CONSTRUCTIVE CAUSAL ABSTRACTION AND A CASE STUDY IN MECHANISTIC INTERPRETABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing frameworks in causal abstraction first propose a high-level causal model, and then verify whether its interventional effects agree with those of a given micro-level model - the two models are declared causally consistent when the discrepancy is zero. However, proposing a consistent macro causal model is difficult, because macro causal relations are almost never well-defined, since distinct micro-level interventions associated with the same apparent macro intervention can produce different effects. Our central insight is that causally consistent macro-level models can be obtained *constructively*: ill-defined macro interventions can be interpreted as missing variables on the macro level, and can be accounted for by introducing a complementary variable q that retains omitted causally relevant state. In the linear case, q has a unique form and we provide sufficient conditions for valid implementations to exist; in the nonlinear case, we provide an ODE-based algorithm for constructing implementations. We call this framework *Constructive Causal Abstraction* (CCA), and use CCA to formulate the difference between machine-learning model auditing and editing. With the Bias-in-Bios biographies dataset, we demonstrate that CCA allows practitioners to articulate assumptions when detecting the model's gender bias. Finally, CCA resolves a well-known paradox known as 'interpretability illusion' which points to a mismatch between the subspace for editing and the one where the model implements a signal-to-output pathway.

1 INTRODUCTION

Scientific reasoning routinely operates at multiple levels of resolution. A macroeconomist models sector-level aggregates, not individual firms; a neuroscientist reasons about brain regions, not single synapses; an AI safety researcher targets high-level capabilities of a neural network, not individual weights. In each case, practitioners intervene on a coarse-grained *macro* model and expect the outcome to be predictable from that model alone, without tracking every microscopic degree of freedom.

Causal abstraction (Rubenstein et al., 2017; Beckers & Halpern, 2019; Rischel, 2020; Kekic et al., 2024) formalises this expectation. A high-level model $\bar{M}$ is an exact transformation of a low-level model M if there exists a mapping τ between their variable spaces such that a commutative diagram holds: intervening at the micro level and then aggregating yields the same outcome as aggregating first and then intervening at the macro level.

However, existing frameworks are *verificational*: they assume the macro model is given a priori, which implicitly entails that effect of interventions are well-defined, and they check for consistency with the micro model. If the proposed macro model is not consistent, the existing frameworks do not provide guidance on the constructive question:

> *Given a micro-level causal model and a set of macro-variables aggregated from the micro-variables, how do we come up with a macro model that is an exact transformation of the micro model?*

The hardness of the problem arises because the aggregation map $\boldsymbol{p} : \mathcal{X} \to \bar{\mathcal{X}}$ is many-to-one: for a given macro-intervention $\bar{\boldsymbol{i}}$, there are generically infinitely many micro-level realisations. Unless further structure is imposed on the macro model, different realisations produce different macro-level causal effects, making the macro-intervention *ill-defined*. In other words, there is an *implementation gap*.

Macro level noise. In a structural causal model (SCM), exogenous noise variables encode variation not explained by the causal mechanisms. When micro variables are aggregated into macro variables, the micro exogenous noise and the ambiguity of implementations must be coherently aggregated; otherwise the macro model has no well-defined noise, and no well-defined counterfactuals. We show that the missing ingredient is a *complementary variable* $\boldsymbol{q}$: a function of the micro state and noise that encodes information discarded by $\boldsymbol{p}$. Specifying $\boldsymbol{q}$ pins down which micro-realisations are admissible, collapsing the set of valid implementations to those that all produce the same macro-level causal effect.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Constructive Causal Abstraction. Specifying the macro level noise $\boldsymbol{q}$ along with the aggregation map $\boldsymbol{p}$ can be viewed as projecting micro-level variables onto a subset of coordinates under a coordinate transformation of the underlying micro variable system; exact transformation arises by construction, since all we have done is a coordinate transformation. Therefore, we term our framework *constructive causal abstraction* (CCA).

Our contributions are three-fold.

