(ρro)Metheus: Probabilistic-Causal Resolution Oracle for Covenant-Bound Artificial Intelligence

Ling Xiao. [1]
Metheus

Abstract

This technical document introduces the (ρro)Metheus Oracle, a protected adjudication environment whereby analysts mobilize:

- audited artificial intelligence (AI) models,
- validated information pipelines,
- to compute *causal relations* from data.

Although (ρro)Metheus is a general purpose computational oracle, its initial deployment targets are the regulators, central/commercial banks, insurers, and other entities serving transnational trade in the Global South. Specifically, outputs of (ρro)Metheus will inform high-valued decisions pertaining to the disbursement of funds and claim resolutions. These are business-critical decisions that require both:

- regulator oversight *a priori* oracle deployment;
- and proper assignment of liabilities *a posteriori* oracle judgment.

This implies every step of the computation inside the (ρro)Metheus Oracle must be properly governed by state-level entities, so that the oracle satisfies the strong guarantee of:

- correct or covenant-compliant computation runs to completion;
- while false or covenant-violating computations fail outright.

This requirement is addressed by statically typed programming languages, where programs that do not terminate fail at compile time. We enforce this design pattern in (ρro)Metheus: at the heart of the oracle is a *domain-specific programming language* called: Typed ρ-Calculus for covenant-bound probabilistic causal intelligence. It is a two-tiered computational model or proof system whereby:

- *in the term universe* or the "analytic" level, (ρro)Metheus uses Pearl's do-Calculus to evaluate causal relations. In particular, the oracle uses information from validated data sources and summaries generated by audited large language models (LLMs) to compute causal proofs.
- *at the type universe* or the compliance level rests a meta-language extended by state regulators. The type-system governs the quality or legality of computation, so that illegal and false computations terminate immediately. Thereby providing a static guarantee that (ρro)Metheus-bound AI agents are well-behaved in accordance to legal constraints.

When verifying programs for contractual compliance, the (ρro)Metheus compiler uses a legalistic analogue of Floyd–Hoare logic, or Loare-Logic to generate proofs witnessing the Covenant-Compliance of Typed ρ-Calculus code written by analysts. This automated proof system shifts the burden of writing legally-compliant programs from analysts to the (ρro)Metheus compiler, thereby extending the operation domain of AI agents into more sensitive and regulated settings.

In other words, if code is law and AI agents are junior employees, then the (ρro)Metheus oracle is populated by a set of law-abiding workers. Meanwhile the Loare-Logic proof system is the enforcer of data and AI validity. The legal content of Loare-Logic is defined in the FairCovenant Foundation, a legislative body governing AI safety in trade along the Global South. Thus Loare-Logic constitutes to the *lingua-franca* of safeAI, and sits appropriately in the public domain as a public good. We brand this computational model distinguished by 1) validated data pipelines, 2) audited AI models used to discover causal relations, and 3) deep regulatory oversight governing the computation graph of the (ρro)Metheus oracle: **Covenant-Compliant Causal-Intelligence**.

[1] lingxiao@metheusai.xyz

Keywords and phrases Machine learning, statistics, artificial intelligence, causality, causal reinforcement learning, formal verifications, deep AI governance, AI safety, safeAI, programming language design, domain specific programming language, legalistic technology.

Contents

1 Introduction

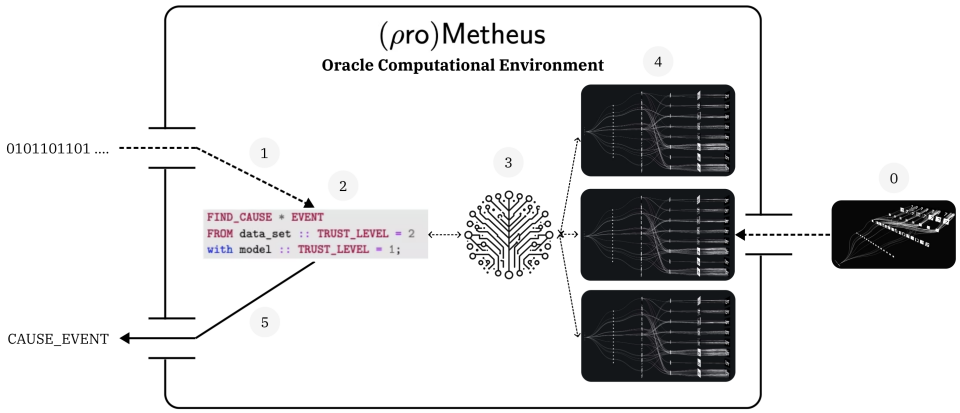
1.1 Technical Stakeholder Perspective

This is a system specification document for the (ρro)Metheus **Provenance Resolution Oracle** for Covenant-Bound Artificial Intelligence. It features:

- A *regulator-compliant* computational environment that enables analysts to run *audited* machine learning (ML) models on *validated* datasets. [2] We refer this set of ML models and datasets as *computational assets.*
- The (ρro)Metheus oracle is especially suitable for recovering casual relationships in datasets in a statistically rigorous manner.
- The theoretical underpinnings of the (ρro)Metheus Oracle is a domain specific programming language named: Typed ρ-Calculus for covenant-bound probabilistic causal intelligence. It combines decades of research and design in programming language theory, classical statistics, and recent advances in machine learning to produce a safer way of interacting with existing AI agents.

1.1.1 Computational Asset Lifecycle in (ρro)Metheus

In concrete terms, (ρro)Metheus is a user interface targeting industry professionals at banks, insurance companies, and regulators.



This user interface is the visible face of the (ρro)Metheus protected computational environment, whereby analysts interact with audited computational assets. The asset lifecycle proceeds as follows:

- (0,1): model and data are audited using stakeholder-defined rules of validation, and subsequently assigned a type, or more precisely a Covenant-Judgement. These legally-tagged computational assets then enter the (ρro)Metheus environment.
- (2) analysts query the data lake with Typed ρ-Calculus in lieu of an LLM-based chat assistant. In particular, they query data with the intention of computing the cause of some real-life `EVENT`.
- (3) (ρro)Metheus compiles the Typed ρ-Calculus language and mobilizes legally-compliant class of data and machine learning models to complete the task.
- (4) (ρro)Metheus draws on open-source and bespoke ML models to compute the likelihood of different causes of the `EVENT`.
- (5) (ρro)Metheus returns probable causes of this `EVENT` to the analyst.

[2] See definition 1 for more on computational environment.

In the figure above, the programming language Typed ρ-Calculus is the top-level user interface. Alternatively we may explore other user interfaces such as:

- A graphical user interface, which may be more appropriate in a consumer setting.
- A natural language interface using LLM's code-generation feature. In this case, we train an LLM to learn a mapping from natural language onto Typed ρ-Calculus.

The advantage of (ρro)Metheus computational environment is three-fold:

- All data and models are audited in accordance to stakeholder-defined laws of validity and security.
- These rule governing validity of information are then propagated along Typed ρ-Calculus's type system from data collection, to model training, fine tuning, and finally inference.
- Additionally, Typed ρ-Calculus has rigorous notions of causality that cannot be expressed in the "classical statistics" underlying machine learning models. Classically trained agents such as auto-regressive models (read: LLMs) can only express correlation, not causation.

The heart of (ρro)Metheus Oracle is the domain specific programming language: Typed ρ-Calculus for covenant-bound probabilistic causal intelligence. Unlike classical statistic language, Typed ρ-Calculus can express causation in addition to correlation.

▶ **Definition 1.** *[Sch13] A **Computational Environment** involves the collection of computer machinery, data storage devices, work stations, software applications, and networks that support the processing and exchange of electronic information demanded by the software solution.*

Next we expand on two use cases of (ρro)Metheus in particular:

- one that targets who query audited data using audited ML models and validated datasets;
- one that targets regulators who define the criteria of valid data and models.

Although they appear to be two orthogonal deployment scenarios, they are in fact part of the same pipeline that takes regulator's legal requirements, and translate them into a compliant artificial intelligence computational environment.

1.1.2 An Example for Analysts

The analysts at banks and insurance companies interact with (ρro)Metheus as a simplified programming language. The language of interface is Typed ρ-Calculus. It is akin to `MySQL`, however instead of rudimentary operations that simply query the database with:

```
SELECT *
FROM employees;
```

The Typed ρ-Calculus under proposal enables analysts to interact with their database in a categorically more sophisticated manner, by seeking the cause of events with:

```
FIND_CAUSE * EVENT
FROM data_set :: TRUST_LEVEL = 2
with model :: TRUST_LEVEL = 1;
```

In this code block, we assume `data_set` and `model` are assets that are pre-loaded in the (ρro)Metheus computational environment. Some additional commentary follows:

1. (line 1): note instead of merely selecting data as is the case with `MySQL`, the analyst uses Typed ρ-Calculus to query the database for the probable cause of `EVENT`.
2. (line 2): the setting `TRUST_LEVEL=2` in line 2 expresses the trustworthiness of the data used to determine the cost of `EVENT` in line 1. In this case `TRUST_LEVEL=2` is a *regulator*-defined trust setting on the validity of the data under consideration. It is effectively a filter on the quality of data used to determine the final cause of `EVENT`.
3. (line 3): the statistical models are also typed with `TRUST_LEVEL=1`, which determines the quality of statistical model used to determine the final cause of `EVENT`. This is critical as the quality of statistical models is a function of:
   - the quality of statisticians training them;
   - the rigor of the assumptions underlying the models;
   - and most importantly, the quality of data the model is trained on.

Naturally models trained on `TRUST_LEVEL=2` data can do no better than attaining a `TRUST_LEVEL=2` itself, as the adage: "garbage in garbage out" attests. [3] The proper propagation of `type` assignment from datasets to models trained on said datasets is a core feature of Typed ρ-Calculus. Now observe that in the Typed ρ-Calculus query above, the quality of dataset is lower than the quality of the model. Then intuitively the final `CAUSE` of `EVENT` determined by this short program should carry `TRUST_LEVEL=2`.

The interaction of datasets and models at inference time is an example of *program composition*. In programming language (PL) parlance, the *inferred type* of the final recommendation correspond to the confidence of the final recommendation in statistical terms. The intellectual core of Typed ρ-Calculus pivots upon this correspondence between type in the PL sense, and a promise or *covenant* of compliance in the legal sense. We term this category of legally-compliant and statically guaranteed artificial intelligence: Covenant-Bound Artificial Intelligence. This correspondence is explored in table (1).

The intellectual core of Typed ρ-Calculus pivots upon the one-to-one correspondence of *covenant* in the legal sense, with inferred the *type* of data in the programming language sense.

1.1.3 An Example for Regulators

In the previous example, we assumed the criteria that determined which computational asset satisfies any particular `TRUST_LEVEL` was a given. In reality, this criteria is determined by regulators. This

[3] For the sake of simplicity we assume `TRUST_LEVEL=2` is less trustworthy than `TRUST_LEVEL=1`.

section provides an example of how regulators would interact with (ρro)Metheus to define the vectors of quality along which computational assets are audited against. Since a regulatory environment with broad stakeholder buy-in is seldomly unilaterally defined, finding consensus on the *criteria to audit* must be an interactive process. Now referencing figure (1), (ρro)Metheus presents a user interface whereby:

1. Multiple parties enter a common covenant proposal environment, so as to submit candidate criteria for community judgement.
2. What follows is an iterative process, whereby the stakeholders debate and elect the most sensible set of criteria to judge data validity and AI model quality.
3. Finally this criteria is reified in the (ρro)Metheus computational environment as a type or Covenant-Judgement.

One example of such a Covenant-Judgement is `TRUST_LEVEL` that will be used by the analysts in the example above. Finally observe that depending on the quality of the dataset or AI model, it may only witness some of the criteria defined in Covenant-Judgement. This fine-grained satisfaction is what give rise to different levels of `TRUST_LEVEL`, i.e. `TRUST_LEVEL` = 1 etc.

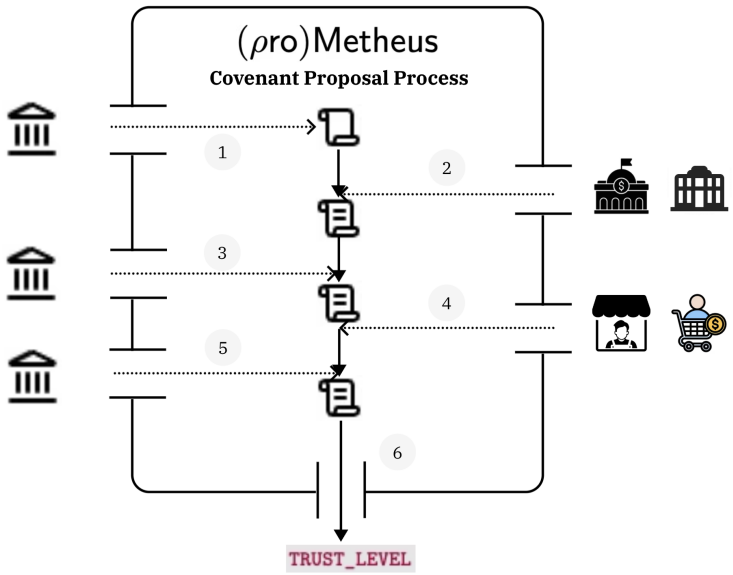


Figure 1 The (ρro)Metheus covenant proposal process is an iterative procedure where the regulator proposes criteria of audit in step (1). In step (2), other stakeholders such as financial institutions amend the proposal. This is sent to the regulator for review in step (3). Then more stakeholders are invited to the roundtable for review in step (4). In step (5), regulators give final approval. And in step (6), this law is translated into code as part of the (ρro)Metheus Typed ρ-Calculus Covenant-Judgement system. This community-defined type or Covenant-Judgement can now be used to audit AI models and datasets for stakeholder compliance.

(ρro)Metheus replace the notion of "code is law" with "`type is law`." Here the subjects under legal audit are computer programs, not people.

The most important aspect of this Covenant-Judgement construction process is that there is a one-to-one correspondence between drafting and enforcing regulations, and designing types and type-checking code. This is (ρro)Metheus' interpretation of the adage "code is law," or alternatively the *Covenant-Type Correspondence*. This correspondence is outlined in table 1.

Law	**Type**
Assign Legal Status	Type Judgment or Covenant-Judgement
Drafting Legislation	Type Design
Core Constitution	Type Primitives
Legislation Amendment	Type Extension
Legislation Enforcement	Type Checking by Compiler
Upholding Covenant	Does Type check
Violating Covenant	Does not Type Check
Relationship before the Covenant	Data/Model Composition before the Compiler
Proving Action is Legal under Terms of Covenant	Proving Typed ρ-Calculus Maintain Logical Invariants

Table 1 The correspondences between legal covenants and type system in Typed ρ-Calculus is the theoretical underpinnings that give a *computational interpretation* to the notion that *compiling code is the same as applying the law*. Under this correspondence, when the (ρro)Metheus compiler type checks the Typed ρ-Calculus code, it is also auditing the AI agents and datasets for legal compliance.

1.2 Broad Stakeholder Perspective

The technical perspective may be sufficient for those with a background in both machine learning and programming language theory, however it is too abstract for the non-technical audience. This section pulls the lens back and recontextualize (ρro)Metheus within the technical and political milieu of the present moment. We answer the question: why is (ρro)Metheus necessary when a diverse set of tools exists to characterize the confidence, validity, and scope of statistical models and datasets?

Why (ρro)Metheus

The demand arises with the latest tranche of breakthroughs in AI around large language models (LLMs). Nominally LLMs are trained via regression on time series data, therefore they are not categorically different from i.e., deep convolution neural networks (CNNs) trained on non-time-series data, or a soft-object manipulations algorithm sampled via reinforcement learning. However the *user-behavior* that have spawned around LLMs applications differ significantly from earlier machine learning tools. Presently users are not only using LLMs as general purpose information retrieval systems, but they are also prompting them as "intelligent reasoning" engines. This user behavior spans across many settings:

- *Financial setting*: banks and insurance companies are using LLMs in high-valued contexts such as producing analyst reports to inform buy/sell decisions. They have never used CNNs in this way. Moreover, the prior generation of financial models used to price assets are either simple, i.e. linear regression on a few factors, or reasonably explainable models based on heat-diffusion equations. In contrast, LLMs are complex and opaque.
- *Medical setting*: doctors are using LLMs to summarize patients' past to inform an opinion, while patients are using LLMs to diagnose their illness. This has far reaching consequences extending past the hospital setting, including the pricing of insurance and other ancillary services.

- *Consumer setting*: people are treating LLMs as friends, teachers, and personal psychologists. This use case may appear the most trivial, however social media has already demonstrated its capacity to silo citizens into misinformation echo chambers and throw elections. Moreover, whereas the harm due to social media is bounded by the number of people who posts, LLM-based social media tools have no such upper bound. This presents a long-term challenge to the health of any society.

In summary, unlike prior deep learning models, LLMs function as "artificial people," or more to the point: "entry level employees." Now given the variety and sensitivity of LLMs' application domains, we assert LLM-based applications require stricter oversight, perhaps even legal oversight, from regulators beyond rudimentary internet guidelines.

1.2.1 The (ρro)Metheus Oracle Desiderata: a High Level Attempt

The (ρro)Metheus computational environment is a protected sandbox with strictly defined and enforced rules. In particular, the *type-system* of the (ρro)Metheus Typed ρ-Calculus is crafted to satisfy the following desiderata:

1. *Flexibility*: regulators from across jurisdictions may define laws governing the behavior of LLMs in a way that suit their local conditions. This is important since language models summarize the internet and with it: culture. There cannot be one set of universal "LLM-laws," as different cultures carry different standards of "good behavior." Therefore the defined boundary conditions must be different.
2. *Composability*: it would be prohibitively expensive, not to mention tedious, to define laws governing every invocation of an AI agent. Thus Typed ρ-Calculus language is defined so that regulators can specify simple rules, and when composed they form a rich system spanning the behavior of an AI agent.
3. *End-to-end oversight*: once the boundary conditions are defined, regulator-privilege extends deep into any AI-based system over the course of its application lifecycle. This means regulators may define high-level laws governing what is good AI behavior. Then the laws would propagate from data collation, to model-training, refinement, and finally inference as well as post-inference data-collection steps. If one conceptualizes an AI agent as a living being, then (ρro)Metheus regulates agent-behavior from cradle to grave.

In conclusion, the (ρro)Metheus Typed ρ-Calculus computational framework operationalizes the adage "code is law." In this case the subjects under legal audit are not people but computer programs. In more technical terms:

> The (ρro)Metheus Typed ρ-Calculus language is a new way of realizing *code as law*. It does so with a novel application of type systems common in statically-typed programming languages. Whereby AI agents and data constitutes the term universe, while the regulator legal framework defines the type universe. This type system, or legal framework, governs:
>
> - how AI learns;
> - once learned, how AI interacts with people, data, and each other.

1.2.2 The (ρro)Metheus Stakeholder Stack

The fact that (ρro)Metheus interacts with regulators at the sovereign level is a critical differentiator and unique go-to-market channel. Presently both financial and medical institutions are inundated with "AI offerings" from major firms and startups alike.

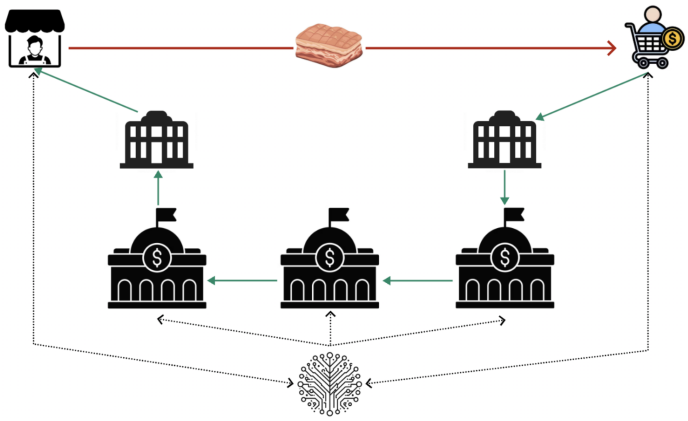


Figure 2 The (ρro)Metheus oracle applied to supply chain orchestration, whereby vendors sell foodstuff to buyers across national borders. The (ρro)Metheus oracle informs trade across stakeholders along this transnational network servicing: buyers, sellers, commercial/central banks, and insurers.

- The startups sell GPT wrappers: that is new way of selling the same old enterprise workflows. This game is stale and soon they shall be cannibalized.
- Major firms such as Microsoft and Google use "AI-agents" as cheese in the trap to sell what they have always sold: API calls and data egress. Ambitious managers use these projects to pad their promo-package. And once promoted, the projects are typically orphaned and whither away.

We escape this red ocean completely and appraise the stakeholder stack as follows:

- *Sovereign regulators*: (ρro)Metheus sell protected computational environments, whereby all code execution is compliant with respect to local laws and norms. The importance of security cannot be overstressed, especially as we enter a multipolar era whereby all things on the chessboard enter a state of motion. Since laws are public goods that rightly belong in the public domain, we engage with regulators through the nonprofit foundation FairCovenant. See the whitepaper at http://bit.ly/44EiQV0 for details on how we engage with regulators [LP25].
- *Enterprises*: (ρro)Metheus sells compliance as a service, so that the downsides of said institutions using AI are protected. This is critical: if AI is to make high valued decisions, then someone must be liable - (ρro)Metheus answers this question in a principled manner.
- *Enterprises Analysts*: (ρro)Metheus offers a high level interactive environment whereby analysts may query the data for causal relations. This dovetails how people are using LLMs anyways. However the statistical language describing this class of auto regressive models *cannot express causation*. The (ρro)Metheus Typed ρ-Calculus language fills this gap by expressing causality via a principled mathematical language.

The last point bears repetition. Recall Typed ρ-Calculus is not just for regulator-compliance, but also for *probabilistic causal reasoning*. We did not stress this aspect of (ρro)Metheus in the introduction, but it shall be specified in the main body shortly.

