage formulates the causal question and records its assumptions. A candidate causal structure or mechanism map can then represent hypothesised pathways, mediators, moderators, contextual variables, and sources of bias. Graph-based or multimodal representations may help experts integrate heterogeneous evidence, but they remain conceptual structures requiring empirical scrutiny and human evaluation. Their visual coherence must not be interpreted as proof that the represented mechanisms are true [28].

The third stage connects causal structure with intervention and counterfactual design. It asks whether the intended explanation concerns an intervention effect, a model-level counterfactual, a contrastive account, or possible recourse. Causal methods may function as explanatory targets, resources for constructing explanations, or tools for evaluating

The fourth stage evaluates confounding, selection, measurement, identifiability, sensitivity, external validity, and transportability. It then communicates uncertainty and assumptions in forms matched to users' causal literacy and decision roles. Documentation records provenance, structural choices, disputed assumptions, changes, and decision boundaries. These requirements respond to wider XAI challenges involving context, interdisciplinary evaluation, uncertainty, stakeholder adaptation, and responsible use [30].

Table 3. Proposed Causal Explainability Framework for Decision Validity: Use-Case Definition, Prediction-Intervention Distinction, Correlational Explanation Review, Causal Question Formulation, Assumption Specification, Mechanism Mapping, Confounding Assessment, Identifiability, Interventional Explanation, Counterfactual Reasoning, Feasibility, Uncertainty, Human Interpretation, Decision Boundaries, Auditability, Evaluation, and Deployment Boundaries

Framework stage	Core purpose	Required input	Causal explainability function	Potential output	Validation or evaluation need	Failure risk
Use-case and context definition	Specify purpose, users, consequences, and scope	Decision, population, outcome, risk	Establishes explanatory context	Context statement	Stakeholder review	Undefined or shifting objective
Prediction–intervention distinction	Classify the intended claim	Prediction and intended action	Prevents target conflation	Question classification	Methodological review	Prediction used as intervention evidence
Correlational explanation review	Identify what the model explanation shows	Attribution, saliency, rules, examples	Limits evidence to model behaviour	Explanation inventory	Fidelity and stability testing	Association described as cause
Causal question formulation	Define action, comparator, outcome, population, and time	Decision objective	Establishes estimand	Structured causal question	Domain review	Ambiguous target
Assumption specification	Record claim dependencies	Structural and identification assumptions	Makes conditions visible	Assumption register	Plausibility review	Hidden assumptions

Mechanism mapping	Organise candidate causal pathways	Evidence and domain knowledge	Represents causes, mediators, and moderators	Candidate mechanism map	Empirical and expert validation	Plausible narrative treated as truth
Confounding and bias assessment	Identify open bias pathways	Causal structure and sampling process	Assesses exchangeability	Bias-risk map	Graph and domain review	Omitted causes or bad controls
Identifiability assessment	Determine whether the target is recoverable	Estimand, assumptions, and data	Separates identifiable and non-identifiable claims	Identification status	Formal verification	Estimation without identification
Data and measurement review	Assess construct and sample validity	Provenance, measures, missingness	Connects data with causal concepts	Data-validity statement	Reliability and construct review	Proxy or measurement error
Interventional explanation design	Explain qualified action effects	Identified intervention and uncertainty	Supports intervention-aware interpretation	Conditional explanation	Effect and domain validation	Unsupported recommendation
Counterfactual explanation design	Construct relevant contrasts	Factual case, target, and constraints	Supports why/why-not reasoning	Counterfactual explanation	Validity, plausibility, and stability checks	Unrealistic alternative
Feasibility and recourse review	Assess attainability and consequence	Action set and causal dependencies	Separates explanation from recourse	Feasibility-qualified candidate	User and longitudinal evaluation	Impossible or ineffective action
Uncertainty and sensitivity communication	Show fragility and residual risk	Estimates and violation scenarios	Calibrates confidence	Uncertainty profile	Comprehension testing	Caveats hidden or misunderstood
Human interpretation assessment	Test causal literacy and understanding	User roles and explanation format	Evaluates meaning and reliance	Comprehension assessment	Behavioural testing	Satisfaction mistaken for understanding
Decision-validity assessment	Determine whether evidence supports action	Causal evidence, utility, feasibility, risk	Links inference to bounded decision support	Decision-validity statement	Multidisciplinary evaluation	Causal result treated as prescription
Decision-boundary definition	Specify abstention and escalation conditions	Risk and scope thresholds	Prevents unsupported use	Boundary statement	Scenario testing	Boundary overridden