1. **Constructive foundation.** We achieve causal exact transformation by construction, through formulating causal abstraction as a change of coordinates followed by a projection from the micro model onto macro variables $\boldsymbol{p}$ and macro noise variable $\boldsymbol{q}$. This perspective provides the necessary foundation for reliably controlling micro-level systems through macro-level interventions (Section 4).
2. **Well-defined macro interventions and existence of implementations.** We point out and resolve ill-defined macro models due to ambiguous interventions, formally prove well-definedness of causal effects under our framework (Theorem 2, Theorem 4). For linear causal models with linear aggregation maps, $\boldsymbol{q}$ admits a unique-form expression and we establish mild sufficient conditions for existence of micro-level implementations of any macro shift intervention (Section 5, Theorem 3). For nonlinear causal models, we derive an ODE-based algorithm that constructs valid implementations (Section 6).
3. **Case study in mechanistic interpretability.** Applying our theoretical framework to model auditing and editing, we showcase that it helps capture distributional assumptions in the well-studied sensitive-attribute auditing scenario (Section 7.1) and demonstrate that it resolves a well-known paradox in mechanistic interpretability through an example in Indirect Object Identification (Section 7.2).

2 Related Work

Causal abstraction. Rubenstein et al. (2017) is first to define exact transformations between SCMs. Beckers & Halpern (2019) strengthen this account, while Rischel & Weichwald (2021) develop a category-theoretic treatment of abstraction error. These accounts are verificational: given a micro model, macro model, and cross-level maps, they assess whether the abstraction is consistent. We take a constructive perspective: starting from a micro model and macro variables, we derive the complementary structure needed to make macro-interventions well-defined. Kekic et al. (2024); Zhu et al. (2024b) also construct abstractions, jointly learning macro variables, intervention maps, and mechanisms from interventional data; their identifiability results require conditions including linearity, independent high-level noise, and sufficient intervention coverage. Our work addresses a logically prior question: when the macro variables are specified, what makes an intervention on them well-defined? We expose the ambiguity among micro-level implementations and resolve it through the macro noise $\boldsymbol{q}$, without additional assumptions, providing a general foundation upon which further learning or identifiability constraints may be imposed. Finally, Zhu et al. (2024a) identify ambiguity in macro-level interventions and illustrate resulting structural paradoxes, while our work constructively resolves that ambiguity.

Causal coarsening and emergence. Chalupka et al. (2016) study causal feature learning from a multi-resolution perspective although they work solely in the discrete case, and Hoel et al. (2013) quantify the "causal emergence" that arises when macro models have more predictive power than micro models. Shalizi & Moore (2003) analyses when macroscopic variables can be treated as autonomous dynamical systems. These works ask *when* macro variables have autonomous causal dynamics; we address the complementary question of *how* to construct them.

Mechanistic Interpretability. Neural networks are studied through interventions on macro variables. Sensitive-attribute audits ask whether demographic information is encoded or causally used (De-Arteaga et al., 2019; Vig et al., 2020; Ravfogel et al., 2020); interchange interventions test causal abstractions (Geiger et al., 2021); activation steering modifies behaviour through learned directions (Turner et al., 2023; Li et al., 2023; Zou et al., 2023); and locate-then-edit methods modify factual associations (Meng et al., 2022; 2023). Their conclusions can depend on the intervention's implementation. Hase et al. (2023) find successful fact edits need not occur where causal tracing localises the fact, while Makelov et al. (2024) show that patches can act through dormant pathways. Wu et al. (2024) dispute interpreting these cases as "illusions", citing training and evaluation choices. Similar sensitivities affect activation patching and circuit ablation (Zhang & Nanda, 2024; Miller et al., 2024), while Tan et al. (2024) find that steering effects vary across inputs and distributions. Meanwhile, CCA formalises one source of sensitivity as *implementation ambiguity*: different micro-level realisations of the same purported macro-intervention need not agree. Requiring implementations to preserve the complementary variable $\boldsymbol{q}$ makes the macro-level effect well-defined and implementation-invariant. Articulating $\boldsymbol{q}$ lets practitioners specify preservation constraints defining a well-defined macro model.

3 PRELIMINARIES

We build on Peters et al. (2017) to define the class of **structural causal models** we consider in this paper.