1.3 The Greater Metheus Ecosystem

Although (ρro)Metheus is a general purpose computational oracle, it is initially deployed for stakeholders along specific supply chains to support transnational trade organizations (see figure 2). Additionally, (ρro)Metheus will activate the larger open-source AI community by drawing both from

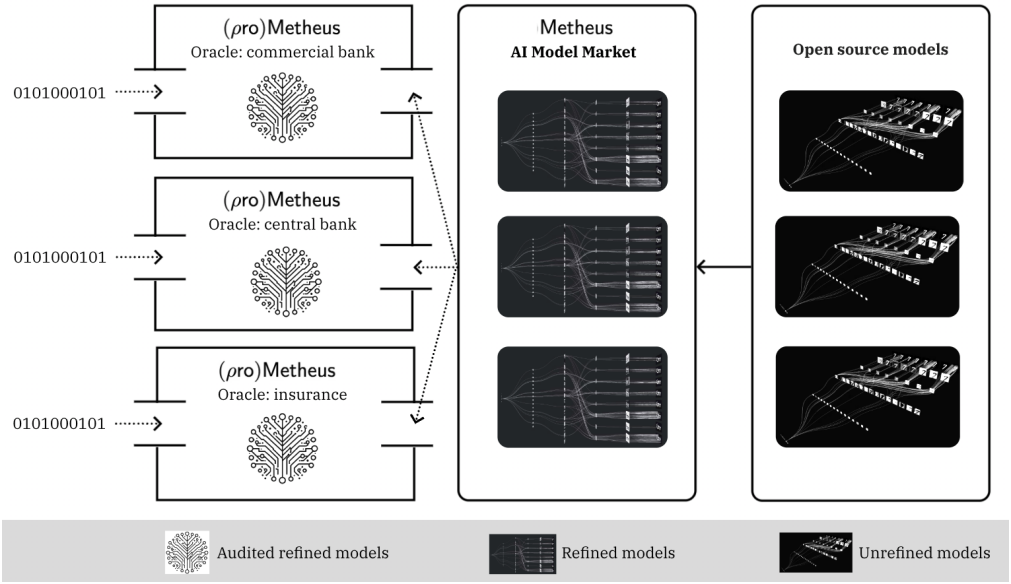


Figure 3 The (ρro)Metheus Oracle connects enterprises with the broader AI foundational model ecosystem by embedding refined AI models within a rigorous legal framework.

freely available models, as well as incentivizing the community to fine-tune or train new models. The process is outlined in figure 3:

- *left column*: there will be separate (ρro)Metheus oracle environment for each industry stakeholder. This is appropriate as each player has different standards of data validity, and AI model security. These separate requirements find congruence in the Typed ρ-Calculus type system.
- *center column*: (ρro)Metheus is connected to an open source AI model market, whereby foundational models are refined by distributed teams of ML engineers. [4]
- *right column*: this moment in technology is unique because of the gluttony of open source foundational models. We funnel this set of commodified assets into the privileged environment of (ρro)Metheus oracle.

Observe that (ρro)Metheus's moat is the regulator and stakeholder-defined laws of data validity and model security. That is say: **their wall is our moat.**

In (ρro)Metheus, the regulator's wall is our moat.

---

[4] See monograph at http://bit.ly/4lFn8Sf for details on AI model market [Lin25].

2 The Covenant-Type Correspondence

This section underlines the principled basis of Typed ρ-Calculus, that is the correspondence between:

- a legal covenant,
- and types that mediate how program assets interface with each other.

This correspondence formalizes the notion that *compiling code is the same as applying the law*; or simply "code is law." Some definitions and examples are listed below, followed by more in depth discussion on the correspondence.

In (ρro)Metheus, code that compiles is also code that obeys the law. That is to say compliant computations run to completion, while illegal computations fail *before* execution. This is know as: `static guarantee of covenant compliance`.

2.1 Preliminary Definitions and Examples

▶ **Definition 2.** *[25a] In law, a **legal covenant** is a binding agreement or promise, often included in a contract or deed, that obligates one party to perform or refrain from a specific action. It can be 1) positive covenant requiring an action, or 2) a negative covenant restricting an action.*

▶ **Example 3.** [25b] In (Property Law), there are two types of restrictive covenants: affirmative and negative:

- An affirmative covenant obligates a person to act. For example, a covenant that requires the homeowner to keep the trees trimmed in the yard is an affirmative covenant.
- A negative covenant prohibits a person to act. For example, a negative covenant can forbid a homeowner from building a fence.

▶ **Definition 4.** *[Pie02] A **Type System** is a tractable syntactic method for proving the absence of certain program behaviors by classifying phrases according to the kinds of values they compute. Its theoretical basis lies in early 20th century study of logic, mathematics, and philosophy. In computational terms, programs equipped with types allows the software engineer and the compiler to reason about the run-time behavior of the program.* [5]

▶ **Example 5.** (`Int` and `String`) In *statically typed languages* such as `Haskell`, every data and function in the language is associated with a type. The type specifies the proper behavior the function, while the compiler ensures the program type-checks, or in more informal terms: "behaviorally correct" *before* the code runs. This removes entire class of errors from run time and ensure the code does not behave poorly as it executes. The following code block defines a function `strLen` counts the length of a `String`, in this case written as `[Char]`.

[5] A compiler is a software program that translates one programming language to another.

```
strLen :: [Char] -> Int
strLen [] = 0
strLen x:xs = 1 + strLen xs
```

Observe this function is typed so that it can only accept a string (in this case represented as list of Characters), so that we have: `strLen "hi" = 2`. If instead the user passes a Boolean value into the function with `StrLen true`, then the function `strLen` will fail outright since it is "contractually obligated" to only accept values of type `[Char]`.

2.2 The Correspondence

The type-covenant correspondence occurs in two phases: 1) the drafting of legislation or type system design, and 2) the enforcement of legislation or static type checking. [6] A summary of the similarities is found in table 2.

Law	Type
Assign Legal Status	Type Judgment
Drafting Legislation	Type Design
Core Constitution	Type Primitives
Legislation Amendment	Type Extension
Legislation Enforcement	Type Checking by Compiler
Upholding Covenant	Does Type check
Violating Covenant	Does not Type Check
Relationship before the Covenant	Data/Model Composition before the Compiler
Proving Action is Legal under Terms of Covenant	Proving Typed ρ-Calculus Maintain Logical Invariants

Table 2 The correspondences between legal covenants and type system in Typed ρ-Calculus is the theoretical underpinnings that give a *computational interpretation* to the notion that *compiling code is the same as applying the law*. Under this correspondence, when the Typed ρ-Calculus compiler type checks the Typed ρ-Calculus code, it is also auditing the code for legal compliance.

Now we expand table 2 as follows:

- *Legislation design*: the drafting of legislation corresponds to designing type system within Typed ρ-Calculus.
  1. In real life one finds a core constitution, with many extensions that flow from the core structure. Similarly within the (ρro)Metheus computational environment, one will find primitive type and semantics that define entities and their relations. This is followed by extensions defined by the non-profit foundation defined in [LP25].
  2. In particular, in (ρro)Metheus Typed ρ-Calculus, the correspondence between type and covenant arises in the negative direction. That is to say the Typed ρ-Calculus compiler prevent non-compliant AI agents from running.

[6] Normally we abhor the egregious anthropomorphization of machine learning in popular literature. But in this setting there is a fine parallelism of how human societies relate to the *legal machinery*, and how AI agents in (ρro)Metheus computational environment relate to the Typed ρ-Calculus type system. That is to say in both cases we focus on the rationalization or mechanization of relationships.

3. Finally just as covenants govern the behavior of people w.r.t each other before the state, in (ρro)Metheus the type system governs the *composition* of data with data, data with models, and model-to-model composition.

- *Legislation enforcement*: unlike covenants in the context of people, inside the (ρro)Metheus computational environment, AI agents and associated software cannot violate covenants at all. This is because as Typed ρ-Calculus mobilize computational assets such as data and AI agents, said assets are type-checked before they are used in regulated computations. The advantage of AI agents over people is that while people may violate covenants, digital covenants defined by Typed ρ-Calculus logic are absolute: no AI agents may break the type system provided they are bound to (ρro)Metheus. Therefore compliant computations run to completion, while illegal computations fail *before* execution. This property of the (ρro)Metheus environment is called *static guarantee of covenant fulfillment*.

3 Formal Specification of the (ρro)Metheus Covenant

3.1 Introduction

The (ρro)Metheus Typed ρ-Calculus language is an example of a domain-specific programming language or DSL. DSLs are common in industry, and span from simple scripting languages for ad-hoc tasks, to complex languages embedded into a far-reaching ecosystems. The primary example here is MatLab. A complete (ρro)Metheus computational environment is comparable to MatLab in complexity. However whereas MatLab is a tool with low level integration with the legal structure of academia, (ρro)Metheus is deeply embedded into the financial and legal aspects of transnational trade systems. This requires a broad surface area of engagements with legal, financial, and insurance entities servicing said trade routes.

Consequentially, the determination of core type or covenant primitives in the Typed ρ-Calculus DSL is part of a broader governance process of The FairCovenant Foundation outlined in [LP25]. In particular, FairCovenant collaborate closely with stakeholders to determine the proper regulatory requirements needed to deploy AI agents. These requirements are then encoded into the type system, or legal system of (ρro)Metheus computational environment. See figure 4 for how the process is denoted procedurally. [7].

Moreover, since the type system is shared by multiple stakeholders that span many jurisdictions, the type system itself exists as a *public good*, and therefore is financed as such. Since the exact design of the core type system is delegated to the FairCovenant Foundation, the rest of the paper will speak of types in the generic sense with symbol κ. Covenant type assignment is then written with:

$$\Gamma \models \texttt{agnt} : \kappa_1 \qquad \Gamma \models \texttt{data} : \kappa_2. \tag{1}$$

Expression 1 is read: under the *validation environment* Γ defined by The FairCovenant Foundation, the AI `agnt` satisfies covenant κ_1, and the `data` has covenant value κ_2. The symbol Γ is also referred to as "typing context" in programming language terms. See section 4.1 for more on type judgement.

3.2 Prior Art

Now we reify the type assignment of expression (1) with concrete illustrations. Over the past few years, there has been a growing call to better document machine learning models and the data they are

[7] Link to whitepaper: http://bit.ly/44EiQV0

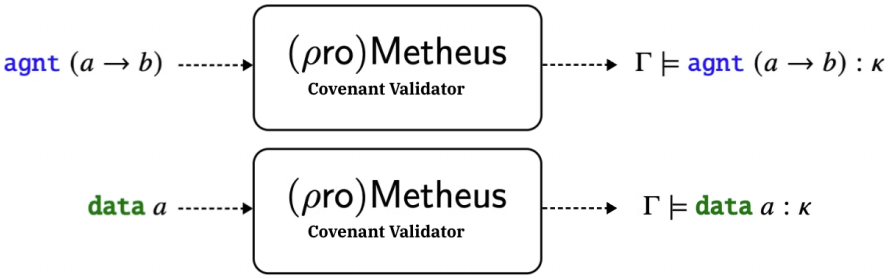


Figure 4 The (ρro)Metheus covenant-invalidator Γ ingests `data` or `agnt`, and assigns some Covenant-Judgement κ to each computational asset. The exact rules governing the content of the covenant κ is determined by The FairCovenant Foundation, so that once judged the `agnt`/`data` is deemed regulator-compliant. This κ now governs how said computational asset can interact with other assets in the (ρro)Metheus protected environment. Here compliant computations run to completion, while illegal computations fail *before* code execution. Thus illegal code cannot even run. This property is called: *static guarantee of covenant fulfillment*.

trained on. The next two examples represent preliminary attempts of assigning types to AI assets. The (ρro)Metheus Typed ρ-Calculus ecosystem will take the flavor of documentation presented in examples 6 and 7, and express them as types.

▶ **Example 6.** [**DBLP:journals/fat/Mitchell19**] (**Model Cards**) are documentations (or meta-data) that determines the application scope and provenance of machine learning models. This is an example of $\Gamma \models$ `agnt` $: \kappa$. From the abstract the authors state:

"Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions...that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information... this framework can be used to document any trained machine learning model."

An example of documentation for a particular model is presented in figure 5.

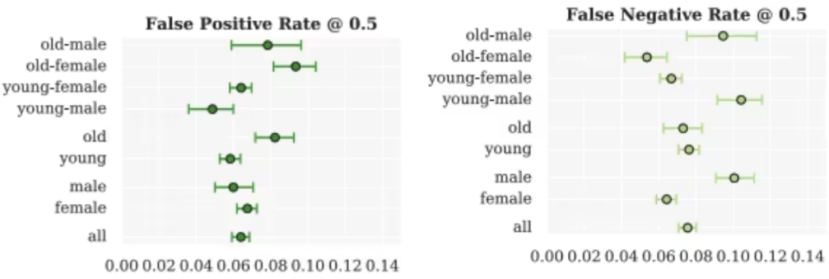


Figure 5 In this model for smiling detection in images reported in [**DBLP:journals/fat/Mitchell19**]. Here the false positive and false negative rates for each model is reported, broken down by demographic data. Documenting the performance envelope of a model is important if the model is to be deployed in sensitive areas, such as detecting the likelihood of a criminal from images. This likelihood could be used to make arrests, which presents a challenge as said estimations are often skewed by the exiting on criminals. In (ρro)Metheus Typed ρ-Calculus environment, the totality of a given model's metadata would be abstracted into a unique type **T**.

▶ **Example 7.** [Mah22] (**Data Cards**). Similar to model cards, data cards document the provenance, validity, and applicability of data used to train machine learning models. This is an example of $\Gamma \models$ `data` $: \kappa$. From the abstract we have:

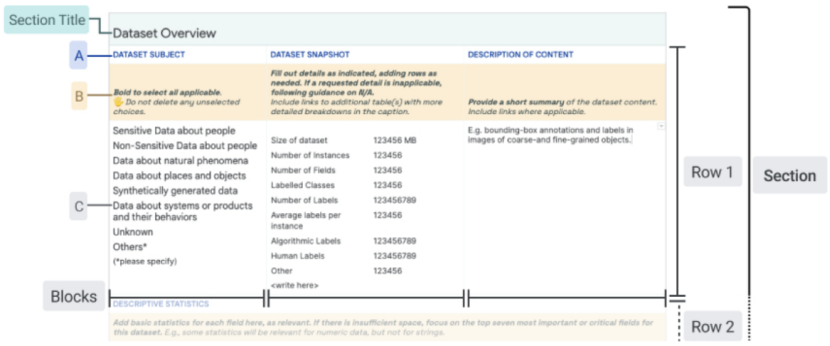


Figure 6 Similar to model cards, data cards document meta-information on the dataset used to train AI models. Since machine "learning" is data compression, this meta information is then appropriately inherited by any AI model trained on the data, along with any additional meta-information accumulated while the model is under training. Image from [Mah22].

"we propose Data Cards for fostering transparent, purposeful and human-centered documentation of datasets within the practical contexts of industry and research. Data Cards are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset's lifecycle for responsible AI development. These summaries provide explanations of processes and rationales that shape the data and consequently the models—such as upstream sources, data collection and annotation methods; training and evaluation methods, intended use; or decisions affecting model performance." An example data card format is show in figure 6.

The cited paper [**DBLP:journals/fat/Mitchell19**] on model cards is not just academically forward, the careers of the authors are also cautionary tales for those who push for regulation on AI agents without proper institutional support. In particular, the authors Margaret Mitchell and Timnit Gebru are of an "activist mindset." They would later author a different paper challenging the ability of deep learning models to think. No serious scholar in the field confuses deep learning model inference for "thinking." However the conduct of the authors nonetheless drew the ire of management at Google. They found themselves in the crosshair of a Distinguished Google Fellow, the two women were subsequently fired abruptly.

Example 6 is a case study of how *not* to push for closer scrutiny on AI models, especially when so many livelihoods, reputations, money, and egos are on the line. This is why (ρro)Metheus deploy AI regulation and legal framework through the nonprofit foundation FairCovenant. Only by working closely with regulators, insurers, and central banks in their domain of operation, can we externalize the necessary *legitimacy* to regulate artificial intelligence.

3.3 Formal Specification of the (ρro)Metheus Covenant

This section formally specifies the (ρro)Metheus Covenant as a privileged datatype in the Typed ρ-Calculus universe. Recall the details of the data validation criteria are designed and maintained by The FairCovenant Foundation. The format of said validation criteria is akin to a checklist. In the event whereby a criterion is real valued, then it can be discretized into regions. This region is then expressed as a binary tree, so that a real value falling within one of the zones correspond to a particular path down the tree. [8]

▶ **Definition 8.** *(An* `agnt` *or* `data` ***validity criteria*** *C) is a check-list of quality measures that the* `agnt` *or* `data` *must satisfy to partake in the* (ρro)Metheus *computational environment.*

The data and model cards of examples 7 and 6 are instances of validity criteria C.

▶ **Definition 9.** *(A* Covenant-Judgement κ*). For every set of* `agnt` *or* `data` *validity criteria C of length* $|C| = n$ *defined by The* FairCovenant *Foundation, so that each criteria in C is satisfied independently of another. Now let the* (ρro)Metheus *Oracle judges some* `agnt` *or* `data` *according to this criteria C, so that it generates a verdict* $\kappa_{2^n,i}$ *of length n, whereby:*

- $\kappa_{n,i}[j] = 1$ *if the* `agnt` *or* `data` *satisfies criterion j.*
- $\kappa_{n,i}[j] = 0$ *otherwise.*

This n-bit vector $\kappa_{n,i}$ *is a* Covenant-Judgement *w.r.t the criteria C. Moreover, the following three statements are equivalent:*

- `agnt` *or* `data` *has* Covenant-Judgement $\kappa_{n,i}$*;*
- `agnt` *or* `data` *witnesses* $\kappa_{n,i}$*.*
- *In formal notation:*

$$\Gamma \models \texttt{agnt} : \kappa_{n,i} \quad \text{or} \quad \Gamma \models \texttt{data} : \kappa_{n,i}.$$

We denote the set of all possible covenants generated by criteria C with the symbol $\mathbb{K}$. Observe that given a criteria set of length n, $\mathbb{K}$ is exactly the set of all paths down a binary tree of height n:

$$\mathbb{K}(C) = \{\kappa_{n,0}, \ldots, \kappa_{n,2^n-1}\}.$$

And each Covenant-Judgement $\kappa_{n,\mathrm{i}}$ is one path down this tree (see figure 7).

▶ **Example 10.** (**A covenant set** $\mathbb{K}(C)$) over three criteria is the set of paths down a binary tree of depth 3, and may be drawn:

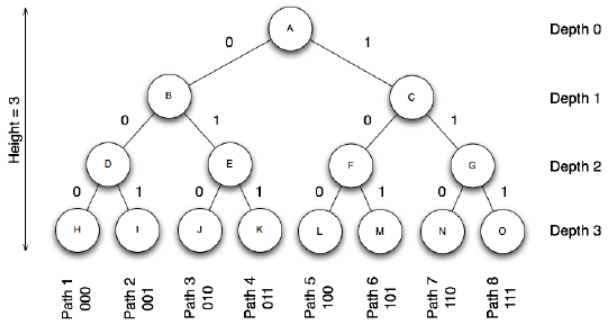


Figure 7 A binary tree of depth 3 induces $2^3 = 8$ possible 3-bit strings. Each string is an example of a covenant indexed by 3, written: $\kappa_{3,\mathrm{i}}$.

[8] This formulation is elegant because the covenant itself is also structured as a binary tree, see definition 9.

In this example, we may enumerate the possible Covenant-Judgements with:

$$\mathbb{K}(C) = \{\kappa_{3,0}, \kappa_{3,1}, \ldots, \kappa_{3,7}\} \text{ where}$$
$$\kappa_{3,0} = 000 \quad \kappa_{3,1} = 100 \quad \kappa_{3,2} = 010 \quad \kappa_{3,3} = 001 \quad \kappa_{3,4} = 110 \quad \kappa_{3,5} = 101 \quad \kappa_{3,6} = 011 \quad \kappa_{3,7} = 111,$$

so that each $\kappa_{3,x}$ is a particular Covenant-Judgement. For example if a particular `data` a satisfies the first and third criteria only, then `data` a *witnesses* 101. Its Covenant-Judgement is written:

$$\Gamma \models \texttt{data}\ a : \kappa_{3,5}, \quad \text{where } \kappa_{3,5} = 101.$$

▶ Remark 11. (**Simplified notation on** Covenant**-**Judgement). The notation $\kappa_{3,1}$ is cumbersome, so we prefer to write κ_1 instead, and leave the length of the bit-string 3 implicit.

Additionally, Typed ρ-Calculus must judge covenants by comparing them numerically, so that (ρro)Metheus may price covenant witnesses accordingly. Thus we will define a notion of distance and "less than." Hence we place an algebra on $\mathbb{K}$ as follows.

▶ **Definition 12.** *(**A covenant space** $\mathbf{K}$) with respect to some* FairCovenant *Foundation defined criteria C is a metric space equipped with a distance function. Written:*

$$\mathbf{K}(C) = (\mathbb{K}(C), \ ||\ ), \quad \text{where } ||\ : \mathbb{K} \times \mathbb{K} \to \mathbb{R}. \tag{2}$$

Where the distance function $||$ is defined by a suitable arithmetization of the Covenant-Judgement*s in $\mathbb{K}$. This arithmetization is defined by The* FairCovenant *Foundation, so that given arithmetization function* arith*, we have:*

$$\kappa_i \,||\, \kappa_j = \mathsf{arith}(\kappa_i) - \mathsf{arith}(\kappa_j). \tag{3}$$

From this point forward, when we say "covenant," we refer to the covenant space $\mathbf{K}(C)$, so that for every FairCovenant criteria of length n, there is an associated covenant space $\mathbf{K}(C)$. Now we consider the issue of Covenant-Judgement composition. This arises in settings when:

- Datasets are composed with other datasets, and they have different covenant criteria. For example, `data` a may be a set of images, while `data` b may be a set of image-aligned captions. They will satisfy different covenant criteria, however the combined dataset [9] will carry Covenant-Judgement information from the source datasets.
- Trained `agnt` that witnesses some Covenant-Judgement κ is refined on `data` -set of Covenant-Judgement κ'. Then the resultant model must have some type κ'', the content of this κ'' must be defined.
- The output of one `agnt` $(a \to b)$ is input some other `agnt` $(b \to c)$. If `agnt` $(a \to b)$ witness κ, while `agnt` $(b \to c)$ witness κ'. Then what is the resultant value c output by the model?