Documentation and auditability	Preserve provenance and responsibility	Versioned assumptions and evidence	Enables traceability	Framework dossier	Completeness review	Untraceable change
Evaluation and feedback updating	Revise assumptions and boundaries	New evidence and outcomes	Supports iterative correction	Updated framework	Prospective evaluation	Feedback entrenches bias
Deployment boundary	Preserve conceptual status	Separate operational evidence	Prevents conceptual-to-operational leap	Research-use designation	Independent implementation testing	Framework presented as validated system

Figure 2 presents the proposed causal explainability framework, integrating prediction-intervention distinction, causal assumptions, mechanism mapping, counterfactual reasoning, confounding assessment, identifiability, human interpretation, decision validity, and feedback updating.

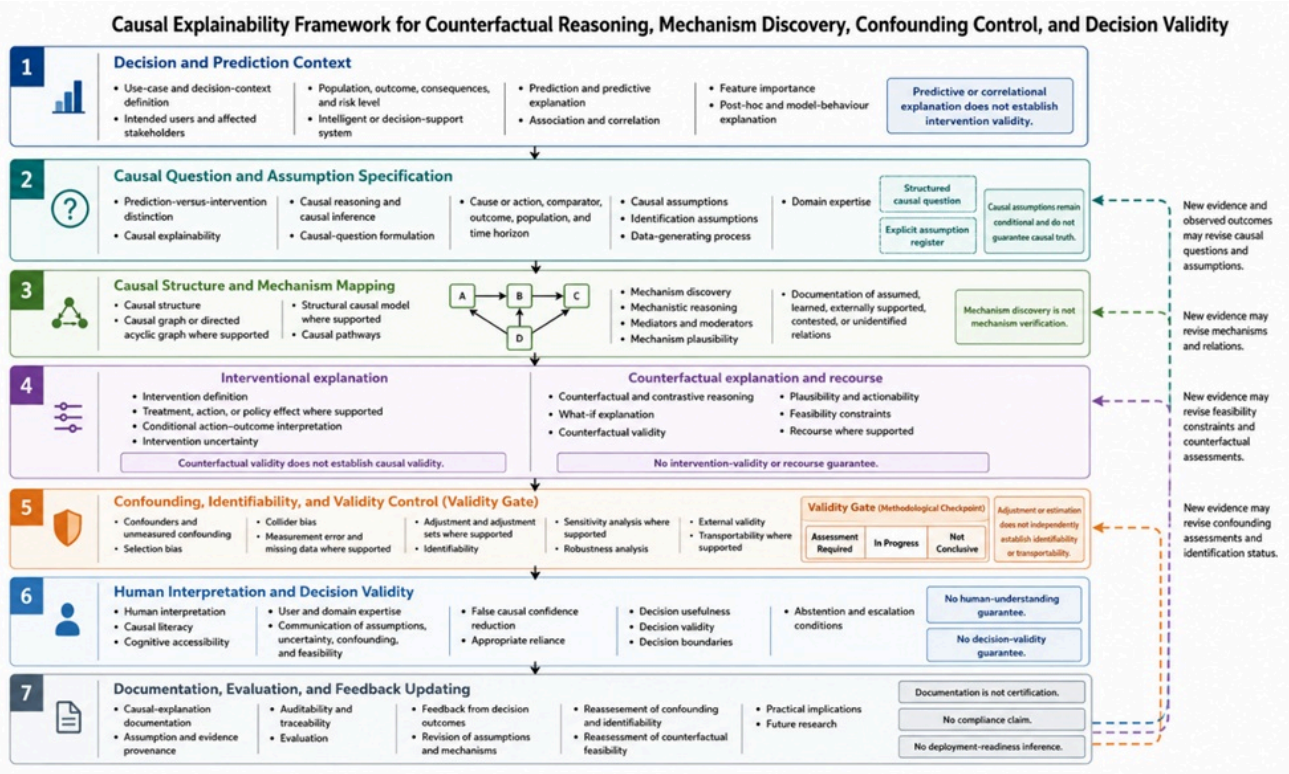


Figure 2. Causal Explainability Framework for Counterfactual Reasoning, Mechanism Discovery, Confounding Control, and Decision Validity

Alt-text description.