Definition 1 (SCM without measure)**.** *A structural causal model (SCM) is a tuple* $(\mathcal{K}, \mathcal{J}, \mathcal{X}, \mathcal{N}, \boldsymbol{f})$ *where* $\mathcal{K}$ *is a finite index set of endogenous variables;* $\mathcal{J}$ *is a finite index set of exogenous variables;* $\mathcal{X} = \prod_{k\in\mathcal{K}} \mathcal{X}_k$ *and* $\mathcal{N} = \prod_{j\in\mathcal{J}} \mathcal{N}_j$ *are measurable product spaces; and* $\boldsymbol{f} : \mathcal{X} \times \mathcal{N} \to \mathcal{X}$ *is a measurable function called the causal mechanism.*

We do not assume a probability measure over $\mathcal{N}$ and we do not assume mutual independence of the exogenous variables (the model is semi-Markovian in general). This allows our results to apply to a wide class of causal models.

We write $\boldsymbol{s} : \mathcal{N} \to \mathcal{X}$ for the solution function satisfying $\boldsymbol{s}(\boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}), \boldsymbol{n})$ for all $\boldsymbol{n} \in \mathcal{N}$.

Assumption 1 (Unique solutions)**.** *For every* $\boldsymbol{n} \in \mathcal{N}$*, the equation* $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n})$ *has a unique solution. The same holds for all intervened structural equations considered in this paper.*

This assumption is a pointwise, measure-free counterpart of unique solvability (Bongers et al., 2021, Definition 3.3 and Theorem 3.6). Since unique solvability need not be preserved under intervention (Bongers et al., 2021, Section 3.4), we impose it for every intervention considered here. Note that we only require that a solution exists, which is always true for directed acyclic graphs (DAGs), but we do not assume that the causal models considered must form DAGs.

Definition 2 (Intervention)**.** *An intervention on variables* $X_{[k]}$ *is a map* $i_{[k]} : \mathcal{X}_{[k]} \to \mathcal{X}_{[k]}$*. It transforms the structural equation from* $f_{[k]}(X_{\mathrm{pa}([k])}, N_{[k]})$ *to* $i_{[k]} \circ f_{[k]}(X_{\mathrm{pa}([k])}, N_{[k]})$*. The identity map* id *represents no intervention.*

This covers shift interventions ($i_k(x) = x + c$) and do-interventions ($i_k(x) = c$) as special cases. To the best of our knowledge, we are the first to directly formulate interventions as maps on variable spaces that act on structural mechanisms by post-composition. This is a special case of the soft interventions defined by Geiger et al. (2025, Definition 10) and Massidda et al. (2023, Definition 6), with the replacement mechanism being $\tilde{f}_{[k]} = i_{[k]} \circ f_{[k]}$.

The **aggregation map** is a function $\boldsymbol{p} : \mathcal{X} \to \bar{\mathcal{X}}$, where $\bar{\mathcal{X}}$ the space of macro variables. We write $\upsilon : \mathcal{K} \to \bar{\mathcal{K}}$ for the index assignment; $\upsilon(k) = \bar{k}$ indicates that micro-variable k contributes to macro-variable $\bar{k}$.

Assumption 2 (Block-decomposable aggregation)**.** *The aggregation map* $\boldsymbol{p} : \mathcal{X} \to \bar{\mathcal{X}}$ *with* $\bar{\mathcal{X}} = \mathbb{R}^{\bar{k}}$ *decomposes as* $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{\bar{k}})$*, and each* $\boldsymbol{p}_{\bar{k}} : \mathcal{X}_{\upsilon^{-1}(\bar{k})} \to \bar{\mathcal{X}}_{\bar{k}}$ *depends only on the micro-variables assigned to macro-variable* $\bar{k}$*.*

Crucially, what distinguishes our work from existing works is that we do not assume a priori that interventions on macro variables, and by extension, macro models, are well-defined.