▶ **Definition 13.** *(**Covenant Composition**) between two covenant spaces $\mathbf{K}$ and $\mathbf{K}'$ generated by two different criteria C and C', is the Cartesian product of the two spaces. Written:*

$$\mathbf{K} \ \otimes \ \mathbf{K}' = \Big(\big\{ \boldsymbol{\kappa}_m = (\kappa_i, \kappa_j) \ : \ \kappa_i \in \mathbf{K}, \kappa_j \in \mathbf{K}' \big\}, \ \ || \Big). \tag{4}$$

Now we can define the distance $||$ between two product covenants by arithmetizing each of the elements in the vector with:

$$\boldsymbol{\kappa}_m = (\mathsf{arith}(\kappa_i), \mathsf{arith}(\kappa_j)). \tag{5}$$

[9] Sometimes called a multi-modal dataset.

Thereby interpreting this κ_m as a real-valued vector. Then the distance function is defined as the dot product between the two vectors:

$$\kappa_m \,||\, \kappa_n = \kappa_m \cdot \kappa_n. \tag{6}$$

Some further observations follow:

- Covenants are closed under composition. That is to say the product of covenants are a covenant themselves.
- Covenant composition is not necessarily commutative. That is to say we should not expect this to be true: $(\kappa_\mathrm{i}, \kappa_\mathrm{j}) = (\kappa_\mathrm{j}, \kappa_\mathrm{i})$.
- We may introduce an identity covenant for every Foundation-defined covenant criteria, by defining this identity as the all-ones vector of suitable length. In vector space language, this corresponds to lifting some sub-space into the affine-space.

One final definition will round out the chapter.

▶ **Definition 14.** *(**The** (ρro)Metheus **Covenant Universe** $\mathfrak{K}$) is generated as follows:*

1. *Given the set of all possible criteria $C, \ldots$ The* FairCovenant *Foundation defines, we generate the set of all possible covenant spaces with:*

$$\mathbb{S} = \left\{ \mathbf{K}(C_i), \ldots \;\middle|\; C_i \in \{C_1, \ldots\} \right\}.$$

2. *Now construct the sigma-algebra over $\mathbb{S}$ with $\sigma(\mathbb{S})$.*
3. *And complete each subset $\mathbb{S}$ in $\sigma(\mathbb{S})$ under $\otimes$ with:*

$$\mathfrak{K} = \left\{ \otimes_{\mathbf{K}(C_i)\in\Sigma} \mathbf{K}(C_i) \;\middle|\; \Sigma \in \mathbb{S} \right\}. \tag{7}$$

In the remainder of the paper, we define what it means to assign every program in Typed ρ-Calculus some Covenant-Judgement in $\mathfrak{K}$. This concludes the discussion on the formal specification of Covenant-Judgements. In summary:

1. Covenants are specified by The FairCovenant Foundation, it is a list of criterion that `agnt` and `data` must satisfy to partake in the (ρro)Metheus computational enviornment.
2. A particular instance of a Covenant-Judgement is just a bit string, which when appropriately arithmetized, becomes a real number that can be subtracted.
3. When a computational asset such as `agnt` witnesses some Covenant-Judgement κ, we assign the `agnt` with this Covenant-Judgement with the notation: $\Gamma \models$ `agnt` $: \kappa$.
4. When computational assets compose, the Covenant-Judgements also compose accordingly. Composition occurs when:
   - `data` compose with other `data` to augment information content.
   - `agnt` compose with `data` during training or inference.
   - `agnt` compose with other `agnt` when the output of one `agnt` is used as input into another `agnt`.

   Composition allows us to build complex Covenant-Judgements, or *legal obligations* from atomic Covenant-Judgement definitions. This greatly augment (ρro)Metheus' ability to audit `data` and `agnt`s. The FairCovenant Foundation member's time is valuable, composition allows them to define primitive obligations that code must satisfy, so that the these covenant-primitives *span* a set of acceptable AI behavior in the (ρro)Metheus universe.

4 Formal Specification of Simply Typed ρ-Calculus

The rest of the paper recursively descends downwards, so as to define Typed ρ-Calculus for covenant-bound probabilistic causal intelligence. The language is introduced by layers as follows:

1. We begin at the highest level of abstraction, and define how models and data interact with each other in the aggregate. In particular, we are sensitive to how stakeholders in the FairCovenant Foundation defines how AI assets interact with each other.
2. Next we descend down one level, and express how Typed ρ-Calculus can be used to build a probable causal graph given information drawn from raw data and information summarized by LLM models. We then show how the type system interacts with the causal graph to properly assign legal culpability as Typed ρ-Calculus computes probable cause of an event.
3. Finally, we propose future directions whereby the type system may find coverage, for example in the computational graph of deep learning models in i.e. TensorFlow.

In this chapter, we first introduces the basic technical knowledge needed to understand Typed ρ-Calculus and Loare-Logic, and then specify layer one in point (1) above.

4.1 Technical Background

This section provides the necessary background from the field of programming languages and formal verification to understand Typed ρ-Calculus used to compute causal relations from data, and the (ρro)Metheus Loare-Logic used to generate proofs of Covenant-Compliance.

Language Syntax and Typing Relations

Given a typed programming language, its *typing relations* is an inference rule that describes how the type system assigns a type to a syntactic construction.

- These rules are applied by the type system to determine if a program is well-typed and what type the expressions have.
- If the program is well-typed, then it should satisfy certain properties such as all programs can either step forward or terminate with a value.
- Types restrict the universe of allowed programs so that all well-typed programs is "well behaved" according to the designer's desiderata [Pie02].

An example follows to demonstrate the interplay between the syntax of a language and its typing relations.

▶ **Example 15.** [Pie17](**Simply-Typed Lambda Calculus**). Given a language with syntax:

expressions	$\mathsf{e} := \mathsf{x} \mid \lambda \mathsf{x} : \tau.\mathsf{e} \mid \mathsf{e}_1\mathsf{e}_2 \mid \mathsf{n} \mid \mathsf{e}_1 + \mathsf{e}_2 \mid ()$
values	$\mathsf{n} := \lambda \mathsf{x} : \tau.\mathsf{e} \mid \mathsf{n} \mid ()$
types	$\tau := \texttt{Int} \mid \texttt{Unit} \mid \tau_1 \to \tau_2.$

We introduce a relation or judgment over typing contexts Γ, expressions e, and types τ. The judgment:

$$\Gamma \models \mathsf{e} : \tau,$$

is read as "expression e has type τ in context Γ." Next we write the type judgement over all expressions:

$$\frac{}{\Gamma \models n : \texttt{Int}} \qquad \frac{}{\Gamma \models () : \texttt{Unit}} \qquad \frac{\Gamma(x) = \tau}{\Gamma \models x : \tau}$$

$$\frac{\Gamma \models e_1 : \texttt{Int}, \Gamma \models e_2 : \texttt{Int}}{\Gamma \models e_1 + e_2 : \texttt{Int}} \qquad \frac{\Gamma, x : \tau \models e : \tau'}{\Gamma \models \lambda x : \tau . e : \tau \to \tau'} \qquad \frac{\Gamma \models e_1 : \tau \to \tau' \; \Gamma \models e_2 : \tau}{\Gamma \models e_1 e_2 : \tau'}$$

Observe this set of typing rules are applied to each expression and determines how the Lambda Calculus expressions can be typed. Furthermore, line two shows the inductive typing rules whereby complex expressions featuring function application and addition are also typed.

The (ρro)Metheus Loare-Logic replicates this design pattern, by applying FairCovenant-defined Covenant-Judgement to the set of Typed ρ-Calculus expressions defined in eqn (8).

Hoare Logic

Hoare logic is a formal system with a set of logical rules for reasoning rigorously about the correctness of computer programs. Hoare-style verification is based on the idea of a specification as a contract between the implementation of a function its clients:

- The specification consists of the precondition and a postcondition.
- The precondition is a predicate describing the condition of the code or functions relies on for correct operation. The client must fulfill this condition.
- The post condition is a predicate describing the condition of the function establishes after correct running; the client can rely on this condition being true after the call to the function.

The goal of Hoare-style verification is thus to statically prove that, given a pre-condition, a particular post-condition will hold after a block of code executes. We do so by generating a logical formula known as a *verification condition*, constructed so that if true, we know that the program behaves as expected. From a syntactic standpoint, Hoare logic is presented to the programmer as *Hoare triples*.

▶ **Definition 16.** *[Pie17] (**A Hoare Triple**) is a claim about the state before and after executing a command. The standard notation is:*

$$\{\{P\}\}\ \varphi\ \{\{Q\}\}.$$

Meaning:

- *If command φ begins execution in a state satisfying assertion P,*
- *and if φ eventually terminates in some final state,*
- *then that final state will satisfy the assertion Q.*

Assertion P is called the precondition of the triple, and Q is the postcondition. [10]

▶ **Example 17.** [Pie17] (**Hoare Triple**) The following:

$$\{\{\mathbf{x} = 0\}\}\ \mathbf{x} := \mathbf{x} + 1\ \{\{\mathbf{x} = 1\}\},$$

is a valid Hoare triple, stating that command $\mathbf{x} := \mathbf{x} + 1$ will transform a state in which $\mathbf{x} = 1$ to a state in which $\mathbf{x} = 1$.

[10] We use double brackets for pre and post conditions, following [Pie17].

▶ **Example 18.** (**Hoare Logic**) applied to a programming language would proceed as follows. Given a simple language:

arithmetic expressions	$\mathsf{e} := \mathsf{x} \mid \mathsf{n} \mid \mathsf{e}_1 + \mathsf{e}_2 \mid \mathsf{e}_1 \times \mathsf{e}_2$
boolean expressions	$\mathsf{b} := \texttt{true} \mid \texttt{false} \mid \mathsf{e}_1 < \mathsf{e}_2$
commands	$\mathsf{c} := \texttt{skip} \mid \mathsf{x} := \mathsf{e} \mid \mathsf{c}_1; \mathsf{c}_2.$

Its set of Hoare triples are:

$$\frac{}{\{\{P\}\}\ \texttt{skip}\ \{\{Q\}\}} \qquad \frac{}{\{\{P[\mathsf{e}/\mathsf{x}]\}\}\ \mathsf{x} := \mathsf{e}\ \{\{P\}\}}$$

$$\frac{\{\{P\}\}\ \mathsf{e}_1\ \{\{R\}\},\ \{\{R\}\}\ \mathsf{e}_2\ \{\{Q\}\}}{\{\{P\}\}\ \mathsf{e}_1; \mathsf{e}_2\ \{\{Q\}\}}$$

Now given any program written in the language above, we can use these Hoare rules to verify the program satisfy certain conditions laid out in Q and P.

In the context of (ρro)Metheus:

- the precondition P defines the quality of data and statistical models that are input into the analysis;
- the command φ is the Typed ρ-Calculus computation that determines the output;
- while the post condition Q is quality of output that the analyst-written (ρro)Metheus program is contractually obligated to output.

The terms of contracts are determined by FairCovenant in accordance to all stake-holder demands. The core feature of Loare-Logic is to guarantee that Typed ρ-Calculus programs that cannot satisfy contractually-obligated standards of quality fail at compile time, *before* the program is run. This *static guarantee* of quality satisfies the dual mandate of:

- ensuring the safety and validity of model outputs;
- and saving money as full analytic runs and/or training runs are expensive.

4.2 Typed ρ-Calculus **Expressions**

At the top level, Typed ρ-Calculus formally specify the notion of atomic computational assets, and computation over said assets in the (ρro)Metheus universe. In particular, we define the following three type of asset composition:

- *(`data` -`data` composition)*: arises when data scientists augment datasets by combining it with other sources of data. Here the pressure point is to ensure the dataset does not degrade in quality under augmentation.
- *(`data` -`agnt` composition)*: arises when an `agnt`s consume data either in training, or in inference.
- *(`agnt`-`agnt` composition)*: arises when one `agnt` calls on another `agnt` to complete a sub-task. Now since the output of `agnt` could also be interpreted as `data` , this composition may be decomposed as a sequence of `data` -`agnt` compositions.

▶ Remark 19. We present the salient aspect of Typed ρ-Calculus while keeping the syntactical rigor to a minimum. This elevates the core features of the (ρro)Metheus computational environment, without inundating the readers with implementation details of a fully featured domain-specific programming language.

4.2.1 Elementary Typed ρ-Calculus Expressions

$$\begin{aligned} &\textit{atomic assets:} && \texttt{asset}\ a\ b := \texttt{data}\ a \mid \texttt{agnt}\ (a \to b) \\ &\textit{elementary functions:} && \texttt{fns} := \texttt{train} \mid \texttt{infer} \\ &\textit{covenant judgments:} && \kappa \in \mathfrak{K} \end{aligned} \tag{8}$$

In the rest of this paper, we will use φ to refer to *atomic assets* and *elementary functions* in Typed ρ-Calculus, as well as inductively defined expressions built from said code. Some more commentary follows:

- (*line 1*): in Typed ρ-Calculus, the atomic `asset` are either datasets `data` or AI agents `agnt`s. The expressions are parameterized by generic types a and b, which could be i.e. `String` or `Int` types. Note these parametrized types are not the same as the Covenant-Judgement, which specifies quality of data or model that has been audited by The (ρro)Metheus Oracle.
- (*line 2*): the elementary functions are `train` which takes an `agnt` $(a \to b)$ and (`data` a, `data` b) as input and outputs a trained `agnt` $(a \to b)$. Or `infer` which takes in `agnt` $(a \to b)$ as input along with query `data` a, and output inferred response `data` b.
- (*line 3*): the Covenant-Judgement of Typed ρ-Calculus is the entire covenant judgement universe as defined in def (14).

▶ **Example 20.** (**Document Data**) A set of raw conversation transcripts has type '`data String`'.

▶ **Example 21.** (**Dialogue Systems**) An automated dialogue system or chat-bot that accepts queries in string, and respond with raw text has type '`agnt` (`String` $\to$ `String`)'.

4.2.2 Group Composition over `asset` Expressions

Now given a set of words, we can "place a grammar" over this set, so that the words may compose into sentences. Similarly, if one conceptualizes `asset` as elementary programs or words, then one can "place an algebra" on `asset` so that they compose to form larger programs. The algebra of choice in Typed ρ-Calculus is the `Semigroup`.

▶ **Definition 22.** *(`Semigroup`) A semigroup is an ordered pair* $(\mathbb{S}, \times)$ *such that* $\mathbb{S}$ *is a non-empty set, and* $\times$ *is an associative binary operation on* $\mathbb{S}$.

▶ Remark 23. Addition and multiplication over integers are both examples of semigroups. In the case of integers, there is also an identity element associated with summation in 0. So that addition forms a *group* in (`Int`,+,0), and similarly for (`Int`,×,1) in the case of multiplication. In the case of `asset` , the identity asset is not necessarily defined, hence the restriction to semigroup.

The next definition specifies `asset` composition by placing the Semigroup algebra on `data` :

$$\begin{aligned} &\texttt{Class Semigroup}\ (\texttt{data}\ a)\ \ \text{where} \\ &\qquad \diamond :: \texttt{data}\ a \times \texttt{data}\ a \to \texttt{data}\ a \end{aligned} \tag{9}$$

▶ **Example 24.** (`data` **Composition**). Given two block of raw texts s_1 and s_2, each of type `data String`, we can concatenate the two blocks with $\diamond$ to form a new data set of type `data String`. Written:

$$s_1 \diamond s_2 :: \texttt{data String}.$$

Similarly, we place Semigroup operation on `agnt` as follows:

$$\texttt{Class Semigroup}\ (\texttt{agnt}\ (\cdot \to \cdot))\ \text{where}$$
$$\rightarrowtail :: \texttt{agnt}\ (a \to b) \times \texttt{agnt}\ (b \to c) \to \texttt{agnt}\ (a \to c) \tag{10}$$

▶ **Example 25.** (`agnt` **Composition**). Given an AI agent `agnt` (`String` → `String`) that consumes text and output text (i.e. a chat-bot), and another AI agent `agnt` (`String` → `Image`)′ that consumes text and generates images. Then we can compose them with:

$$\left(\texttt{agnt}\ (\texttt{String} \to \texttt{String}) \rightarrowtail \texttt{agnt}\ (\texttt{String} \to \texttt{Image})'\right) :: \texttt{agnt}\ (\texttt{String} \to \texttt{Image}).$$

And now we have a new AI agent that outputs images when prompted.

4.2.3 Elementary Functions over `asset` Expressions

Now we define two functions that operate on `asset` . These correspond to elementary operations in the (ρro)Metheus machine learning ecosystem.

$$\texttt{train} :: \texttt{data}\ a \times \texttt{data}\ b \times \texttt{agnt}\ (a \to b) \to \texttt{agnt}\ (a \to b) \tag{11}$$
$$\texttt{infer} :: \texttt{data}\ a \times \texttt{agnt}\ (a \to b) \to \texttt{data}\ b \tag{12}$$

Commentary below:

- Function (11) takes in some dataset (`data` a, `data` b), and machine learning model `agnt` $(a \to b)$, and `train`s the model to output a trained ML function `agnt` $(a \to b)$. This signature is written for training discriminative models with labeled input pairs. In the case of generative models, `data` b is simply passed into the `train` function as the empty set.
- Function (12) takes a model `agnt` $(a \to b)$, and accepts an input (or prompt in the case of LLMs) `data` a, `infer`s its value and outputs a prediction or response of type `data` b.

This ends the discussion on Typed ρ-Calculus top level expression. In summary:

- The atomic `asset` of the (ρro)Metheus computational environment are `data` and `agnt`. They must be audited according to FairCovenant-defined rules before use client-side.
- The elementary machine learning functions that operate over said `asset` are `train` and `infer`, they take `data` and `agnt` as inputs, and output the appropriate `asset` in response.
- The composition operators $\rightarrowtail$ and $\diamond$ compose `agnt` and `data` respectively. These semigroup functions, along with `train` and `infer`, allow the analyst to build complex Typed ρ-Calculus programs from atomic Typed ρ-Calculus `asset` .

In the next section, we introduce a novel concept termed Loare-Logic. It is the workhorse logical system used by the (ρro)Metheus compiler to validate Typed ρ-Calculus programs for Covenant-Compliance, before any AI-related computation is run.

4.3 Loare-Logic: **Axiomatic Semantics with Legalistic-Hoare Logic**

This section lays out the mathematical tools needed to enforce Covenant-Compliance of Typed ρ-Calculus expressions within the (ρro)Metheus computational environment. Section 4.3.1 defines Loare-Logic syntactically as a set of Covenant-Judgement rules, as well the mechanisms of how pre and post-conditions are enforced to ensure Covenant-Compliance. Then we enumerate all the rules in the subsequent section:

- Section 4.3.2 assigns Covenant-Judgement onto each atomic `asset` . This is equivalent to designating legal status to each entity in (ρro)Metheus, thereby reifying the type-covenant correspondence asserted in chapter 2.
- Sections 4.3.3 and 4.3.4 map `asset` composition onto Covenant-Judgement composition as defined by the product operation of definition (13). This ensures `asset` composed under $\diamond$ and $\rightarrowtail$ also witness the appropriate Covenant-Judgements, and are therefore Covenant-Compliant to the appropriate degree.
- Section 4.3.5 defines how Covenant-Judgement maps over the elementary function `infer` and `train`. That is to say: how does `infer`ence and `train`ing affect the output values' Covenant-Judgements.
- Section 4.3.6 provide some example proofs, demonstrating how Loare-Logic enforces Covenant-Compliance in complex Typed ρ-Calculus programs built inductively from `asset` primitives.

4.3.1 Loare-Logic **Introduction**

Loare-Logic is a formal system used by the (ρro)Metheus compiler to analyze the quality of data and model assets used in Typed ρ-Calculus programs. Loare-Logic ensures the analytic output of said programs are covenant compliant according to FairCovenant-defined rules.

▶ **Definition 26.** *(*Loare-Logic*) is a two-layered proof system that either generates a proof stating some* Typed ρ-Calculus *is covenant compliant, or terminates program execution for covenant violations. The system is layered as follows:*

- *A set of atomic and inductively defined rules assigning* Covenant-Judgement *to each expression in the* Typed ρ-Calculus *universe. Specifically, the following set of rules:*

$$rules := \{\texttt{data-L}, \texttt{agent-L}, \texttt{compa-L}, \texttt{compb-L}, \texttt{seql-L}, \texttt{seqr-L}, \texttt{seqd-L}, \texttt{train-L}, \texttt{infer-L}\},$$

assign each user-defined expression in Typed ρ-Calculus *onto some covenant κ in the covenant universe $\mathfrak{K}$. In the case where the expression φ is some atomic `asset` , then the rule is written:*

$$\textit{rule:} \quad \frac{}{\Gamma \models \varphi : \kappa}$$

If the expression is inductively defined using Semigroups operations $\diamond$ or $\rightarrowtail$, and/or using elementary functions such as `train` and `infer`, then the rule is written:

$$\textit{rule:} \quad \frac{\Gamma \models \varphi_1 : \kappa_1, \Gamma \models \varphi_2 : \kappa_2}{\Gamma \models \varphi_1\varphi_2 : \kappa_3}$$

Stating that if expression φ_1 witness Covenant-Judgement *κ_1 while φ_2 witness κ_2, then their combined expression $\varphi_1\varphi_2$ has* Covenant-Judgement *κ_3. The exact rules to induce the judgement κ_3 is defined by* (ρro)Metheus *compiler in conjunction with* FairCovenant *Foundation regulations.*

- *Building upon the rule primitives are a set of contractual obligations decorating* Typed ρ-Calculus *programs. They govern the execution of* Typed ρ-Calculus *programs, guaranteeing that the pre and postcondition of said programs are covenant compliant. Syntactically, this is written:*

$$\{\{\Gamma \models x_1 : \kappa_1, \ldots, x_n : \kappa_n\}\}\ \varphi(x_1, \ldots, x_n)\ \{\{\Gamma \models y : \kappa' \geq \kappa''\}\},$$
$$\text{where } y = \varphi(x_1, \ldots, x_n), \quad \text{with } \kappa_1 \in \mathbf{K}_1, \ldots \kappa_n \in \mathbf{K}_n, \quad \text{and } \kappa', \kappa'' \in \mathbf{K}'.$$

Stating that:

- *given some input assets* $x_1, \ldots, x_n$ *of* Covenant-Judgement $\kappa_1, \ldots, \kappa_n$*;*
- *and* Typed ρ-Calculus *expression* φ *that transforms the inputs onto output* $y = \varphi(x_1, \ldots, x_n)$*;*
- *the output* y *must witness some* Covenant-Judgement $\kappa' \geq \kappa''$.