A conceptual causal explainability framework showing prediction review, causal questions, assumptions, mechanisms, interventions, counterfactuals, confounding, identifiability, uncertainty, human interpretation, decision validity, auditability, and feedback loops.

Practical Implications and Future Research

A central practical implication is that organisations and researchers should separate explanation-generation capability from causal-claim validity. Feature attributions, saliency displays, counterfactual generators, and fluent narratives should carry explicit scope statements identifying whether they describe model behaviour, statistical association, hypothetical prediction, or an identified causal effect. Future work should test whether such distinctions reduce causal overclaiming without making explanations inaccessible.

Causal assumption transparency and structure validation require further research. Candidate graphs and mechanisms should be evaluated across environments, expert groups, measurement choices, and contradictory evidence. Counterfactual feasibility and recourse assessment should incorporate temporal, economic, institutional, and causal dependencies rather than proximity alone. Sensitivity analysis should also be adapted for explanations so users can see how assumptions or unmeasured confounding could change an interpretation [20].

External validity, transportability, and human interpretation remain particularly underdeveloped. Causal explanations should be studied across populations and decision contexts rather than evaluated only on the data used to construct them [21]. Research should distinguish causal literacy, comprehension, confidence, reliance, and decision performance because improvement in one outcome may not improve another [25]. Foundation-model explanations require especially cautious evaluation because fluent causal narratives may appear mechanistic even when they are generated from statistical regularities.

These implications are research directions rather than operational prescriptions. The framework does not select causal discovery algorithms, adjustment procedures, counterfactual generators, datasets, metrics, platforms, interventions, or deployment processes. It does not provide legal, compliance, procurement, auditing, clinical, financial, safety-engineering, or policy guidance. Its contribution is to identify the assumptions, evidence, human evaluation, and decision boundaries that future causal explainability research should make visible [30].

Conclusion

This article proposed a causal explainability framework for moving from prediction toward intervention by connecting correlational-explanation review, causal-question formulation, explicit assumptions, mechanism mapping, interventional explanation, counterfactual reasoning, feasibility assessment, confounding control, identifiability, sensitivity, uncertainty communication, human interpretation, decision validity, documentation, and feedback updating. The framework emphasises that feature importance, saliency, attention, correlations, local perturbations, and post-hoc plausibility may explain model behaviour without establishing causes or valid actions. Causal explainability should instead be evaluated through transparent assumptions, defensible structures, validity checks, feasible counterfactual constraints, human-understanding assessment, and explicit decision boundaries. Even when these requirements are satisfied, the framework does not independently prove causal truth, guarantee intervention validity or recourse, establish decision validity, ensure appropriate reliance, or demonstrate operational readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 12 Nov 2025

Revised: 29 Dec 2025

Accepted: 16 Jan 2026

Published online: 15 March 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>.

Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38.
<https://doi.org/10.1016/j.artint.2018.07.007>.

Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. *Proc IEEE*. 2021;109(5):612-34.
<https://doi.org/10.1109/JPROC.2021.3058954>.

Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. *Inf Fusion*. 2022;81:59-83.
<https://doi.org/10.1016/j.inffus.2021.11.003>.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018;51(5):1-42.
<https://doi.org/10.1145/3236009>.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.
<https://doi.org/10.1038/s42256-019-0048-x>.

Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11).
[https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).

Chen H, Lundberg SM, Lee SI. Explaining a series of models by propagating Shapley values. *Nat Commun*. 2022;13(1):4512.
<https://doi.org/10.1038/s41467-022-31384-3>.