The threshold value of data quality κ'' *is called* ***the boundary condition on the output of computation*** *as enforced by* Loare-Logic*. It ensures the resultant prediction or response from the* Typed ρ-Calculus *code* φ *meets the minimum standards of quality as defined by the* FairCovenant *Foundation. The pre and postconditions defined in the brackets* $\{\{\cdot\}\}$ *will automatically "decorate" programs written by analysts in some specific instance of the* (ρro)Metheus *computational environment.*

Observe how Loare-Logic combines the syntax of type relations in example (15), and that of Hoare logic in definition (16) to express contractual enforcement of Typed ρ-Calculus programs w.r.t pre-agreed Covenant-Judgement. Now we specify the content of each *rule* enumerated in the set above, and provide examples of the Loare-Logic triple in action.

4.3.2 Elementary Covenant-Judgement over Computational Assets

The judgment of individual `data` and `agnt asset` are done by The (ρro)Metheus Oracle according to FairCovenant Foundation defined rules. In this case we simply assume they are assigned some appropriately-defined Covenant-Judgement κ as follows:

$$\texttt{data-L}\ \frac{}{\Gamma \models \texttt{data}\ a : \kappa} \qquad \texttt{agent-L}\ \frac{}{\Gamma \models \texttt{agnt}\ (a \to b) : \kappa}$$

4.3.3 Semantics of `data` Composition w.r.t Covenant-Judgement

The following set of expressions takes the semigroup composition over `data` and `agnt` defined in expressions (9) and (10), and map them onto Covenant-Judgement composition as defined in expression (4).

$$\texttt{compa-L}\ \frac{\Gamma \models \texttt{data}\ a : \kappa_1, \Gamma \models \texttt{data}\ a : \kappa_2, \kappa_1 \preceq \kappa_2, \kappa_1, \kappa_2 \in \mathbf{K}}{\Gamma \models \texttt{data}\ a\ \diamond\ \texttt{data}\ a : \kappa_1}$$

$$\texttt{compb-L}\ \frac{\Gamma \models \texttt{data}\ a : \kappa, \Gamma \models \texttt{data}\ b : \kappa', \kappa \in \mathbf{K}, \kappa' \in \mathbf{K}'}{\Gamma \models \texttt{data}\ a\ \diamond\ \texttt{data}\ b : \kappa\ \otimes\ \kappa'}$$

- In `compa-L`, we are given some `data` a of Covenant-Judgement κ_1, and `data` a judged to be κ_2, so that the first judgement κ_1 is less "valuable" than the second. Then in this case, the entire dataset default to the less valuable judgment κ_1.

- In `compb-L`, there is some `data` a and `data` b, observe they are of different parameterized types. For example we may have `data String` and `data Image`, then their concatenation is the multi-modal dataset `data String` $\diamond$ `data Image`. In this case their covenants are drawn from different spaces, that is to say $\kappa \in \mathbf{K}$ and $\kappa' \in \mathbf{K}'$. Naturally the multi-modal dataset is in the product space $\kappa \otimes \kappa' \in \mathbf{K} \times \mathbf{K}'$.

▶ **Example 27.** (`compa-L` **and** `compb-L`) Suppose the following computational assets are loaded in the (ρro)Metheus environment:

$$\Gamma \models \texttt{data Int} : \kappa_1, \quad \Gamma \models \texttt{data Int}' : \kappa_2, \quad \Gamma \models \texttt{data Image} : \kappa,$$
$$\text{with } \kappa_1 < \kappa_2, \kappa_1, \kappa_2 \in \mathbf{K},\ \kappa \in \mathbf{K}'.$$

Then if the analyst build composite datasets, they will have the following inductively-defined Covenant-Judgement:

$$\Gamma \models \texttt{data Int} \diamond \texttt{data Int}' : \kappa_1, \qquad \Gamma \models \texttt{data Int} \diamond \texttt{data Image} : \kappa_1 \otimes \kappa.$$

Note that since `data Int` and `data Int`′ witness different elements of the same Covenant-Judgement space, the lower-valued Covenant-Judgement takes precedent. Where as `data Int` and `data Image` are drawn from different Covenant-Judgement spaces, thus the composite dataset rests in the product space of Covenant-Judgements or $\mathbf{K} \times \mathbf{K}'$.

▶ Remark 28. (**Notation on type and** Covenant**-Judgement**) In the code given in example 27, we use the notation $\Gamma \models$ `data` $a : \kappa$ to signify a computational `asset` of type `data` a has Covenant-Judgement κ. Technically, it is more complete to write, i.e.:

$$\Gamma \models \mathsf{x} :: \texttt{data}\ a : \kappa_1,$$

to signify that some dataset of value x of type `data` a has Covenant-Judgement κ_1. However this introduces another symbol x, which is cumbersome. Hence in all of our examples we leave the variable name x implicit.

4.3.4 Semantics of `agnt` Composition w.r.t Covenant-Judgement

Now we consider how sequencing over `agnt` by $\rightarrowtail$ affect Covenant-Judgement.

$$\texttt{seql-L}\ \frac{\Gamma \models \texttt{agnt}\ (a \to a) : \kappa_1, \Gamma \models \texttt{agnt}\ (a \to a)' : \kappa_2, \kappa_1 \leq \kappa_2, \kappa_1, \kappa_2 \in \mathbf{K}}{\Gamma \models \texttt{agnt}\ (a \to a)'' : \kappa_1}$$

$$\texttt{seqr-L}\ \frac{\Gamma \models \texttt{agnt}\ (a \to a) : \kappa_2, \Gamma \models \texttt{agnt}\ (a \to a)' : \kappa_1, \kappa_1 \leq \kappa_2, \kappa_1, \kappa_2 \in \mathbf{K}}{\Gamma \models \texttt{agnt}\ (a \to a)'' : \kappa_1}$$

$$\texttt{seqd-L}\ \frac{\Gamma \models \texttt{agnt}\ (a \to b) : \kappa_1, \Gamma \models \texttt{agnt}\ (b \to c) : \kappa_2, \kappa_1 \in \mathbf{K}, \kappa_2 \in \mathbf{K}'}{\Gamma \models \texttt{agnt}\ (a \to c) :\ \exists \kappa_3 \in \mathbf{K}'' \text{ s.t. s.t. } \kappa_3 = \texttt{proj}(\kappa_1 \otimes \kappa_2)}$$

$$\text{where } \texttt{proj}(\kappa_1 \otimes \kappa_2) := \mathsf{arith}(\mathbf{K}'') \cdot \sqrt{\left(\frac{\mathsf{arith}(\kappa_1)}{\mathsf{arith}(\mathbf{K})}\right)^2 + \left(\frac{\mathsf{arith}(\kappa_2)}{\mathsf{arith}(\mathbf{K}')}\right)^2}.$$

- In `seql-L` and `seqr-L`, one of the `agnt` witnesses a less valuable Covenant-Judgement than the other. In both cases the combined `agnt` $(a \to a)''$ defaults to the less valuable judgement κ_2.

- In `seqd-L`, the `agnt` are parametrized by different types: `agnt` $(a \to b)$ v.s. `agnt` $(b \to c)$. Consequentially, their Covenant-Judgement are drawn from different covenant spaces, and the composed `agnt` $(a \to c)$ is drawn from a third covenant space $\mathbf{K}''$. Thus, we have to search for the most suitable Covenant-Judgement $\kappa_3 \in \mathbf{K}''$ that we *expect* `agnt` $(a \to c)$ to witness. We use the following procedure:
  1. We interpret the Covenant-Judgement `agnt` $(a \to b)$ and `agnt` $(b \to c)$ as two sides of a triangle, so that the Covenant-Judgement of `agnt` $(a \to c)$ is the hypotenuse. The objective is now to find the length of this hypotenuse.
  2. Next we arithmetize κ_1 and κ_2 using the 'arith' function, and determine the length of the hypotenuse with the Pythagorean theorem.
  3. This arithmetized value of κ_3 is then un-arithmetized onto some $\kappa_3 \in \mathbf{K}''$.

▶ **Example 29.** (`seqd-L` **Composition**). Suppose we have this set of (ρro)Metheus computational `asset` s, and their Covenant-Judgement:

$$\Gamma \models \texttt{agnt}\ (a \to b) : \kappa_2 \in \mathbf{K}, \quad \text{s.t. } \mathsf{arith}(\kappa_2) = 3, \quad \text{and } \mathsf{arith}(\mathbf{K}) = 10,$$
$$\Gamma \models \texttt{agnt}\ (b \to c) : \kappa_6' \in \mathbf{K}', \quad \text{s.t. } \mathsf{arith}(\kappa_6) = 7, \quad \text{and } \mathsf{arith}(\mathbf{K}') = 20,$$
$$\text{and } \mathsf{arith}(\mathbf{K}'') = 12, \quad \text{with } \mathbf{K}'' = \{\kappa_0'', \kappa_1'', \ldots, \kappa_{11}''\}.$$

Then we know that under $\rightarrowtail$ composition, the Covenant-Judgement of the combined value is:

$$\texttt{proj}(\kappa_2 \otimes \kappa_6') := \mathsf{arith}(\mathbf{K}'') \times \sqrt{\left(\frac{3}{10}\right)^2 + \left(\frac{7}{20}\right)^2} = 5.53,$$
$$\text{let } \texttt{agnt}\ (a \to c) := \texttt{agnt}\ (a \to b) \rightarrowtail \texttt{agnt}\ (b \to c),$$
$$\Gamma \models \texttt{agnt}\ (a \to c) : \kappa_4''.$$

Where going into the last line, we round 5.53 down to the value 5, which correspond to $\kappa_4'' \in \mathbf{K}''$.

4.3.5 Semantics of Function Application w.r.t Covenant-Judgement

The next rule defines how `train` changes Covenant-Judgement of the program output.

`train-L`

$$\frac{\Gamma \models \texttt{agnt}\ (a \to b) : \kappa_1, \Gamma \models \texttt{data}\ a : \kappa_2, \Gamma \models \texttt{data}\ b : \kappa_3, \kappa_1 \in \mathbf{K}, \kappa_2 \in \mathbf{K}', \kappa_3 \in \mathbf{K}''}{\Gamma \models \texttt{train}(\texttt{data}\ a, \texttt{data}\ b, \texttt{agnt}\ (a \to b)) : \ \exists \kappa_3 \in \mathbf{K} \text{ s.t. } \kappa_3 = \texttt{proj}(\kappa_1 \otimes \kappa_2 \otimes \kappa_3)}$$

$$\text{where } \texttt{proj}(\kappa_1 \otimes \kappa_2 \otimes \kappa_3) := \mathsf{arith}^{-1}\left(\frac{\mathsf{arith}(\kappa_1) \cdot \mathsf{arith}(\kappa_2) \cdot \mathsf{arith}(\kappa_3)}{\mathsf{arith}(\mathbf{K}) \cdot \mathsf{arith}(\mathbf{K}') \cdot \mathsf{arith}(\mathbf{K}'')}\right).$$

In `train-L` the `train` function `train`s some `agnt` $(a \to b)$ using `data` a and `data` b, each one judged as κ_1 and κ_2. In this case, the `train`ed `agnt` $(a \to b)$ is in the product space of the three Covenant-Judgements. The last rule defines how `infer` changes Covenant-Judgement of the program output.

`infer-L`

$$\frac{\Gamma \models \texttt{agnt}\ (a \to b) : \kappa_1, \Gamma \models \texttt{data}\ a : \kappa_2, \kappa_1 \in \mathbf{K}, \kappa_2 \in \mathbf{K}'}{\Gamma \models \texttt{infer}(\texttt{data}\ a, \texttt{agnt}\ (a \to b)) : \ \exists \kappa_3 \in \mathbf{K}'' \text{ s.t. } \kappa_3 = \texttt{proj}(\kappa_1 \otimes \kappa_2)}$$

$$\text{where } \texttt{proj}(\kappa_1 \otimes \kappa_2) := \text{arith}^{-1}\left(\frac{\text{arith}(\kappa_1)\cdot\text{arith}(\kappa_2)}{\text{arith}(\mathbf{K})\cdot\text{arith}(\mathbf{K}')}\right).$$

In `infer-L`, the `infer` function runs some `agnt` $(a \to b)$ against input `data` a to output prediction or response `data` b. The resultant `data` b is in the product space.

▶ **Example 30.** (**`train-L` Composition**) Suppose we have this set of (ρro)Metheus computational `asset` s, and their Covenant-Judgement:

$$\Gamma \models \texttt{agnt}\ (a \to b) : \kappa \in \mathbf{K}, \quad \text{s.t. arith}(\kappa) = 5, \quad \text{and arith}(\mathbf{K}) = 10, \ \mathbf{K} = \{\kappa_0, \ldots, \kappa_9\}.$$
$$\Gamma \models \texttt{data}\ a : \kappa' \in \mathbf{K}', \quad \text{s.t. arith}(\kappa') = 25, \quad \text{and arith}(\mathbf{K}') = 30.$$
$$\Gamma \models \texttt{data}\ b : \kappa'' \in \mathbf{K}'', \quad \text{s.t. arith}(\kappa'') = 18, \quad \text{and arith}(\mathbf{K}'') = 20.$$

Then we know that under `train`ing, the Covenant-Judgement of the output `agnt` $(a \to b)'$ is:

$$\texttt{proj}(\kappa \otimes \kappa' \otimes \kappa'') := \frac{5 \cdot 25 \cdot 18}{10 \cdot 30 \cdot 20} \cdot 10 = 3.75$$
$$\text{let } \texttt{train}(\texttt{data}\ a, \texttt{data}\ b, \texttt{agnt}\ (a \to b)) := \texttt{agnt}\ (a \to b)'$$
$$\Gamma \models \texttt{agnt}\ (a \to b)' : \kappa_2 \in \mathbf{K}'.$$

Where going into the last line, we round down to the value 3, which correspond to $\kappa_2 \in \mathbf{K}$. Observe that training on lower quality data lowers the quality of the model `agnt` $(a \to b)$.

▶ **Example 31.** (**`infer-L` Composition**) Suppose we have this set of (ρro)Metheus computational `asset` s, and their Covenant-Judgement:

$$\Gamma \models \texttt{agnt}\ (a \to b) : \kappa \in \mathbf{K}, \quad \text{s.t. arith}(\kappa) = 5, \quad \text{and arith}(\mathbf{K}) = 10,$$
$$\Gamma \models \texttt{data}\ a : \kappa' \in \mathbf{K}', \quad \text{s.t. arith}(\kappa') = 19, \quad \text{and arith}(\mathbf{K}') = 30,$$
$$\mathbf{K}'' = \{\kappa_0, \ldots, \kappa_{24}\}.$$

Then we know that under `infer`ence, the Covenant-Judgement of the output `data` b is:

$$\texttt{proj}(\kappa \otimes \kappa') := \frac{5 \cdot 19}{10 \cdot 30} \cdot 24 = 22.799$$
$$\text{let } \texttt{infer}(\texttt{data}\ a, \texttt{agnt}\ (a \to b)) := \texttt{data}\ b$$
$$\Gamma \models \texttt{data}\ b : \kappa_{21} \in \mathbf{K}''.$$

Where going into the last line, we round 22.799 down to the value 22, which correspond to $\kappa_{21} \in \mathbf{K}''$.

▶ Remark 32. (**Computational efficiency of `seqd-L`, `train-L`, and `infer-L` Composition**) Observe that in the examples above, we used simple arithmetic to find the Covenant-Judgement of the composed value. Notably, in example 30 we did not run the `train` function, which would be quite expensive. [11] And in example 31, we did not run the `infer` function, whose cost is also not trivial. This satisfies the *static check* aspect of the (ρro)Metheus environment feature, whereby code that could potentially lead to "bad" effects are not run at all. In this case the "bad" outcome may include:

- inference on poor input data, leading to low quality output;

[11] Indeed a single training run may cost 10s of millions of USD.

- training on poor quality data, leading to lower quality model `agnt`.

▶ Remark 33. (**The role of** FairCovenant **Foundation in defining inductive** Loare**-**Logic judgements) Observe that the `proj()` function we defined for `train` can never improve the Covenant-Judgement of the trained `agnt`, it can only degrade its value. This `proj()` may be too stringent in production. In reality the stringency of `proj()` is context dependent, and will be defined in conjunction with stakeholders as work streams within the FairCovenant Foundation.

4.3.6 Proof of Covenant-Compliance with Loare-Logic

This section places the elementary concepts of Loare-Logic introduced in the previous section in the context of an example, and walk through how a (ρro)Metheus compiler uses Loare-Logic to enforce Covenant-Judgement compliance.

▶ **Example 34.** (**Proof of Covenant-Compliance**) Given Covenant-Judgements sets:

$$\begin{aligned}\text{data Covenant-Judgement}: \quad & \mathbf{K}_a = \{\kappa_1^a, \ldots, \kappa_{10}^a\}, \quad \mathbf{K}_b = \{\kappa_1^b, \ldots, \kappa_{10}^b\}, \\ & \mathbf{K}_c = \{\kappa_1^c, \ldots, \kappa_{10}^c\}, \quad \mathbf{K}_d = \{\kappa_1^d, \ldots, \kappa_{10}\},\end{aligned}$$

$$\begin{aligned}\text{function Covenant-Judgement}: \quad & \mathbf{K}_{f_1} = \{\kappa_1^{f_1}, \kappa_2^{f_1}, \kappa_3^{f_1}, \kappa_4^{f_1}, \kappa_5^{f_1}\}, \quad \mathbf{K}_{f_2} = \{\kappa_1^{f_2}, \kappa_2^{f_2}, \kappa_3^{f_2}\}, \\ & \mathbf{K}_{f_3} = \{\kappa_1^{f_3}, \kappa_2^{f_3}, \kappa_3^{f_3}\}.\end{aligned}$$

And available computational `asset` with Covenant-Judgement:

$$\begin{aligned}&\Gamma \models \mathsf{x}_1 :: \texttt{data}\ a : \kappa_9^a, \quad \Gamma \models \mathsf{x}_2 :: \texttt{data}\ a : \kappa_6^a, \quad \Gamma \models \mathsf{x}_3 :: \texttt{data}\ a : \kappa_7^a, \\ &\Gamma \models \mathsf{y} :: \texttt{data}\ b : \kappa_8^b \\ &\Gamma \models \mathbf{fn}_1 :: \texttt{agnt}\ (\texttt{data}\ a \to \texttt{data}\ c) : \kappa_5^{f_1}, \\ &\Gamma \models \mathbf{fn}_2 :: \texttt{agnt}\ (\texttt{data}\ c \to \texttt{data}\ d) : \kappa_2^{f_2}.\end{aligned} \tag{13}$$

Additionally, we know that there is some:

- $\kappa_{\mathsf{j}}^{f_3} \in \mathbf{K}_{f_3}$ s.t. $\Gamma \models \texttt{agnt}\ (\texttt{data}\ a \to \texttt{data}\ d) : \kappa_{\mathsf{j}}^{f_3}$.
- $\kappa_{\mathsf{i}}^{d} \in \mathbf{K}_{d}$ s.t. $\Gamma \models \texttt{asset}\ : \kappa_{\mathsf{i}}^{d}$.

Now suppose the user builds a Typed ρ-Calculus expression as follows:

$$\mathsf{d}_* = \texttt{infer}\ \mathsf{x}_1 \underbrace{\Big(\underbrace{\texttt{train}\ \underbrace{(\mathsf{x}_2 \diamond \mathsf{x}_3}_{(2)},\ \mathsf{y})\ \underbrace{(\mathbf{fn}_1 \rightarrowtail \mathbf{fn}_2)}_{(1)}}_{(3)}\Big)}_{(4)}. \tag{14}$$

Prompting the program to train some model $\mathbf{fn}_1 \rightarrowtail \mathbf{fn}_2$ on dataset $(\mathsf{x}_2 \diamond \mathsf{x}_3, \mathsf{y})$, before doing inference on input x_1. Now given a client-defined set of boundary condition on the quality of the output to expression (14) with:

$$\kappa_*^d \geq \kappa_5^d.$$

Stating the output of expression (14) must witness some covenant-value greater than or equal to κ_5^d. Then we can construct a proof that witnesses the Covenant-Compliance of expression (14) as follows:

1. Using the rule `seqd-L`, function composition under $\rightarrowtail$ reduces to:

$$\texttt{seqd-L}\ \frac{\Gamma \models \mathbf{fn}_1 : \kappa_5^{f_1}, \quad \Gamma \models \mathbf{fn}_2 : \kappa_2^{f_2}}{\Gamma \models (\mathbf{fn}_1 \ \rightarrowtail\ \mathbf{fn}_2) : \kappa_3^{f_3}}$$

   Where we determine the value of $\kappa_3^{f_3}$ with:

$$\texttt{proj}(\kappa_5^{f_1} \ \otimes\ \kappa_2^{f_2}) = 3 \times \sqrt{\left(\frac{5}{5}\right)^2 + \left(\frac{2}{3}\right)^2} = 3.60 \sim 3.$$

   So we have in the conclusion: $\Gamma \models (\mathbf{fn}_1 \ \rightarrowtail\ \mathbf{fn}_2) : \kappa_3^{f_3}$.

2. Using the rule `compb-L`, we have:

$$\texttt{compa-L}\ \frac{\Gamma \models \mathsf{x}_2 : \kappa_6^a, \quad \Gamma \models \mathsf{x}_3 : \kappa_7^a}{\Gamma \models (\mathsf{x}_2 \ \diamond\ \mathsf{x}_3) : \kappa_6^a}$$

3. Using the rule `train-L`, we have:

$$\texttt{train-L}\ \frac{\Gamma \models (\mathsf{x}_2 \ \diamond\ \mathsf{x}_3) : \kappa_6^a, \quad \Gamma \models \mathsf{y} : \kappa_8^b, \quad , \Gamma \models (\mathbf{fn}_1 \ \rightarrowtail\ \mathbf{fn}_2) : \kappa_3^{f_3}}{\Gamma \models \texttt{train}\ (\mathsf{x}_2 \ \diamond\ \mathsf{x}_3, \mathsf{y})(\mathbf{fn}_1 \ \rightarrowtail\ \mathbf{fn}_2) : \kappa_2^{f_3}}$$

   Where we determined the value of projected Covenant-Judgement with:

$$\texttt{proj}(\kappa_6^a \ \otimes\ \kappa_8^b \ \otimes\ \kappa_3^{f_3}) = 3 \times \frac{6 \times 8 \times 3.60}{10 \times 10 \times 3} = 1.72 \sim 2.$$

4. Finally using `infer-L` we have:

$$\texttt{infer-L}\ \frac{\Gamma \models \mathsf{x}_1 : \kappa_9^a, \quad \Gamma \models (\mathbf{fn}_1 \ \rightarrowtail\ \mathbf{fn}_2) : \kappa_2^{f_3}}{\Gamma \models \mathsf{d}_* : \kappa_6^d}$$

   Where we inductively define the output Covenant-Judgement with:

$$\texttt{proj}(\kappa_9^a \ \otimes\ \kappa_2^{f_3}) = 10 \times \frac{9 \times 2}{10 \times 3} = 6.$$

   And so we have:

$$\Gamma \models \mathsf{d}_* : \kappa_6^d,$$

   witnessing the expected output value of expression (14) to be greater than the boundary condition of κ_5^d. Thus the (ρro)Metheus has used Loare-Logic to generate a proof stating the computation is Covenant-Compliant

We can rewrite this entire proof in shorter format with:

$$\frac{\dfrac{\dfrac{\Gamma \models \mathbf{fn}_1 : \kappa_5^{f_1} \qquad \Gamma \models \mathbf{fn}_2 : \kappa_2^{f_2}}{\Gamma \models (\mathbf{fn}_1 \ \rightarrowtail\ \mathbf{fn}_2) : \kappa_3^{f_3}} \qquad \dfrac{\Gamma \models \mathsf{x}_2 : \kappa_6^a \qquad \Gamma \models \mathsf{x}_3 : \kappa_7^a}{\Gamma \models (\mathsf{x}_2 \ \diamond\ \mathsf{x}_3) : \kappa_6^a} \qquad \Gamma \models \mathsf{y} : \kappa_8^b}{\Gamma \models \texttt{train}(\mathsf{x}_2 \ \diamond\ \mathsf{x}_3, \mathsf{y})\ (\mathbf{fn}_1 \ \rightarrowtail\ \mathbf{fn}_2) : \kappa_2^{f_2}} \qquad \Gamma \models \mathsf{x}_1 : \kappa_9^a}{\Gamma \models \mathsf{d}_* = (\texttt{infer}\ \mathsf{x}_1 (\mathbf{fn}_1 \ \rightarrowtail\ \mathbf{fn}_2)) : \kappa_6^d}$$

This is known as a proof tree, or in this context: a Loare-Logic proof tree.

This exact proof is generated by the (ρro)Metheus compiler before any training or inference is conducted. So that if the proposed computation in expression (14) is Covenant-Compliant, then this proof can be produced as a witness. Otherwise the compiler halts the computation and asks the analyst to input higher quality `asset` into the Typed ρ-Calculus expression, so as to satisfy client-defined notions of quality. This is called static guarantee of Covenant-Compliance. The Typed ρ-Calculus language thus has satisfies the following desiderata:

- (*Monetary*) save money on training and inference by checking the *expected* quality of the final computation, *before* any expensive computation is done client side.
- (*Legality*): Loare-Logic has shifted the burden of proof from the analyst to the compiler. Thereby opening up a greater set of inference tasks that can be done by analysts in high-stake settings with legal responsibilities.
- (*Composability*): the proof tree generated by Loare-Logic used elementary definitions of Covenant-Judgement generated by The FairCovenant Foundation, and boundary conditions of inference quality defined by the client. It then inferred the final Covenant-Judgement of the computation without further human input, thereby greatly augmenting the capacity of regulators to audit the computation trees written by analysts deep in the bowels of their work routine.
- (*End-to-end-oversight*): although the computation in expression (14) appear simple and "one lined," it in fact may span many days or even weeks. As the `train` function alone may be an industrial training run of some deep learning model. The (ρro)Metheus compiler could audit this industrial routine a-priori in its entirety, or audit aspects of the computation over the course of its realization. This gives flexibility as there may be some discrepancy between the *expected* quality of the output under Covenant-Judgement, and its actual quality after a training run. This flexibility is particularly salient in the case of deep learning models, which feature nonconvex objective functions with indeterminate quality a-priori training.
- (*Flexibility*): in this example we used author-defined notions of `proj()` and arithmetization functions. However in real life, it would be the stakeholders that define said functions. They may dial up or down the restrictiveness of the proof at each step along the proof tree generated above. Thereby controlling the amount of computation they deem acceptable.

▶ **Example 35.** (Typed ρ-Calculus **program decorated by** Loare-Logic **triples**) We can rewrite example (34) by decorating the expression with Loare-Logic triples with:

$$
\begin{aligned}
&\left\{\left\{\Gamma \models \mathsf{x}_1 : \kappa_9^a, \mathsf{x}_2 : \kappa_6^a, \mathsf{x}_3 : \kappa_7^a, \mathsf{y} : \kappa_8^b, \mathbf{fn}_1 : \kappa_5^{f_1}, \mathbf{fn}_2 : \kappa_2^{f_2}\right\}\right\} \\
\mathsf{d}_* = \ &\texttt{infer} \\
&\quad \mathsf{x}_1 \\
&\quad \texttt{train}\ \left(\mathsf{x}_2 \diamond \mathsf{x}_3, \mathsf{y}\right)\ \left(\mathbf{fn}_1 \rightarrowtail \mathbf{fn}_2\right) \\
&\quad \left\{\left\{\Gamma \models \kappa_*^d \geq \kappa_5^d\right\}\right\}
\end{aligned}
\tag{15}
$$

This conforms to the syntax presented in definition (26).

5 Extending Typed ρ-Calculus with Causal-Intelligence

This section accomplishes four items:

- Extends the basic Typed ρ-Calculus language to incorporate Pearl's do-Calculus for causal discovery and counterfactual queries.
- Enrich the definition of Loare-Logic so that its Covenant-Judgements cover the language extension stated in point one.
- Demonstrate some practical examples of how Typed ρ-Calculus enriched with causal discovery logic can be used to diagnose the causality of events in practice. And how Loare-Logic is threaded through this analysis to guarantee the outcome of causal inference contractually meets the standards set forth by the client.

5.1 Why Estimating Causality is Valuable in Applied AI

5.1.1 A Brief History of AI

The march towards general artificial intelligence has been a long one. In "the beginning," scholars built AI using hard-coded logical rules that require an immense amount of effort, but cannot generalize outside of a narrow predefined domain. Two prominent examples are the dialogue system of ELIZA in the 1960s, and early chess programs that used tree search to compute the best next move given the current board. In both cases, the combinatorial explosion of possibilities proved to be a *hard ceiling* for classical methods. The chatbot ELIZA could not converse beyond a few sentences. And chess programs cannot scale to solve Go: a far more subtle game featuring 10^{170} positions, versus chess' 10^{50} possible moves.

Figure 8 ELIZA was an early chatbot developed at MIT in the late 1960s. It used hard coded grammar and semantics rules to fool the human eye, however it was quite brittle in real life. The developers thought they could improve the system to pass the Turing test within the summer, but this proved to be too ambitious given the techniques and hardware of the era. Decades later, this outcome was finally achieved with large language models (LLMs). And yet the work is far from finished. From a technical standpoint, LLMs are information retrieval systems with high recall and dubious precision. From an allegorical standpoint, LLMs are "ghosts of the internet" echoing endlessly to possess the human interlocutor. They cannot think anymore than ghosts can die (again).

What was needed was not more rules, but also a more principled way to express uncertainty about knowledge of the underlying reality itself. This underlying reality could manifest itself in i.e. the possible next best move on some game board, or the ambiguous meaning of words in a sentence. One principled way of reasoning about uncertainty is expressed with Bayesian inference over graphical models, whose structure represents intuitive or subjective notions of uncertainty of the underlying reality. The advent of internet-scale datasets and GPUs pushed Bayesian-network-based methods into

industry, culminating in the deep learning revolution of the 2010s. The Bayesian perspective proved flexible enough to express classical statistical workhorses such as regression. And the combination of these disciplines led to several major breakthroughs:

- Breakthroughs such as convolutional neural networks (CNNs), which could identify objects in images and video with higher rate of accuracy than humans. Thereby laying the ground work for truly automated drones and self-driving cars, which could now navigate in human-environments using cameras as an elementary sensor.
- Breakthroughs such as AlphaGo, which defeated the world's top-ranked human players. This feat was previously unthinkable with classical methods that relied on brute force or even heuristically-pruned tree search. However, with a combination of CNNs that represent the Go-board in compressed format, and using deep reinforcement learning to define the best strategy to do tree search, the game of Go has finally been solved by a computer in a definitive manner.
- Breakthroughs such as large language models, which could converse in "general" terms with a human interlocutor. Thus effectively defeating the Turing test, previously the "most hallowed bar" of human intelligence. Moreover, the methods underlying LLMs proved robust enough to handle a variety of time-series data, including multi-modal data (read: music and video) . This makes it possible to generate digital avatars, music, not to mention computer code itself. This is profound: LLMs are automating a significant chunk of the artisanal market for human talent. Thus AI has penetrated into the vocational class of artists, musicians, and junior software programmers.

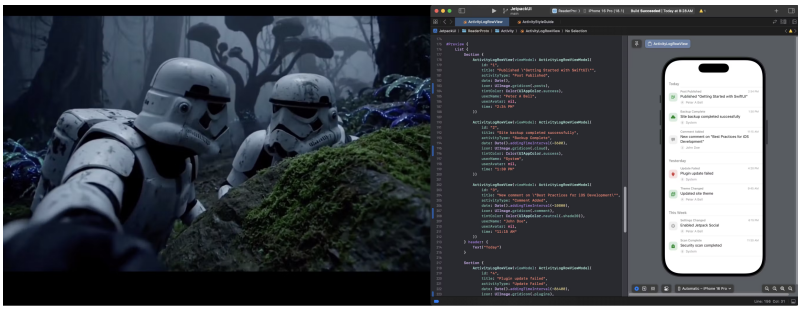


Figure 9 Multi-modal LLMs are penetrating the artisanal class of jobs. (Left): The Instagram account **StormtrooperVlogs** produce entertaining Star Wars videos with high production value. All for a fraction of the cost: sans green screen, special effects capex, actors who cannot act, or George-Lucas-flavored dialogue. (Right): vibe coding has become both an internet meme and very real productivity driver. Entry level coders used to write the same boiler plate code over and over again. Now they are automated away with LLM-agents that sample from GitHub's vast repo of existing code.

5.1.2 The Next Frontier of Applied AI is Causal-Intelligence

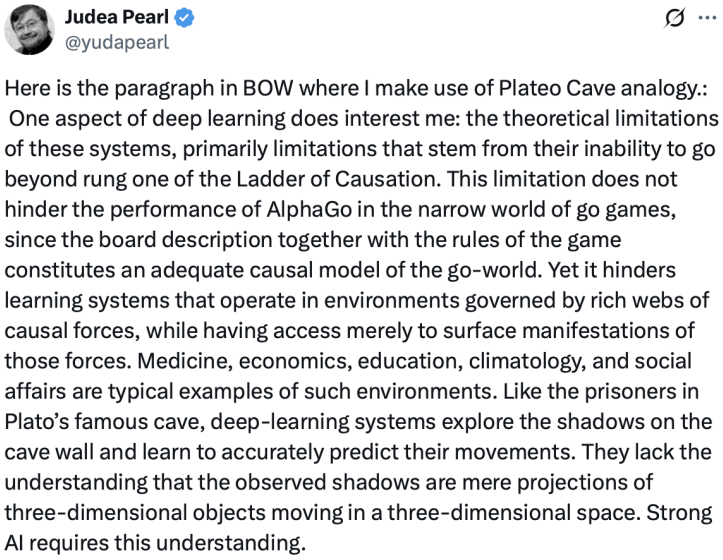
The next frontier in artificial intelligence is to penetrate the professional caste of white-collar workers. In this domain, LLMs may suffice at the entry level, whereby most of the work is document retrieval and summarization. That is to say: grunt work. [12] However at the associate level and above, AI agents must develop a refined understanding of how and why the world works if it is to be truly productive. [13] And yet it is at this exact level that current deep learning methods shall hit a ceiling, as hard and definite as the ceiling that plagued the classical AI methods. This glass-ceiling in intelligence is *the*

[12] Cue "plz fix, k thnx" joke.
[13] One has to wonder: do electric sheep dream of office politics?

barrier of causality. Unlike prior walls in deep learning over the last decade, this wall is baked into the syntax of the mathematical language underpinning deep learning itself. Therefore this ceiling cannot be shattered by bigger data or more compute.

Deep-learning-based AI models such as LLMs will inevitably hit the ceiling of causality. This ceiling cannot be shattered by bigger data or more compute.

The limitation of deep learning is not lost to those who study causality:



- Despite its success, the system is fundamentally limited in that the statistical language underpinning said models can only express correlation, not causation. Pearl was an early proponent of applying Bayesian inference over probabilistic graphical models to express correlation in data. Although he was awarded the Turing prize for this work, Pearl nonetheless developed the field of causality to address the limitation of Bayesian inference that he himself pioneered. [14]
- Presently, the elementary concepts all machine learning scientists must master are Bayesian networks, optimization techniques expressed using linear algebra, and various formulations of regression. However neither advanced master students nor PhD's in machine learning are exposed to the questions of causality; in fact they are often actively discouraged from looking in this direction. Consequentially, the current field of AI is somewhat blind to the topic at the human level, especially in light of the LLM craze.
- In more blunt terms, the contribution of Causal-Intelligence is *undervalued* by a factor of ten, if not more. Here the magnitude is important, as the median compensation at OpenAI among machine learning scientists is 1.2 million USD, while Facebook is paying some engineers 100 million USD over four years to develop native video generation pipelines.

[14] The Turing award and the Gödel Prize are the highest honor in computer science, comparable to the Nobel prize.

- However no amount of gratuitous capex will help said models find causality. This does not affect Facebook, whose use case for generative AI is in the realm of the mundane. However in critical decision making roles where (ρro)Metheus occupies, causal inference is a must-have. Over time, the market will rationalize and reappraise the value of Causal-Intelligence, and Pearl's work on causation will surely be added to any core curriculum in machine learning.
- Determining causality is particularly relevant w.r.t (ρro)Metheus' target market. For example, in the context of insurance, proving beyond a reasonable doubt that some party caused an event to occur is the first step in assigning responsibility to this party, and soon thereafter: liability. Thus there is a direct line from causality to the exchange of money that does not exist in more primitive deep learning models. The monetary case of Causal-Intelligence could not be more clear:

Causality = Responsibility = Liability

Once machine intelligence can understand causality, the industry can finally move beyond artificial intelligence inference to *machine intelligence adjudication*.

5.2 Technical Preliminary

This section introduces the basic concepts needed to speak of causality in a formal manner. Unlike prior chapters whose intellectual basis is drawn from programming language design and classical logic, the content of this section is largely drawn from Pearl's do-Calculus and those who extend it. Namely [Bar25], [J P17], and [Pea09]. We take a narrow pedagogical pathway through the subject so as to present a minimal slice, with an emphasis on defining the concepts needed to specify causality. Knowledge of probability, graphical models, and reinforcement learning is assumed. Readers who are unfamiliar with these concepts, or wish to see how we use notation to express said concepts, should read section 6. Finally, any metaphysical discussion on the nature of causality itself, while entertaining, is simply ignored.

5.2.1 The Forward and Backward Question of Causality

All causal questions posed against (ρro)Metheus can be divided into two categories:

- **The Forward Question**: in this direction, the system is given a proposition of causality, and must construct a proof from the available dataset witnessing this proposition. Additionally, the system may be asked to present a confidence estimation along with the proof given.
- **The Backward Question**: in the reverse direction, the system is given a dataset with embedded causal information, and must determine one or several causal propositions evidenced in data. Moreover, the system may be asked to rank the proposition by how well they fit the evidence.

Observe: the computational complexity or demand of the two questions are categorically different. The forward question may be determined, while the backward question is often times ill-posed. While the forward question is solved by a simple application of scoring datasets against a proposed causal graph, the backward question may be parsimoniously solved by causal reinforcement learning.

▶ Remark 36. (**Hierarchy of Proofs**) Observe that the proposition and proofs in this case are different from the Loare-Logic proof trees of section 4.3.6. Where as Loare-Logic generates type-level proofs witnessing the validity or Covenant-Compliance of the computation in "cyberspace," the

Causal-Intelligence proof is a term-level proof demonstrating the likely causality of some event in "real space." In short, the Loare-Logic-proof is a judgement on the process of arriving at a conclusion, while the Causal-Intelligence proof is a judgement of the underlying reality. This is what we mean by "two-tiered proof system" in the abstract.

The rest of this chapter will introduce the relevant aspect of do-Calculus used to infer causality from data. This is followed by causal reinforcement learning (CRL), which for our purpose is a "strategic" way to determine the most likely causal graph. Then we will express CRL with Typed ρ-Calculus, so as to augment (ρro)Metheus with the capacity to understand causality.

5.2.2 The Ladder of Causation

This section contextualizes classical machine learning techniques w.r.t hierarchy of causal concepts.

Layer	Activity	Machine Learning
Association $\mathfrak{p}(y\|x)$	Seeing: what does an observation say about the underlying state?	(Un)supervised learning
Intervention $\mathfrak{p}(y\|do(x))$	Doing: how would an action lead to a different outcome.	Reinforcement learning
Counterfactual $\mathfrak{p}(y_x\|y', x')$	Imagining: why? What if the agent acted differently.	Explanation/Transparency

Table 3 Pearl's causal ladder and their colloquial analogues.

The causal hierarchy of table 3 tiers the sophistication of questions queried on data into three tiers:

1. (*Association*): invoking purely statistical relationships defined by raw data. For instance, observing a customer who buys toothpaste makes it more likely that they buy floss. Such association can be inferred directly from the observed data using conditional expectation.
2. (*Intervention*): The second level ranks higher than association because it involves not just seeing what is, but changing what is seen. One question at this level is: "what happens if we double the price of toothpaste?" Such questions cannot be answered from sales data alone, because they involve a change in customers behavior in response to the new price. Customer choices under the new price structure may differ substantially from the past.
3. (*Counterfactual*): A typical question in this category is "what if I were to act differently?" Thus necessitating retrospective reasoning. Counterfactual models are placed at the top of the hierarchy because they subsume intervention and associational questions. If we have a model that can answer counterfactual queries, we can also answer questions about interventions and observations. However, the reverse is not true. Associational questions cannot be used to answer intervention or counterfactual questions. Therefore this ladder is a hierarchy delineating the explanatory power of statistical models [25d; Pea09].

The aim of standard statistical analysis (i.e. regression and estimation) is to assess parameters of a distribution from samples drawn from said distribution. Then using said parameters, one infers associations among variables, and update parameters given new data. One underlying assumption here is that data is drawn i.i.d, which proves restrictive in reality and severely limit the capacity of many statistical models to generalize out of domain. Causal analysis not only infers probabilities under static conditions, but also the dynamics of beliefs under changing conditions. These changes are induced by treatments, new policies, or external interventions [Pea10]. The various levels of the causation ladder is outlined in table 3.

Furthermore, the relationship between associational/correlational and causal concepts with statistical methods is as follows:

- **An associational concept** is a relationship that can be defined in terms of a joint distribution of observed variables. Examples include correlation, regression, dependence, conditional independence, likelihood, odds ratio, marginalization, conditionalization, etc. All statistical models utilizing deep learning tools fall under this category. Language models in particular are a kind of auto-regression on time-series data.
- **A causal concept** is any relationship that cannot be defined from the distribution alone. Examples of causal concepts are randomization, influence, effect, confounding, "holding constant," structural coefficients, faithfulness/stability, instrumental variables, intervention, explanation, and attribution.

Associational concepts are readily expressed by the language of probability. However, probabilistic models, no matter how complex, cannot express simple concepts such as "wet pavement does not cause rain." In theory, causal relations can only be determined with randomized controlled experiments. However such experiments may be expensive to run, or its collected data may be gated under privacy considerations. Sometimes it is simply unethical to run such experiments altogether, i.e.: does inhaling vape cause cancer in children 10 and under?

In order to express causation, the language of probability must be augmented with new symbols or algebra to express causal concepts that are intuitive to the human mind. This is the language of do-Calculus applied to the transformation of directed graphical models, it allows (ρro)Metheus to estimate causal relations from observational data alone. It does so by translating basic intervention or counterfactual queries into operations on graphical models, questions such as does treatment **x** cause outcome **y**. Finally it uses observational data alone to accept or reject said queries.

This ends the prerequisite preliminary needed to set up the (ρro)Metheus Causal-AI adjudication environment. The next subsection augments (ρro)Metheus with causal intelligence. Those who are unfamiliar with do-Calculus and causal reinforcement learning may read section 6 for further technical preliminaries.

5.3 Case Study: Transnational Supply Chain Insurance

This section proceeds in this order:

1. First, it presents one use case for (ρro)Metheus in the context of transnational trade.
2. Then it abstracts the use cases into a general system diagram and workflow.
3. Next, it poses the system flow as an off-policy causal reinforcement learning problem.
4. Finally, key aspects of the workflow is expressed directly in the syntax of Typed ρ-Calculus itself, so that the (ρro)Metheus compiler can generate static proofs confirming the validity of the adjudication process itself.

This section expands upon the basic example given in section 1.1.2, and presents more detailed use cases of how the (ρro)Metheus oracle determines causality given data from multiple sources.

While LLMs replace analysts, (ρro)Metheus causal-AI replaces associates.