Guo R, Cheng L, Li J, Hahn PR, Liu H. A survey of learning causality with data: problems and methods. *ACM Comput Surv*. 2020;53(4):1-37.
<https://doi.org/10.1145/3397269>.

Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A. A survey on causal inference. *ACM Trans Knowl Discov Data*. 2021;15(5):1-46.
<https://doi.org/10.1145/3444944>.

Nogueira AR, Pugnana A, Ruggieri S, Pedreschi D, Gama J. Methods and tools for causal discovery and causal inference. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2022;12(2).
<https://doi.org/10.1002/widm.1449>.

Lagemann K, Lagemann C, Taschler B, Mukherjee S. Deep learning of causal structures in high dimensions under data limitations. *Nat Mach Intell*. 2023;5(11):1306-16.
<https://doi.org/10.1038/s42256-023-00744-z>.

Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell*. 2020;2(7):369-75.
<https://doi.org/10.1038/s42256-020-0197-y>.

Karimi AH, Barthe G, Schölkopf B, Valera I. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Comput Surv*. 2022;55(5):1-29.
<https://doi.org/10.1145/3527848>.

Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min Knowl Discov*. 2024;38(5):2770-824.
<https://doi.org/10.1007/s10618-022-00831-6>.

Verma S, Boonsanong V, Hoang M, Hines KE, Dickerson JP, Shah C. Counterfactual explanations and algorithmic recourses for machine learning: a review. *ACM Comput Surv*. 2024;56(12):1-42.
<https://doi.org/10.1145/3677119>.

Baron S. Explainable AI and causal understanding: counterfactual approaches considered. *Minds Mach*. 2023;33(2):347-77.
<https://doi.org/10.1007/s11023-023-09637-x>.

Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*. 2021;50(2):620-32.

Cinelli C, Forney A, Pearl J. A crash course in good and bad controls. *Sociol Methods Res*. 2024;53(3):1071-104.
<https://doi.org/10.1177/00491241221099552>.

Cinelli C, Hazlett C. Making sense of sensitivity: extending omitted variable bias. *J R Stat Soc Series B Stat Methodol*. 2020;82(1):39-67.
<https://doi.org/10.1111/rssb.12348>.

Degtiar I, Rose S. A review of generalizability and transportability. *Annu Rev Stat Appl*. 2023;10:501-24.
<https://doi.org/10.1146/annurev-statistics-042522-103837>.

Bertsimas D, Kallus N. From predictive to prescriptive analytics. *Manage Sci*. 2020;66(3):1025-44.
<https://doi.org/10.1287/mnsc.2018.3253>.

Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans Interact Intell Syst*. 2021;11(3-4):1-45.
<https://doi.org/10.1145/3387166>.

Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Comput Surv.* 2023;55(13s):1-42.
<https://doi.org/10.1145/3583558>.

Wang X, Yin M. Effects of explanations in AI-assisted decision making: principles and comparisons. *ACM Trans Interact Intell Syst.* 2022;12(4):1-36.
<https://doi.org/10.1145/3519266>.

Gentile D, Donmez B, Jamieson GA. Human performance effects of combining counterfactual explanations with normative and contrastive explanations in supervised machine learning for automated decision assistance. *Int J Hum Comput Stud.* 2025;196:103434.
<https://doi.org/10.1016/j.ijhcs.2024.103434>.

Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2019;9(4).
<https://doi.org/10.1002/widm.1312>.

Holzinger A, Malle B, Saranti A, Pfeifer B. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. *Inf Fusion.* 2021;71:28-37.
<https://doi.org/10.1016/j.inffus.2021.01.008>.

Carloni G, Berti A, Colantonio S. The role of causality in explainable artificial intelligence. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2025;15(2).
<https://doi.org/10.1002/widm.70015>.

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Del Ser J, et al. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inf Fusion.* 2024;106:102301.
<https://doi.org/10.1016/j.inffus.2024.102301>.

Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. *Nat Med.* 2024;30(4):958-68.
<https://doi.org/10.1038/s41591-024-02902-1>.