5.3.1 User Journey

In the transnational supply chain setting, parcels carrying perishable goods move from a commodity center in East Africa (E.A.C.C.) to a special economic development zone in East Asia (SEZ). The stakeholders are:

1. The *seller* from E.A.C.C. is situated in an underdeveloped economic zone. Here, documentation for quality of parcel is sparse, and technology penetration is low. The seller grows perishable commodity locally, and sells them to the E.A. *reseller*. The reseller marks up the parcel by 50% and ships it across the ocean, with a transit time of 40 days.
2. The *E.A. reseller* deposits the parcel to the *local reseller* in the SEZ, who marks it up an additional 75% and warehouses the product for 14 days.
3. The local reseller then delivers the parcel to *point of sale* (P.O.S) in the SEZ. The P.O.S sells it at retail price with 10% markup.
4. The *buyer* buys parcel and takes it home. He opens the product and finds that it is *spoiled*.

The questions here are:

- Who is responsible for the final condition of the product?
- How much should the responsible parties pay?
- Whomst are they paying? How are the funds disbursed?
- Which players along the supply chain should be investigated further? If the players have reputation rating, how much should their ratings be impacted due to failure to maintain parcel quality?

This is a fundamentally a risk management problem *a-priori* parcel purchase, and claims adjudication process *a-posteriori* parcel delivery. In order the prosecute the outcome appropriately, (ρro)Metheus Causal-Intelligence has the following user touch points:

- Both the seller and reseller in E.A.C.C will have a mobile app equipped with a camera, he takes a picture of the package before dropping it off with the reseller. The image is GPS and timestamped. This is a trivial technical requirement as mobile phone penetration in East Africa is 100%.
- The local resellers at the S.E.Z. shares data from their warehouse over the 14 day period, this may include:
  - continuous measurements of warehouse temperature,
  - as well as timestamp and duration of when products are loaded, etc.
- (ρro)Metheus will also have access to temperature gauge within the shipping container as it completes its transit across the ocean.
- Finally, (ρro)Metheus will have access to a stream of general world events that is ancillary to the parcel transit process. This may include weather reports, news reports of exogenous shock, etc.

In summary, the parcel lifecycle is a data-generation process. The (ρro)Metheus Causal-Intelligence oracle will assign liability based on this data stream. Now there are two possible adjudication regimes: the interactive and the Causal-Intelligence automated.

1. **Stakeholder Interactive Regime**: in this arrangement, each stakeholder along the supply chain interacts with a natural-language-based dialogue-based system. Here are asked to offer:
   - Additional data proving the condition of the product before and after their touch point.
   - A plausible explanation for why the parcel's quality failed the predefined standard.

- Finally, a certified examiner will perform a final review of the information and decide on an outcome. Payouts will be allocated according to defined policies, and the associated parties' insurance premium will be adjusted appropriately.

The interactive adjudication process is also a ground truth, or data labeling process that will be used to train a Causal-Intelligence agent that enables the automated regime.

2. **Causal-Intelligence Automated Regime**: in this setting, the (ρro)Metheus oracle validates the parcel history, and decide on an outcome based on its internal Causal-Intelligence engine.

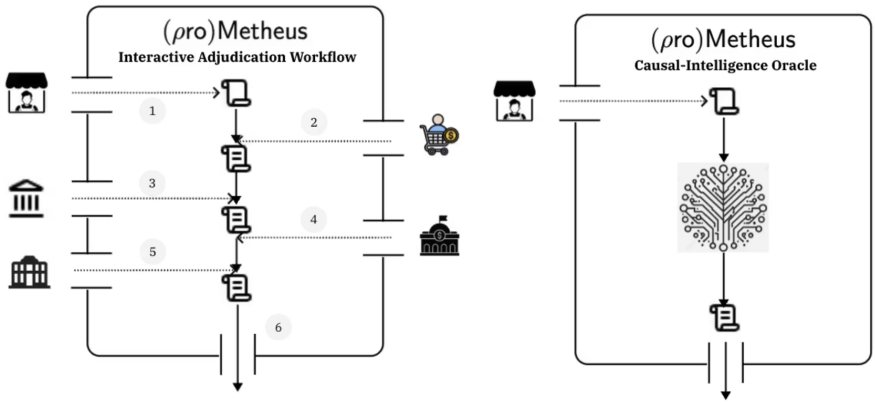


Figure 10 (*Left*): the (ρro)Metheus oracle interactive adjudication workflow follows the classic insurance adjudication process, whereby each stakeholder is asked to either verify or file additional data documenting the quality of parcel under their custody. Finally a certified examiner passes judgement and closes the case with appropriate payouts and adjustments. (*Right*): the (ρro)Metheus Causal-Intelligence automated adjudication process ingests the claim along with data flow generated by trade, and decide on the outcome directly sans additional human input.

5.3.2 Posing the Causal-Intelligence Technical Problem

5.4 Expressing Causality with Typed ρ-Calculus

5.4.1 Typed ρ-Calculus with do-Calculus Primitives

5.4.2 Typed ρ-Calculus with Causal Reinforcement Learning

5.4.3 Loare-Logic Governing Typed ρ-Calculus for Causal-Intelligence

6 Technical Background On Causal Reinforcement Learning

This chapter introduces the relevant aspect of do-Calculus used to infer causality from data. This is followed by causal reinforcement learning (CRL), which for our purpose is a "strategic" way to determine the most likely causal graph.

6.1 Causal Bayesian Networks and do-Calculus

In order to express causation, the language of probability must be augmented with new symbols or algebra to express causal concepts that are intuitive to the human mind. This is the language of do-Calculus applied to the transformation of directed graphical models, it allows (ρro)Metheus to estimate causal relations from observational data alone. It does so by translating basic intervention or counterfactual queries into operations on graphical models, questions such as does treatment $\mathbf{x}$ cause outcome $\mathbf{y}$. Finally it uses observational data alone to accept or reject said queries.

6.1.1 Bayesian Network as the Underlying Data Structure

In inferring causality, the elementary data structure of choice to encode the expert's assumptions and/or express relationships found amongst data is the *causal Bayesian network*, which is a specific type of Bayesian network that explicitly models causal relationships between variables. Governing the relationship among the variables are *structural equations*. They are *deterministic* expressions of how the variables under consideration relate to each other. Moreover, we assume the observed uncertainty in outcome arises from external factors that are not account for by the model. This is the Laplace interpretation of probability. This set of parameterized structural equations relating the factors form a *causal model*.

▶ **Definition 37.** ***(Structural Causal Model)****[Pea09] is a pair* $\mathcal{M} = (\mathcal{G}, \mathcal{F}, \mathcal{U})$ *consisting of:*

- *a causal structure* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *is a directed acyclic graph (DAG) where each edge represent some functional relationship amongst the vertices* $\mathbf{x} \in \mathbf{V}$.
- *A set of endogenous variables* $\{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots\}$ *inside the model.*
- *A set of exogenous variables* $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots\}$ *outside the model.*
- *a set of parameters* $\mathcal{F}$ *consistent with* $\mathcal{G}$*. They assign functions:*

  $$\mathbf{x}_i = f(\mathsf{pa}(\mathbf{x}_i), \mathbf{u}_i) \ ,$$

  to each vertex $\mathbf{x}_i \in \mathbf{V}$*, where* $\mathsf{pa}(\mathbf{x}_i)$ *are the parent vertices of* $\mathbf{x}_i$*.*
- *A probability distribution* $\mathfrak{P}(\mathbf{u}_1, \ldots)$ *over exogenous variables.*

▶ **Example 38.** (**Structural Causal Model**) [Pea09] An example causal model is presented in fig (11). Its structural equations are:

$$\begin{aligned}
&\mathbf{x}_1 \in \{\mathsf{winter}, \mathsf{spring}, \mathsf{summer}, \mathsf{fall}\},\\
&\mathbf{x}_2 := \mathtt{f}(\mathsf{pa}(\mathbf{x}_2), \mathbf{u}_2) \ = [(\mathbf{x}_1 = \mathsf{winter}) \vee (\mathbf{x}_1 = \mathsf{fall}) \vee \mathbf{u}_2] \wedge \neg \mathbf{u}_2',\\
&\mathbf{x}_3 := \mathtt{f}(\mathsf{pa}(\mathbf{x}_3), \mathbf{u}_3) \ = [(\mathbf{x}_1 = \mathsf{summer}) \vee (\mathbf{x}_1 = \mathsf{spring}) \vee \mathbf{u}_3] \wedge \neg \mathbf{u}_3',\\
&\mathbf{x}_4 := \mathtt{f}(\mathsf{pa}(\mathbf{x}_4), \mathbf{u}_4) \ = (\mathbf{x}_2 \vee \mathbf{x}_3 \vee \mathbf{u}_4) \wedge \neg \mathbf{u}_4',\\
&\mathbf{x}_5 := \mathtt{f}(\mathsf{pa}(\mathbf{x}_5), \mathbf{u}_5) \ = (\mathbf{x}_4 \vee \mathbf{u}_5) \wedge \neg \mathbf{u}_5'.
\end{aligned} \tag{16}$$

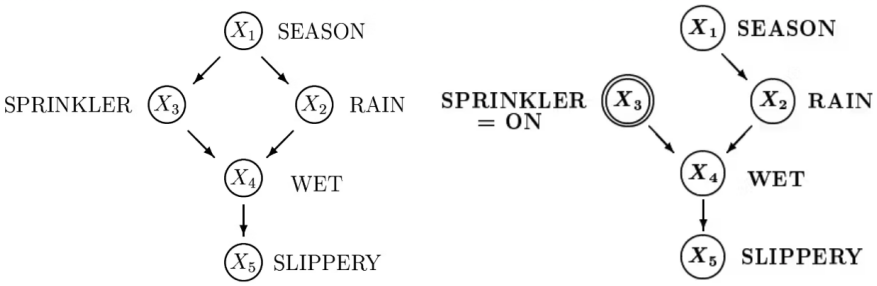


Figure 11 Bayesian networks are directed acyclic graphs (DAGs), where vertices represent variables of interest, and edges represent informational or causal dependency amongst the variables. The strength of the dependency is represented by the conditional probabilities labeling each edge. The example graph on the left expresses the various factors that lead to a slippery pavement [Pea09].

Where $\mathbf{x}_i \in \{0, 1\}$, and $\mathsf{pa}(\mathbf{x}_i)$ refer to the parents of variable $\mathbf{x}_i$. Each $\mathbf{u}_k = 1$, with $\neg\mathbf{u}_k = 0$. They are exogenous terms that make the outcome probabilistic. For example if are given these odds:

$$\mathfrak{p}(\mathbf{u}_2) = 0.6, \qquad \mathfrak{p}(\neg\mathbf{u}_2) = 0.4,$$
$$\mathfrak{p}(\mathbf{u}_2') = 0.2, \qquad \mathfrak{p}(\neg\mathbf{u}_2') = 0.8.$$

Now given $\mathbf{x}_1$ is winter so that $\mathbf{x}_2 = (\mathsf{winter} \vee \mathbf{u}_2) \wedge \neg\mathbf{u}_2'$. Then we have this contingency table:

$$\mathfrak{p}\big(\mathbf{x}_2 = 1 \,\big|\, \mathbf{x}_1 = \mathsf{winter}\big) = \mathfrak{p}(\mathsf{winter} \vee \mathbf{u}_2 \wedge \neg\mathbf{u}_2') = \mathfrak{p}\big(\mathsf{winter} \wedge \neg\mathbf{u}_2' \,\big|\, \mathbf{x}_1 = \mathsf{winter}\big) = 0.8,$$
$$\mathfrak{p}(\mathbf{x}_2 = 0) = 0.2.$$

Suppose instead $\mathbf{x}_1$ is spring so that $\mathbf{x}_2 = (\mathsf{spring} \vee \mathbf{u}_2) \wedge \neg\mathbf{u}_2'$. Then we have this contingency table:

$$\mathfrak{p}\big(\mathbf{x}_2 = 1 \,\big|\, \mathbf{x}_1 = \mathsf{spring}\big) = \mathfrak{p}(\mathsf{spring} \vee \mathbf{u}_2 \wedge \neg\mathbf{u}_2') = \mathfrak{p}\big(\mathbf{u}_2 \wedge \neg\mathbf{u}_2' \,\big|\, \mathbf{x}_1 = \mathsf{spring}\big) = 0.6 \times 0.8 = 0.48$$
$$\mathfrak{p}(\mathbf{x}_2 = 0) = 0.52.$$

The exact set of equations specifying the causal relations is not important here, what is salient is that structural equations are deterministic relations that govern the outcome of variables. Any probability distribution defined over $\mathbf{x}_1, \ldots, \mathbf{x}_5$ is over the uncertain values of $\mathbf{u}_k$, which form the source of randomness. Without the $\mathbf{u}$ factors, the values of x_i is completely determined [Pea09].

Similar to conventional Bayes networks, given a model $\mathcal{M}$ its likelihood is factorized with:

$$\mathfrak{p}(\mathbf{x}_1, \ldots) = \prod_i \mathfrak{p}\big(\mathbf{x}_i \,\big|\, \mathsf{pa}(\mathbf{x}_i)\big). \tag{17}$$

The entries of each $\mathfrak{p}\big(\mathbf{x}_i \,\big|\, \mathsf{pa}(\mathbf{x}_i)\big)$ may be computed according to example (38) above.

6.1.2 Operation on Data Structure with do-Calculus

We can query the graph in figure 11 with the three questions on the ladder of causation of table 3:

1. *Prediction*: would the pavement be slippery if *see* the sprinkler is off? That is we query the probability $\mathfrak{p}\big(\mathbf{x}_4 = \mathsf{Slippery} \,\big|\, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 = \mathsf{Off}, \mathbf{x}_4\big)$.
2. *Intervention*: would the pavement be slippery if we *turn* the sprinkler on? The graph to the right expresses such an intervention whereby the underlying state is set to "sprinkler=ON" regardless of

the condition of the season, which under observed data may affect the status of sprinkler under non-intervened conditions. Note the edge connecting the season vertex to the sprinkler vertex has been pruned as the sprinkler variable is now a constant. That is to say we are computing the probability $\mathfrak{p}\big(\mathbf{x}_5 = \mathsf{Slippery} \;\big|\; \mathbf{x}_1, \mathbf{x}_2, \mathsf{do}(\mathbf{x}_3 = \mathsf{On}), \mathbf{x}_4\big)$.

3. *Counterfactual*: would the pavement be slippery had the sprinkler been off, given we observe the pavement is not slippery and the sprinkler is on? That is we query the likelihood $\mathfrak{p}\big(\mathbf{x}_5 = \mathsf{Slippery}\ \&\ \mathbf{x}_3 = \mathsf{Off} \;\big|\; \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 = \mathsf{On}, \mathbf{x}_4, \mathbf{x}_5 = \neg\mathsf{Slippery}\big)$.

This paper focuses on the second rung of the causal ladder: discovering causal factors by intervention. Since the structural equations of eqn (16) determine a "fixed mechanical system," one can simulate, i.e. intervention of form "turning the sprinkler on," written:

$$\mathsf{do}(\mathbf{x}_3 = \mathsf{On}),$$

by removing equation $\mathbf{x}_3 = \mathtt{f}(\mathsf{pa}(\mathbf{x}_3), \mathbf{u}_3)$ to create the "mutilated graph" of the right picture in fig 11. This pruned graph will be used to determine if the sprinkler variable $\mathbf{x}_3$ *causes* the pavement to be wet.

▶ **Definition 39.** *(**Causal Effect**)[Pea09] Given two variables* $\mathbf{x}$ *and* $\mathbf{y}$ *in some causal model* $\mathcal{M}$*, the causal effect of* $\mathbf{x}$ *on* $\mathbf{y}$*, denoted with:*

$$\mathfrak{p}\big(\mathbf{y} = \mathsf{y} \;\big|\; \mathsf{do}(\mathbf{x} = \mathsf{x})\big), \tag{18}$$

is a function from $\mathbf{x}$ *to the space of probability distributions on* $\mathbf{y}$*. Eqn (18) gives the probability of* $\mathbf{y} = \mathsf{y}$ *induced by deleting from the causal model* $\mathcal{M}$ *all structural equations leading to* $\mathbf{x}$*, and substituting* $\mathbf{x} = \mathsf{x}$ *in the remaining equations in* $\mathcal{M}$*. Alternatively, one can define the difference:*

$$\mathbf{E}_{\mathbf{u}}\big[\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x} = \mathsf{x}_1)\big] - \mathbf{E}_{\mathbf{u}}\big[\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x} = \mathsf{x}_2)\big], \tag{19}$$

as the causal or "average treatment effect" of the $\mathsf{do}(\mathbf{x})$ *operation.*

▶ **Example 40.** (**Causal Effect**) Continuing example of fig (11) in example 38, a new graph where $\mathbf{x}_3$ is set to On is parameterized by the causal model:

$$\begin{aligned} &\mathbf{x}_1 \in \{\mathsf{winter}, \ldots\}, \qquad \mathbf{x}_2 := \mathtt{f}(\mathsf{pa}(\mathbf{x}_2), \mathbf{u}_2)\ , \qquad \mathbf{x}_3 := \mathsf{On}, \\ &\mathbf{x}_4 := \mathtt{f}(\mathsf{pa}(\mathbf{x}_4), \mathbf{x}_3 = \mathsf{On}, \mathbf{u}_4)\ = (\mathbf{x}_2 \vee \mathsf{On} \vee \mathbf{u}_4) \wedge \neg\mathbf{u}_4', \qquad \mathbf{x}_5 := \mathtt{f}(\mathsf{pa}(\mathbf{x}_5), \mathbf{u}_5)\ . \end{aligned} \tag{20}$$

Note that the sprinkler variable $\mathbf{x}_3$ no longer depends on $\mathbf{x}_1$, or the seasons. Meanwhile, in equation $\mathbf{x}_4$ the variable $\mathbf{x}_3$ is set to On [Pea09].

Once the edges incident upon some $\mathbf{x}$ in the causal model have been deleted and its value set to $\mathbf{x} = \mathsf{x}$, we can *identify* whether this factor $\mathbf{x}$ is an actual cause of some downstream effect denoted $\mathbf{y}$. In the absence of experimental data, this is done by computing the *post-intervention probability* using observed data from the *pre-intervention probabilities*. The exact nature of the post-intervention probability after $\mathsf{do}(\mathbf{x})$ adjustment depend on the local topology of the graph around the $\mathbf{x}$ variable. Often, additional adjustments must be made to the graph before the post-intervention likelihood is computed. The three rules of do-Calculus makes these adjustments in a principled manner:

▶ **Theorem 41.** *(**Rules of** do-Calculus)[Pea09; Pea12]*

1. *(**Ignoring observations**): that is if we wish to determine if* $\mathbf{x}$ *causes* $\mathbf{y}$ *after applying* $\mathsf{do}(\mathbf{x})$*, then these vertices may be deleted from the causal graph:*
   - *vertices with no path to* $\mathbf{y}$.
   - *vertices that is d-separated from the outcome* $\mathbf{y}$*. See theorem 42 below.*

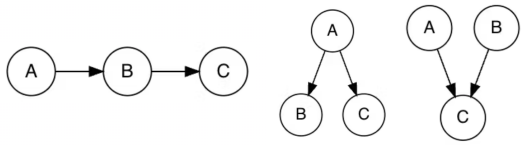


Figure 12 Three types of separation patterns are chain, fork, and collider. In each ase, if the middle vertex in the graph is determined then the path is broken, and the two distal variables are no longer correlated [25c].

- (*Left*): in the chain pattern or "mediation," **b** separates **a** from **c**, s.t. the value of **c** is independent of **a** given **b**
- (*Middle*): in the fork pattern or "mutual dependence," **a** separates **b** from **c**, s.t. the value **c** is independent of **b** given the value of **a**.
- (*Right*): the collider pattern or "mutual causation," **c** separates **a** from **b**, s.t. the values of **a** and **b** are independent given **c**.

Name this vertex **z**. *Now suppose all the edges incident on* **x** *are removed, then this rule is written:*

$$\mathfrak{p}\big(\mathbf{y} \,\big|\, \mathbf{z}, \mathsf{do}(\mathbf{x}), \mathbf{w}\big) = \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x}), \mathbf{w}\big) \qquad \text{if} \quad \mathbf{y} \perp \mathbf{z} \mid \mathbf{w}, \mathbf{x}. \tag{21}$$

Stating that if variable **z** *is independent of outcome* **y** *given the preset values of* **x** *and* **w**, *then* **z** *may be removed from the graph.*

2. ***(Treating interventions as observations)****: this rule states that interventions of form* $\mathsf{do}(\mathbf{x})$ *on outcome* **y** *can be treated as an observation when the causal effect of* **x** *on* **y** *only influences the outcome through directed paths. Suppose we wish to test the causal effect of variable* **z** *on* **y**, *then given a mutilated graph where all the edges going out of* **z** *are removed, we have:*

$$\mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x}), \mathsf{do}(\mathbf{z}), \mathbf{w}\big) = \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}\big) \qquad \text{if} \quad \mathbf{y} \perp \mathbf{z} \mid \mathbf{w}, \mathbf{x}. \tag{22}$$

3. ***(Ignoring interventions)****: sometimes one can ignore an intervention of form* $\mathsf{do}(\mathbf{z})$ *on* **y** *if* **z** *does not influence the outcome* **y** *through any uncontrolled paths. That is to say there is no associational path from from* **z** *to* **y**. *Now given a graph whereby all edges incident on* **x** *has been removed, and all edges outgoing from* **z** *has been removed, then we have:*

$$\mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x}), \mathsf{do}(\mathbf{z}), \mathbf{w}\big) = \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x}), \mathbf{w}\big) \qquad \text{if} \quad \mathbf{y} \perp \mathbf{z} \mid \mathbf{w}, \mathbf{x}. \tag{23}$$

▶ **Theorem 42.** *(d-separation) Given a causal graph with vertices* **V**, *a set* $\mathbb{S} \subseteq \mathbf{V}$ *of vertices is said to block a path from* $\mathbf{x} \in \mathbf{V}$ *to* $\mathbf{y} \in \mathbf{V}$ *if either one of the follow two conditions are true:*

1. *The path contains at least one arrow-emitting vertex that in the set* $\mathbb{S}$. *This is the middle vertex in a chain or fork pattern. See figure (12).*
2. *The path contains at least one collision vertex that is outside* $\mathbb{S}$, *and has no descendant in* $\mathbb{S}$. *This is the middle vertex in a collision pattern, see figure (12).*

If the set of vertices $\mathbb{S}$ *blocks all path from* **x** *to* **y**, *it is said to "d-separate* **x** *and* **y**." *Two sets* $\mathbb{A} \subseteq \mathbf{V}$ *and* $\mathbb{B} \subseteq \mathbf{V}$ *are said to be d-separated by* $\mathbb{S}$ *if any pair* $\mathbf{x} \in \mathbb{A}$ *and* $\mathbf{y} \in \mathbb{B}$ *are d-separated by* $\mathbb{S}$. *If so, then we have the following property:* $\mathbb{A} \perp \mathbb{B} \mid \mathbb{S}$.

▶ **Example 43.** (**Rule 1**)[25c]

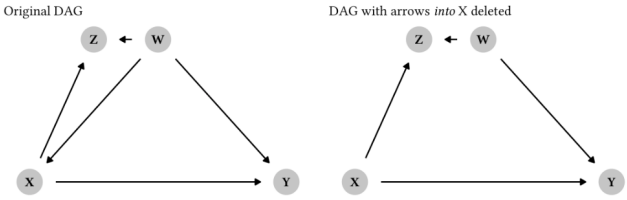


Figure 13 (*Left*): the original DAG with causal functions incident on $\mathbf{x}$ from $\mathbf{w}$. (*Right*): we query the graph for $\mathfrak{p}\left(\mathbf{y} \mid \mathsf{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}\right)$. Since the vertex triple ($\mathbf{z}$, $\mathbf{w}$, $\mathbf{y}$) form the "mutual independence" pattern of fig (12), we have the reduced post-intervention distribution $\mathfrak{p}\left(\mathbf{y} \mid \mathsf{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}\right) = \mathfrak{p}\left(\mathbf{y} \mid \mathsf{do}(\mathbf{x}), \mathbf{w}\right)$ by rule one.

▶ **Example 44.** (**Rule 2**)[25c]

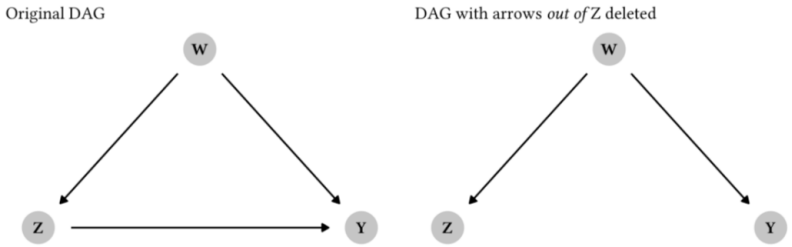


Figure 14 (*Left*): the original graph where we wish to query $\mathbf{z}$ for its causal effect on $\mathbf{y}$. (*Right*): the mutilated graph where edges going out of $\mathbf{z}$ has been removed. Note that this is the fork pattern, so we can write $\mathfrak{p}\left(\mathbf{y} \mid \mathsf{do}(\mathbf{z}), \mathbf{w}\right) = \mathfrak{p}\left(\mathbf{y} \mid \mathbf{z}, \mathbf{w}\right)$ by rule 2.

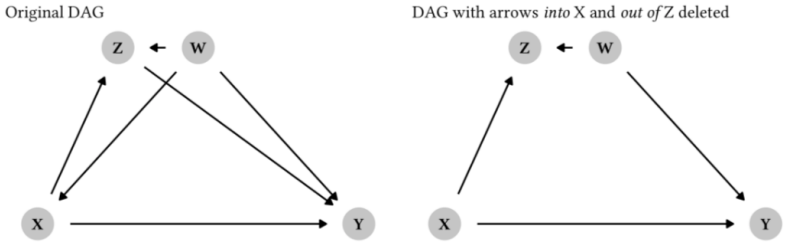


Figure 15 (*Left*): the original graph where we wish to query $\mathbf{z}$ for its causal effect on $\mathbf{y}$. (*Right*): the mutilated graph where edges going out of both $\mathbf{x}$ and $\mathbf{z}$ have been removed. Now observe the fork pattern with the triplet ($\mathbf{z}$,$\mathbf{w}$,$\mathbf{y}$) where $\mathbf{y} \perp \mathbf{z} \mid \mathbf{w}$, so we can write $\mathfrak{p}\left(\mathbf{y} \mid \mathsf{do}(\mathbf{z}), \mathsf{do}(\mathbf{x}), \mathbf{w}\right) = \mathfrak{p}\left(\mathbf{y} \mid \mathsf{do}(\mathbf{x}), \mathbf{z}, \mathbf{w}\right)$ by rule 2.

▶ **Example 45.** (**Rule 3**)[25c]

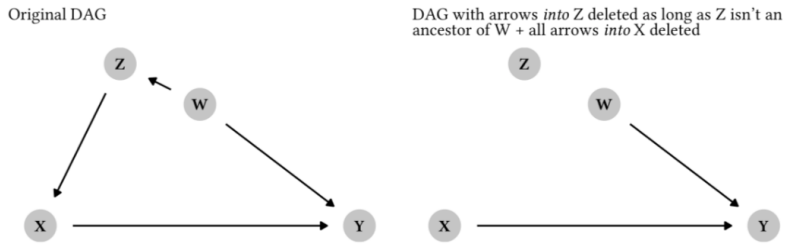


Figure 16 (*Left*): original graph. (*Right*): $\mathsf{do}(\mathbf{x})$ is applied s.t. when the edges incident on $\mathbf{x}$ are removed, $\mathbf{z}$ does not influence $\mathbf{y}$ in any way, so that its edges are deleted via $\mathfrak{p}\left(\mathbf{y} \mid \mathsf{do}(\mathbf{x}), \mathsf{do}(\mathbf{z}), \mathbf{w}\right) = \mathfrak{p}\left(\mathbf{y} \mid \mathsf{do}(\mathbf{x}), \mathbf{w}\right)$.

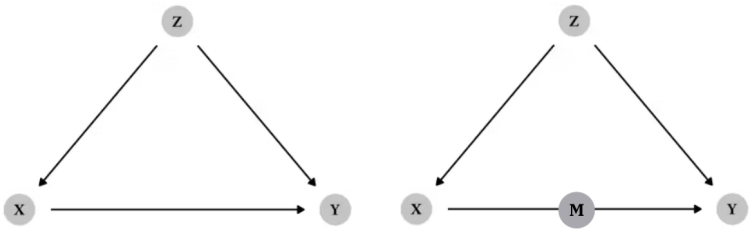


Figure 17 (*Left*) Back-door criterion: here $\mathbf{z}$ is a confounding variable influencing both $\mathbf{x}$ and $\mathbf{y}$, so that $\mathbf{x}$ is independent from $\mathbf{y}$ given $\mathbf{z}$. (*Right*) front-door criterion, where $\mathbf{z}$ is an *unobserved* confounding variable influencing both $\mathbf{x}$ and $\mathbf{y}$, while $\mathbf{m}$ is an observed variable blocking the path from $\mathbf{x}$ to $\mathbf{y}$.

Using the three algebraic rules of do-Calculus above, we can derive the following two expressions that turn any post-intervention distribution into observational distributions, so that the effect of an intervention can be identified from observational data alone.

► **Proposition 46.** *(**Back-door Criterion**)[Pea09] Given the left causal graphical model in fig (17), we wish to query the graph for the effect of treatment* $\mathbf{x}$ *on outcome* $\mathbf{y}$ *by* $\mathsf{do}(\mathbf{x})$*. Then we can reduce the query as follows:*

$$\mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x})\big) = \sum_{\mathbf{z}} \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathbf{x}, \mathbf{z}\big)\mathfrak{p}(\mathbf{z}). \tag{24}$$

Proof.

$$\begin{aligned}
\mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x})\big) &\overset{(1)}{=} \sum_{\mathbf{z}} \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x}), \mathbf{z}\big)\mathfrak{p}\big(\mathbf{z} \,\big|\, \mathsf{do}(\mathbf{x})\big) \\
&\overset{(2)}{=} \sum_{\mathbf{z}} \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathbf{x}, \mathbf{z}\big)\mathfrak{p}\big(\mathbf{z} \,\big|\, \mathsf{do}(\mathbf{x})\big) \\
&\overset{(3)}{=} \sum_{\mathbf{z}} \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathbf{x}, \mathbf{z}\big)\mathfrak{p}\big(\mathbf{z} \,\big|\, \cdot\big) \\
&= \sum_{\mathbf{z}} \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathbf{x}, \mathbf{z}\big)\mathfrak{p}(\mathbf{z}).
\end{aligned}$$

Where going into (1) we use law of total probability and the chain rule. Going into (2), we use rule 2 to treat $\mathsf{do}(\mathbf{x})$ as $\mathbf{x}$. Going into (3) we use rule 3 to ignore $\mathsf{do}(\mathbf{x})$. ◄

► **Proposition 47.** *(**Front-door Criterion**)[Pea09] Given the right causal graphical model of fig (17), the graphical query for the effect of treatment* $\mathbf{x}$ *on outcome* $\mathbf{y}$ *can be reduced with:*

$$\mathfrak{p}\big(\mathbf{y} \,\big|\, \mathsf{do}(\mathbf{x})\big) = \sum_{\mathbf{m}} \mathfrak{p}\big(\mathbf{m} \,\big|\, \mathbf{x}\big) \sum_{\mathbf{x}'} \mathfrak{p}\big(\mathbf{y} \,\big|\, \mathbf{x}', \mathbf{m}\big)\mathfrak{p}(\mathbf{x}'). \tag{25}$$

Proof.

$$
\begin{aligned}
\mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x})\big) &\overset{(1)}{=} \sum_{\mathbf{m}} \mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x}), \mathbf{m}\big)\mathfrak{p}\big(\mathbf{m} \;\big|\; \mathsf{do}(\mathbf{x})\big) \\
&\overset{(2)}{=} \sum_{\mathbf{m}} \mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x}), \mathsf{do}(\mathbf{m})\big)\mathfrak{p}\big(\mathbf{m} \;\big|\; \mathbf{x}\big) \\
&\overset{(3)}{=} \sum_{\mathbf{m},\mathbf{z}} \mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x}), \mathsf{do}(\mathbf{m}), \mathbf{z}\big)\mathfrak{p}\big(\mathbf{z} \;\big|\; \mathsf{do}(\mathbf{x}), \mathsf{do}(\mathbf{m})\big)\mathfrak{p}\big(\mathbf{m} \;\big|\; \mathbf{x}\big) \\
&\overset{(4)}{=} \sum_{\mathbf{m},\mathbf{z}} \mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{m}), \mathbf{z}\big)\mathfrak{p}\big(\mathbf{z} \;\big|\; \mathsf{do}(\mathbf{m})\big)\mathfrak{p}\big(\mathbf{m} \;\big|\; \mathbf{x}\big) \\
&\overset{(5)}{=} \sum_{\mathbf{m}} \mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x})\big)\mathfrak{p}\big(\mathbf{m} \;\big|\; \mathbf{x}\big) \overset{(6)}{=} \sum_{\mathbf{m}} \mathfrak{p}\big(\mathbf{m} \;\big|\; \mathbf{x}\big) \sum_{\mathbf{x}'} \mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{m}), \mathbf{x}'\big)\mathfrak{p}\big(\mathbf{x}' \;\big|\; \mathsf{do}(\mathbf{m})\big) \\
&\overset{(7)}{=} \sum_{\mathbf{m}} \mathfrak{p}\big(\mathbf{m} \;\big|\; \mathbf{x}\big) \sum_{\mathbf{x}'} \mathfrak{p}\big(\mathbf{y} \;\big|\; \mathbf{m}, \mathbf{x}'\big)\mathfrak{p}\big(\mathbf{x}' \;\big|\; \mathsf{do}(\mathbf{m})\big) \overset{(8)}{=} \sum_{\mathbf{m}} \mathfrak{p}\big(\mathbf{m} \;\big|\; \mathbf{x}\big) \sum_{\mathbf{x}'} \mathfrak{p}\big(\mathbf{y} \;\big|\; \mathbf{m}, \mathbf{x}'\big)\mathfrak{p}(\mathbf{x}').
\end{aligned}
$$

Where going into (1) and (2), we applied rule 2. Going into (3) we applied rule 3. Going into (4) we introduce unobserved variable $\mathbf{z}$ and marginalize over it. Going into (5) and (6) we introduce $\mathbf{x}'$ and marginalize it out. Going into (7) and (8) we applied rules 2 and 3. ◂

If repeated application of do-Calculus rules allows one to convert an expression with $\mathsf{do}(\mathbf{x})$ operator to one that is $\mathsf{do}(\mathbf{x})$-free akin to eqn (24) and (25), then the causal effect of $\mathbf{x}$ on the outcome $\mathbf{y}$ is *identifiable*.

▸ **Example 48.** (**Answering the Forward Question**) Once the rules of do-Calculus are established, the forward question is trivially answered in the following setting:

1. The analyst is given data set and structural causal model $\mathcal{M}$ in some (ρro)Metheus environment.
2. The analyst queries $\mathcal{M}$ with i.e. $\mathsf{do}(\mathbf{x}_1)$ to identify the causal effect of $\mathbf{x}_1$ on $\mathbf{y}$.
3. The (ρro)Metheus Typed ρ-Calculus reduces the query using do-Calculus, and computes the likelihood of $\mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x}_1)\big)$. This likelihood is a score denoting the how much $\mathbf{x}_1$ influence $\mathbf{y}$.
4. Other analysts may then repeat this procedure for other factors $\{\mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots\}$.
5. The (ρro)Metheus oracle then ranks the likelihood of $\mathfrak{p}\big(\mathbf{y} \;\big|\; \mathsf{do}(\mathbf{x}_i)\big)$ to establish the cause of outcome $\mathbf{y}$ w.r.t. each $\mathbf{x}_i$.
6. If however $\mathfrak{p}\big(\mathbf{y} \;\big|\; \mathbf{x}_i\big)$ is not identifiable for some $\mathbf{x}_i$, then the (ρro)Metheus program terminates with either:
   - elicitation for more data;
   - or request for expert input in adjusting the assumptions of the graph. That is the edges or functions $\mathtt{f}(\mathsf{pa}(\mathbf{x}_i), \mathbf{u}_i)$ of the structural causal model.

   This aspect is important: request for more data is a *value driver* in the underlying commodity/data economy of the (ρro)Metheus ecosystem. While request for expert advice is a *growth driver* for new user acquisition.

Observe that in many circumstances, the complete dataset may not be revealed to all parties due to privacy and/or legal concerns. In other settings, the data is revealed under differential privacy masks which may reverse the ranking of causal factors. In these cases, the (ρro)Metheus oracle will act as the "neutral chamber" that adjudicates the outcome based on access to all the information.

6.2 Causal Reinforcement Learning

While the forward question may be trivially answered given sufficient data and the identifiable criteria is met, the reverse direction is categorically more complex to answer. In the backward question setting, the (ρro)Metheus oracle is given raw data, and must search for the most likely causal graph *tabula rasa*. This falls in the domain of machine learning, specifically reinforcement learning (RL).

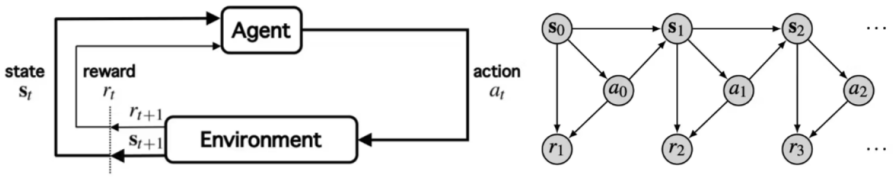


Figure 18 Reinforcement learning defines how an autonomous Agent learns to what to do in order to maximize a collection of sequential numerical rewards. This is an interactive setting whereby an Agent acts on the Environment at each time step t with action a_t. This action a_t changes its underlying state from s_t to s_{t+1}. The Agent observes this delta and sometimes receives a reward r_t at each t, other times the Agent may only receive a final reward when the game terminates at time T [Sut15].

6.2.1 Reinforcement Learning and MDPs

RL is different from supervised learning (i.e.: classification and regression) in that this setting does not provide labeled data. Instead, it is a form of goal-directed learning from repeated interactions with some Environment [C22; Bar25]. Since the RL Agent plays to win, the strategy of how to play the game will be of vital importance, not only for maximizing rewards, but also for computational efficiency. Given some Agent in an Environment, the game proceeds as follows:

1. First the Agent takes an action a_t.
2. Then the Environment responds by:
   - giving the Agent a reward r_t,
   - and increments the state from s_t to s_{t+1}.

So that as the learner must explore the Environment, it constructs an observed state space of form:

$$\{\ldots, (s_t, a_t, r_t, s_{t+1}), \ldots\}. \tag{26}$$

The output of reinforcement learning algorithms are two-fold:

1. A strategy for how to play "the game" to maximize the reward.
2. Sometimes, the algorithm will also output an underlying world model based on the state space it has explored in (26).

The (ρro)Metheus oracle constructs this underlying world model as a structural causal model used to assign liability to various parties.

▸ **Definition 49.** ***Markov Decision Process (MDP)****[Bar25] is a tuple* $(\mathbf{S}, \mathbf{A}, \mathfrak{p}(\cdot), \mathcal{R}, \gamma)$ *where:*

- $\mathbf{S}$*: is a state space which contains all possible states of the* Environment.
- $\mathbf{A}$*: action or policy space containing all possible actions the* Agent *could take.*
- $\mathfrak{p}(s_{t+1} \mid s_t, a_t)$*: is a stochastic transition kernel which gives the likelihood of reaching* s_{t+1} *from* s_t *if the* Agent *takes action* a_t*. Observe the next state* s_{t+1} *is independent of all past states* $s_o, \ldots, s_{t-1}$*, conditioned on current state* s_t *and action* a_t*. This is the "Markov" aspect of MDP.*

- $\mathcal{R} : \mathbf{S} \times \mathbf{A} \to \mathbb{R}$*: a reward function that given state and action, outputs real valued reward.*
- $\gamma \in [0, 1]$*: a discount factor that down-weighs future rewards.*

The MDP sequence is depicted on the right in fig (18). The goal of RL is to derive a policy:

$$\pi(\cdot) : \mathbf{S} \to \mathbf{A}$$

that maximizes the cumulative reward across time-steps:

$$\mathbf{E}\left[\sum_{t}^{T} \gamma^t \mathcal{R}(s_t, a_t) \,\middle|\, s_0 = s_k\right] \qquad \text{where } a_t \sim \pi(s_t). \tag{27}$$

Here the expectation is taken over all possible sequences of gameplay the Agent *may encounter. In the event where the* Agent *is playing some game against an opponent, then implied in the possible game sequences are the set of possible moves the opponent could make.*

▶ Remark 50. (**Online vs offline learning**) are two different settings in RL w.r.t when and how the Environment is shown to the Agent.

- In *online learning*, the Agent is interacting with the Environment in real time, and collect reward whilst exploring its surroundings.
- In *off-policy learning*, the Agent learns an optimal policy from offline data generated by a different behavior policy or agent. This is also known as *imitation learning* or *learning from experts*. This is the setting of the (ρro)Metheus oracle.

▶ Remark 51. (**Managing State Space in RL**) In a simple game such as tic-tac-toe with 9 possible positions on the board, the state space is manageable. In a more complex games such as Go with hundreds squares, the raw state space of the game becomes intractable. In this case machine learning techniques are needed to compress the representation of the state space, this was the part of the breakthrough of Alpha-Go [Sil16].

If the underlying state is completely observed, then this gameplay is a *Markov Decision Process* (MDP), which formalizes the RL setting. See definition 49. If however the underlying state is only partially accessible, then the setting is a *Partially Observed Markov Decision Process* (POMDP). Here the Agent is not playing against a particular state, but a distribution over all possible states. Thus computing the best move may be intractable unless the state space itself is parsimoniously represented.

Once again, neural networks must be used to compress the state space. Another option that further reduces the size of the state space is causal reinforcement learning. It uses structural invariants in the Environment to find a better representation of the world that is smaller and more robust.

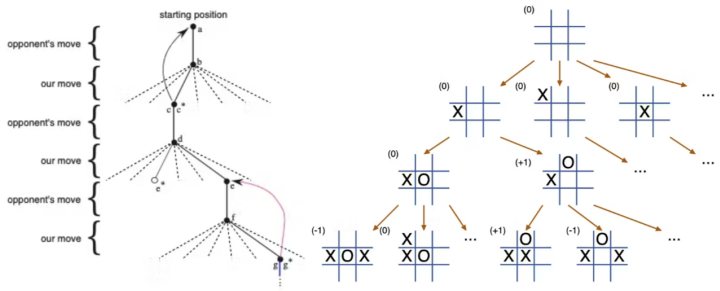


Figure 19 (*Right*): RL game tree for tic-tac-toe, where the state s_t is the current positions on the game board, the actions a_t are where to place the *X* or *O*'s, the reward comes at time T when the game ends and the winner is rewarded. Board games are MDPs because the board is plainly visible. In contrast card games (i.e. poker) are POMDPs because the player cannot observe the opponents hands, but instead keeps a distribution over all possible hands the opposite may hold. (*Left*): a game tree traced by an MDP. The solid lines represent the moves taken during a game; the dashed lines are moves that the player considered but did not make [J P17].

Reinforcement learning is a general setting that applies to many domains with various underlying states. The underlying state could be a game board in Alpha-Go. It may be a mechanical system, i.e. a robot that hops on one leg. In the case of (ρro)Metheus, the underlying state is a world model, whereby the observed data are generated by a set of causal relations waiting to be discovered.

6.2.2 Causal Reinforcement Learning with Known Causal Graph

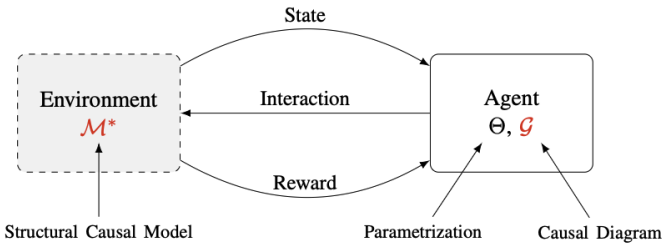


Figure 20 The Agent-Environment interaction from Causal Reinforcement Learning [Bar25].

Causal reinforcement learning (CRL) was formalized to address the fact that although statistical models "should" perform out of sample, they often fail to generalize. One avenue towards more robust models is to build ones that express structural invariances in the world. This entails learning a causal model so that the AI can reason about questions of counterfactual nature, i.e., what would have happened had reality been different, even when no data about this imagined reality is available, and the Environment is not fully observable. In CRL, this causal underlying reality is explicitly expressed. There is an intrinsic duality between how the Agent and Environment view causality as seen in fig (20):

- From the Environment's perspective, causal mechanisms and the probability distribution over the exogenous variables are described by the structural causal model $\mathcal{M}^*$ on the left side. This is a platonic ideal that is assumed to exist.
- From the Agent's perspective, a parsimonious representation of the Environment's invariances is encoded in the form of a causal graph $\mathcal{G}$ on the right side [Bar25].

The (ρro)Metheus oracle builds this causal graph $\mathcal{G}$ using CRL by undertaking a causal discovery process, expressed by the *causal decision model*.

▶ **Definition 52.** ***Causal Decision Model (CDM)*** *[Bar25]*

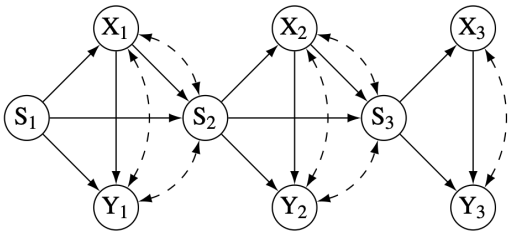


Figure 21 The MDP of fig (18) re-expressed as a CDM, where the Agent acts on the Environment with actions $a_t := x_t$, thus changing the Environment state from s_t to s_{t+1}, yielding an observation of $\mathbf{y}_t$ [Bar25].

CDM is a tuple

$$(\mathcal{M}^*, \mathbf{\Pi}, \mathcal{R}), \qquad \text{where } \mathcal{M}^* = (\mathcal{G}, \mathcal{F}, \mathcal{U}), \tag{28}$$

Over decision horizon or number of rounds $1, \ldots, t, \ldots, T$. *And over variable spaces:*

$$\begin{aligned} &(\textit{state space}) : && \mathbf{S} = \{s_1, \ldots, s_T\}, \\ &(\textit{action/intervention space}) : && \mathbf{x} = \{x_1, \ldots, x_T\}, \\ &(\textit{treatment effect space}) : && \mathbf{y} = \{y_1, \ldots, y_T\}. \end{aligned} \tag{29}$$

So that:

- $\mathcal{M}^*$ *is the true underlying structural causal model.*
- $\mathbf{\Pi}$ *is a policy space denoted:*

$$\big\{(x_1, s_1), \ldots, (x_T, s_T)\big\}, \quad \text{or } \pi_t(x_t|s_t)\ \forall t \tag{30}$$

Such that each action x_t *is non-descendent of* $x_{t+1}, \ldots, x_T$. *And each state* s_t *is non-descendent of* $x_t, \ldots, x_T$. *See example (54) below.*

- $\mathcal{R}$ *is a reward function over effect signals* $\mathbf{y} = \{y_1, \ldots\}$ *defined with:*

$$\mathcal{R} : \mathbf{y} \to \mathbb{R}, \tag{31}$$

mapping treatment effect to real valued reward. See fig (21).

In general, we focus on MDPs and POMDPs, however it is helpful to note that there is an escalation ladder in complexity of planning the Agent is asked to undertake. This ladder is defined w.r.t the planning horizon.

▶ Remark 53. (**CRL Learning Regime** $\mathcal{L}$). The decision Horizon T and ability to perceive confounding variables define the kind of game, or learning regime $\mathcal{L}$ the Agent is playing.

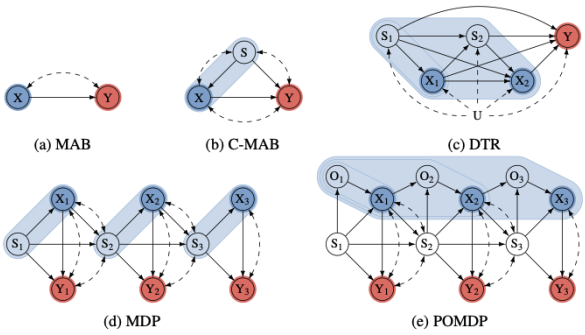


Figure 22 Causal reinforcement learning may be represented by various decision making processes [Bar25].

- (a) In the simplest *multi-arm bandit* setting, the Agent pulls a lever with action $\mathbf{x}_t$ and receives a reward $\mathbf{y}$. The policy space is $\mathbf{\Pi} = \{(\mathbf{x}, \emptyset)\}$, so that Agent simply samples some treatment value x from the the distribution by $x \sim \pi(\mathbf{x})$. The reward is just $\mathcal{R}(\mathbf{y})$.
- (b) In the *contextual bandit* setting with confounding variable $\mathbf{s}$, the policy space is $\mathbf{\Pi} = \{(x, \{s\})\}$, and the Agent sample policy by $x \sim \pi(\mathbf{x}|s)$. The reward is once again $\mathcal{R}(\mathbf{y})$.
- (c) In the *dynamic treatment regime* setting, the policy space is:

  $$\mathbf{\Pi} = \left\{\pi_1(\cdot), \ldots, \pi_T(\cdot)\right\} = \left\{(x_1, \{s_t, \ldots, s_t, x_1, \ldots, x_{t-1}\})\right\}_{t=1}^{T}.$$

  So that the current treatment $\mathbf{x}$ depends on past treatments, as well as all the confounding variables or system states s_t's up to this time step t. Each treatment x_t is now sampled out of the policy:

  $$x_t \sim \pi_t(\mathbf{x} \mid s_1, \ldots, s_t, x_1, \ldots, x_{t-1}).$$

  The reward is: $\sum_t^T \mathcal{R}(y_t)$
- (d) *Markov Decision Process* is a generalization of dynamic treatment regime, whereby the Environment has infinite states (confounding variables), actions, and observations. And we have:

  $$\mathbf{\Pi} = \left\{(x_t,\ \{s_1, \ldots, s_t, x_1, \ldots, x_{t-1}\})\right\}_{t=1}^{\infty}, \quad \text{and } x_t \sim \pi_t(\mathbf{x} \mid s_1, \ldots, s_t, x_1, \ldots, x_{t-1}).$$

  The reward is now discounted by $\sum_t^\infty \gamma^t \mathcal{R}(y_t)$.
- (e) In *partially observed Markov decision process*, the Agent cannot access the underlying state directly, but instead perceives it with observation variable $\mathbf{o}$. So that we have:

  $$\mathbf{\Pi} = \left\{(x_t,\ \{o_1, \ldots, o_t, x_1, \ldots, x_{t-1}\})\right\}_{t=1}^{\infty}, \quad \text{and } x_t \sim \pi_t(\mathbf{x} \mid o_1, \ldots, o_t, x_1, \ldots, x_{t-1}).$$

  The reward is $\sum_t^\infty \gamma^t \mathcal{R}(y_t)$.

▶ **Example 54.** (**DTR Policy Space**)[Bar25] A treatment regime or policy space in healthcare for alcoholics may be expressed with:

$$\mathcal{M}^* = \Big(\mathcal{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_5), \mathcal{V} = (s_1, s_2, x_1, x_2, \mathbf{y}), \mathcal{F}, \mathfrak{P}(\mathcal{U})\Big).$$

Where:

- $\mathcal{U}$ are the exogenous variables outside of the model.
- $\mathcal{V}$ are the endogenous variables inside the model, with interpretations:

- s_1: not sober
- s_2: sober
- x_1: behavioral therapy
- x_2: drug treatment
- $\mathbf{y}$: days sober.

- $\mathcal{F}$ is encoded with:

$$\begin{aligned}
&\mathsf{s}_1 \rightarrow \mathbb{1}(\mathbf{u}_3 > 0),\\
&\mathsf{x}_1 \rightarrow \mathbb{1}(3\mathsf{s}_1 + \alpha_1\mathbf{u}_1 + \mathbf{u}_2 > 0),\\
&\mathsf{s}_2 \rightarrow \mathbb{1}(0.1 + 0.1\mathsf{s}_1 + 0.1\mathsf{x}_1 + \mathbf{u}_4 > 0),\\
&\mathsf{x}_2 \rightarrow \mathbb{1}(3\mathsf{s}_2 + \alpha_2\mathbf{u}_1 + \mathbf{u}_3 > 0),\\
&\mathbf{y} \rightarrow \mathbb{1}(3\mathbf{u}_1 - 3\mathsf{s}_1 - 3\mathsf{x}_1 - 3\mathsf{s}_1\mathsf{x}_1 + 3\mathsf{x}_2 - 3\mathsf{s}_2\mathsf{x}_2 + 3\mathsf{x}_1\mathsf{x}_2 > 0).
\end{aligned}$$

- $\mathfrak{P}(\mathcal{U})$ is a distribution over the exogenous variables defined with:

$$\mathfrak{p}(\mathbf{u}_t < u) = \frac{1}{1 + e^{-u}}.$$

Now we may construct a restricted policy space or treatment regime as follows:

$$\mathbf{\Pi}_1 = \Big(\pi_1(\mathsf{x}_1|\mathsf{s}_1), \pi_2(\mathsf{x}_2|\mathsf{s}_2)\Big).$$

One could also adopt a dynamic or responsive treatment regime with:

$$\mathbf{\Pi}_2 = \Big(\pi_1(\mathsf{x}_1|\mathsf{s}_1), \pi_2(\mathsf{x}_2|\mathsf{s}_1, \mathsf{x}_1, \mathsf{s}_2)\Big).$$

And observe that $\mathbf{\Pi}_1$ does not account for whether treatment x_1 := behavioral therapy was prescribed first, where as $\mathbf{\Pi}_2$ does account for past treatment and responses.

Example 54 can be generalized into the *causal reinforcement learning task*.

▶ **Definition 55.** *(**Causal Reinforcement Learning Task**) [Bar25] For an SCM $\mathcal{M}^* = (\mathcal{G}, \mathcal{F}, \mathcal{U})$, a CRL task $\mathcal{T}$ in* Environment *$\mathcal{M}^*$ is a 4-tuple $(\mathcal{L}, \mathcal{A}, \mathbf{\Pi}, \mathcal{R})$ where:*

- *$\mathcal{L}$ is the* Agent*'s learning regime as seen in remark (53).*
- *$\mathcal{A}$ is a set of structural assumptions about SCM $\mathcal{M}^*$.*
- *$\mathbf{\Pi}$ is the policy space over actions $\mathbf{x}$.*
- *$\mathcal{R}$ is the reward function over effect signals $\mathbf{y}$.*

Similar to RL, the Agent*'s goal is to find the optimal policy w.r.t the underlying causal structure $\mathcal{M}^*$:*

$$\pi^* = \arg\max_{\pi^* \in \mathbf{\Pi}} \mathbf{E}_\pi \Big[\mathsf{Reward}(\mathbf{y}) \;\Big|\; \mathcal{M}^*, \mathcal{A}, \mathcal{L} \Big]. \tag{32}$$

And observe π^* may be different depending on whether Agent *is in (C)MAB, DTR, or (PO)MDP setting, as these are given in the learning regime $\mathcal{L}$.*

Every causal RL task proceeds with the following computational skeleton.

Algorithm 1 Causal Reinforcement Learning Agent [Bar25]

Require: CRL task $\mathcal{T} = (\mathcal{L}, \mathcal{A}, \mathbf{\Pi}, \mathcal{R})$.
Ensure: a policy estimate $\pi^* \in \mathbf{\Pi}$ optimizing a CDM $(\mathcal{M}^*, \mathbf{\Pi}, \mathcal{R})$.
1: $\mathbb{D} \leftarrow \{\}$
2: **for** $t = 1 \ldots T$ **do**
3: Interact with SCM $\mathcal{M}^*$ following regime $\mathcal{L}$ and receive treatment response:

$$y_t \sim \mathfrak{p}\left(\mathbf{y} \mid \mathcal{M}^*\right).$$

4: $\mathbb{D} \leftarrow \mathbb{D} \cup \{y_t\}$
5: **end for**
6: **return** empirical estimate $\pi \in \mathbf{\Pi}$ for π^* from data $\mathbb{D}$ and assumptions $\mathcal{A}$.

In line 3 above, the Agent interacts with $\mathcal{M}^*$ under the (PO)MDP regime as follows:

1. Perceive state s_t or observation o_t.
2. Select an action x_t with $x_t \sim \pi_t(x_t|s_t)$ and perform $\text{do}(\mathbf{x} = x_t)$.
3. Receive treatment response y_t based on $\text{do}(\mathbf{x})$ action.

6.2.3 Causal Reinforcement Learning with Unknown Causal Graph

The algorithm above presumes the Agent is equipped with set of assumptions $\mathcal{A}$ about the underlying causal structure of the world model $\mathcal{M}^*$. This is a fair assumption in many settings. In this section we consider the problem whereby the Agent must learn a causal model *tabula rasa*, with $\mathcal{A} = \emptyset$. This is the backward question of causality as defined in section 5.2.1.

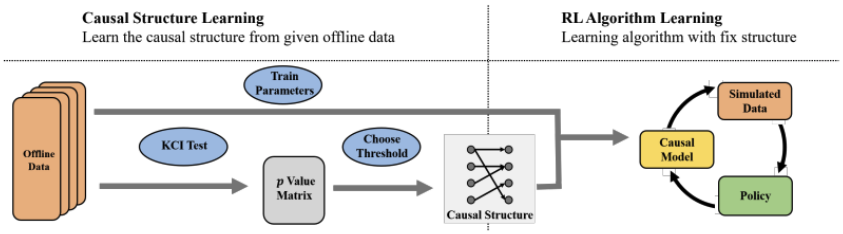


Figure 23 The backward causation problem can be solved by CRL by breaking the process into two steps: 1) a causal graph $\mathcal{G}$ discovery process, and 2) and CRL strategy π optimization process [ZM22].

The computational skeleton for CRL with unknown causal graph proceeds as below.

Algorithm 2 Offline CRL with Unknown Graph Structure (Adapted from [ZM22])

Require: Dataset $\mathbb{D}$
Require: RL learning parameters $(\mathcal{L}, \mathbf{\Pi}, \mathcal{R})$
1: **Stage 1** Causal structure learning
2: $\mathcal{G} \leftarrow \text{learnGraph}(\mathcal{L}, \mathbf{\Pi}, \mathcal{R}, \mathbb{D})$
3: **Stage 2** Offline reinforcement learning
4: Define $\mathcal{A}$ from $\mathcal{G}$
5: $\pi \leftarrow \text{Algorithm1}(\mathcal{L}, \mathcal{A}, \mathbf{\Pi}, \mathcal{R})$
6: **return** policy and causal model pair $(\pi, \mathcal{G})$

Learning the causal graph $\mathcal{G}$ *tabla rasa* is nontrivial, as the greedy approach whereby each vertex $\mathbf{x}$ is independently explored as a causal factor is exponentially suboptimal in the number of interventions required [K19]. Here we review example algorithms that explores the causal factors given certain (mild) assumptions on the structure of true causal graph $\mathcal{G}^*$. In particular, the authors below consider two learning regimes $\mathcal{L}$, multi-arm bandit [K19; Lu21a], and Markov decision process [Lu21b].

Sample Efficient Active Learning of Causal Trees

In [K19], the authors proceeds as follows:

1. First assume the underlying causal graph $\mathcal{G}^*$ is a directed acyclic tree as shown in fig (24).

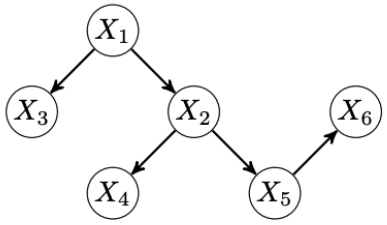


Figure 24 The methods in [K19] assume the underlying causal graph $\mathcal{G}^*$ is a directed acyclic tree. Thus causal factor identification over $\mathcal{G}$ is tree search that can be completed in $\log_2 |\mathcal{G}|$ number of interventions, where $|\mathcal{G}|$ is the number of vertices in the graph.

2. Next, given a set of treatments and effect factors $\mathbf{x}_1, .., \mathbf{x}_n$, and sufficient data to determine the joint distribution $\mathfrak{p}(\mathbf{x}_1, .., \mathbf{x}_n)$. The algorithm initiates a tree $\mathcal{G}_o$ over $\mathbf{x}_1, .., \mathbf{x}_n$ with undirected edges.
3. Then they take a Bayesian approach and assume a prior distribution over the set of all possible graphs given this undirected tree.
4. The goal is to identify the root cause, or vertex $\mathbf{x}_o \in \mathcal{G}$. Define:
   - $\mathsf{N}_{\mathcal{G}}(\mathbf{x})$ as the *neighbor* of $\mathbf{x}$, so that in fig (24) we have $\mathsf{N}_{\mathcal{G}}(\mathbf{x}_2) = \{\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_5\}$.
   - $\mathsf{B}_{\mathcal{G}}[\mathbf{x}_i : \mathbf{x}_j]$ as a *branch* of the graph that can be reached from $\mathbf{x}_j$ when the edge from $\mathbf{x}_i$ to $\mathbf{x}_j$ has been cut. For example in fig (24) the branches of $\mathbf{x}_2$ are:
     - $\mathsf{B}_{\mathcal{G}}[\mathbf{x}_2 : \mathbf{x}_1] = \{\mathbf{x}_1, \mathbf{x}_3\}$, so that the subgraph on $(\mathbf{x}_1, \mathbf{x}_3)$ is isolated from the main graph.
     - $\mathsf{B}_{\mathcal{G}}[\mathbf{x}_2 : \mathbf{x}_4] = \{\mathbf{x}_4\}$, so the vertex $\mathbf{x}_4$ is an island.
     - $\mathsf{B}_{\mathcal{G}}[\mathbf{x}_2, \mathbf{x}_5] = \{\mathbf{x}_5, \mathbf{x}_6\}$, so that the subgraph over $(\mathbf{x}_5, \mathbf{x}_6)$ is an island.
5. For simplicity sake assume $\mathbf{x}_i \in \{0, 1\}$, so that an intervention is written $\mathsf{do}(\mathbf{x}_i = 1)$. The authors define the post-intervention distribution update with:

$$\forall \mathbf{x}_k \in \mathsf{B}_{\mathcal{G}}[\mathbf{x}_i : \mathbf{x}_j],\ \mathfrak{p}\left(\mathbf{x}_o \text{ is } \mathbf{x}_k \,\middle|\, \mathsf{do}(\mathbf{x}_i = 1), \mathbf{x}_1 = x, \ldots\right) \sim \begin{cases} \mathfrak{p}(\mathbf{x}_o \text{ is } \mathbf{x}_k) \frac{\mathfrak{p}(\mathbf{x}_j = x)}{\mathfrak{p}\left(\mathbf{x}_j = x \,\middle|\, \mathbf{x}_i = 1\right)} & \mathbf{x}_j \in \mathsf{N}_{\mathcal{G}}(\mathbf{x}_i) \\ \mathfrak{p}(\mathbf{x}_o \text{ is } \mathbf{x}_k) & \mathbf{x}_j = \mathbf{x}_i \end{cases} \tag{33}$$

   Here the quantity $\mathfrak{p}\left(\mathbf{x}_o \text{ is } \mathbf{x}_k \,\middle|\, \mathsf{do}(\mathbf{x}_i = 1), \mathbf{x}_1 = x, \ldots\right)$ is read: the likelihood that the root cause is vertex $\mathbf{x}_k$, given intervention $\mathsf{do}(\mathbf{x}_i = 1)$ on factor $\mathbf{x}_i$, and let the rest of the factors $\mathbf{x}_1, \ldots$ in $\mathsf{B}_{\mathcal{G}}[\mathbf{x}_i : \mathbf{x}_j]$ take on their observational values x.
6. Now the search can proceeds as follows:

Algorithm 3 Central Node Algorithm, adapted from [K19]

Require: Observational tree $\mathcal{G}_o$ with $|\mathcal{G}_o| = \mathsf{n}$, confidence parameter δ

$\mathsf{t} \leftarrow 0$
$\mathcal{G}(0) = \mathcal{G}_o$
$\mathfrak{p}(\mathbf{x}_\mathsf{o} \text{ is } \mathbf{x}_\mathsf{i})^o = \frac{1}{\mathsf{n}} \; \forall \mathbf{x}_\mathsf{i} = 1, \ldots, \mathsf{n}$
while $\max_\mathsf{i} \mathfrak{p}(\mathbf{x}_\mathsf{o} \text{ is } \mathbf{x}_\mathsf{i}) \leq 1 - \delta$ **do**
    $\mathsf{t} \leftarrow \mathsf{t} + 1$
    identify central vertex $\mathbf{x}_\mathsf{j}$ of $\mathcal{G}(\mathsf{t}-1)$ w.r.t. $\mathfrak{p}(\mathbf{x}_\mathsf{o} \text{ is } \mathbf{x}_\mathsf{i})^{\mathsf{t}-1}$
    Intervene on vertex $\mathbf{x}_\mathsf{j}$ and observe direction of root vertex in $\mathbf{x}_\mathsf{t} \in \{\mathbf{x}_\mathsf{j}\} \cup \mathsf{N}_{\mathcal{G}(\mathsf{t}-1)}(\mathbf{x}_\mathsf{j})$
    Update posterior distribution according to eqn (33)
    $\mathcal{G}(\mathsf{t}) \leftarrow \mathsf{B}_{\mathcal{G}(\mathsf{t}-1)}[\mathbf{x}_\mathsf{j} : \mathbf{x}_\mathsf{t}]$
end while
return vertex remaining in $\mathcal{G}(\mathsf{t})$ as the root cause $\mathbf{x}_\mathsf{o}$ of $\mathcal{G}_o$

[K19] showed that the running time of the algorithm above is $\mathsf{T} \leq \log_2(|\mathcal{G}|)$, and its expected running time is constant factor away from the optimal running time by $\mathbf{E}[\mathsf{T}] \leq 2\mathbf{E}[\mathsf{T}_{opt}]$.

Causal Bandit with Unknown Graph Structure

In [Lu21a],

Causal Markov Decision Processes: Learning Good Interventions Efficiently

In [Lu21b],

bibliography