4 WELL-DEFINED MACRO CAUSAL MODEL

Motivating Example. Consider a retail company with K products, each with sales volume $\mathbf{X}$ and unit price α_k. Total revenue is $Y = \sum_k \alpha_k X_k + N_Y$. The stakeholder tracks a macro-variable $\bar{X} = \sum_k X_k$ (total category sales) and wishes to intervene: "Increase category sales by c." This macro-intervention is ill-defined. If we allocate the sales increase c uniformly, and so each product's sales increases by $\delta_k = \frac{c}{K}$, then Y shifts by $\frac{c}{K} \cdot (\sum_k \alpha_k)$. Meanwhile, concentrating the change on product 1 shifts Y by $c \cdot \alpha_1$. Unless all prices are equal, the two implementations produce different macro-level causal effects on revenue, even though both realize the same shift in $\bar{X}$. The ambiguity arises because the aggregation map $\boldsymbol{p} : \mathbf{X} \mapsto \sum_k X_k$ discards information about the *composition* of sales—information that is causally relevant to profit when products are sold at different prices. Therefore, in this case we allow ourselves to construct aggregation mappings to represent macro-level noise variables, given by the map $\boldsymbol{q} : \mathcal{X} \times \mathcal{N} \to \bar{\mathcal{N}}$, with $\bar{\mathcal{N}} = \mathbb{R}^{\bar{k}}$. $\boldsymbol{q}$ and $\bar{X}$ then together fully determines Y, thus resolving ambiguity; *a priori*, an intervene on $\mathbf{X}$ may reflect on both $\bar{X}$ and $\boldsymbol{q}$. The class of well-defined interventions may emerge by constraining to those which keep $\boldsymbol{q}$ constant.

An implementation must only change the intervened macro variable and its descendants, but not the macro noise $\boldsymbol{q}$.

Definition 3 (Implementation)**.** *Given a micro-model* M*, aggregation maps* $\mathbf{p}$ *and* $\mathbf{q}$*, and a macro-intervention* $\bar{\boldsymbol{i}}_{\bar{k}}$ *on* $\bar{X}_{\bar{k}}$*, an* implementation *of* $\bar{\boldsymbol{i}}_{\bar{k}}$ *with respect to* $(\boldsymbol{p}, \boldsymbol{q})$ *is a micro-level intervention* $i_{\upsilon^{-1}(\bar{k})} : \mathcal{X}_{\upsilon^{-1}(\bar{k})} \to \mathcal{X}_{\upsilon^{-1}(\bar{k})}$ *s.t.:*

1. ***Macro-target:*** $\boldsymbol{p}_{\bar{k}} \circ i_{\upsilon^{-1}(\bar{k})} = \bar{\boldsymbol{i}}_{\bar{k}} \circ \boldsymbol{p}_{\bar{k}}$.
2. ***Noise preservation:*** $\boldsymbol{q}(\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}), \boldsymbol{n}) = \boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}), \boldsymbol{n})$ *for all* $\boldsymbol{n} \in \mathcal{N}$.

Criterion 1 ensures the implementation achieves the desired macro-shift. Criterion 2 ensures the complementary information is preserved, preventing off-target effects at the macro level.

Definition 4 (Well-defined macro-intervention)**.** *A macro-intervention $\bar{\boldsymbol{i}}_{\bar{k}}$ is* well-defined *if there exists at least one implementation, and for all implementations $\boldsymbol{i}, \boldsymbol{i}'$ of $\bar{\boldsymbol{i}}_{\bar{k}}$, $\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) = \boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i}')}(\boldsymbol{n})$ for all $\boldsymbol{n} \in \mathcal{N}$.*

Definition 5 (Well-defined macro causal model)**.** *Given a micro causal model and an aggregation map $\boldsymbol{p}$, a macro causal model exists if there is a function $\boldsymbol{q} : \mathcal{X} \times \mathcal{N} \to \bar{\mathcal{N}}$ and a macro structural function $\bar{\boldsymbol{f}} : \bar{\mathcal{X}} \times \bar{\mathcal{N}} \to \bar{\mathcal{X}}$ such that:*

1. *Every macro-state can be reached from every other via some well-defined macro-intervention.*
2. *For every implementation $\boldsymbol{i}$ of any macro-intervention $\bar{\boldsymbol{i}}$: $\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) = \bar{\boldsymbol{s}}^{(\bar{\boldsymbol{i}})}(\boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}), \boldsymbol{n}))$ for all $\boldsymbol{n}$.*

Criterion 2 parallels the exact transformation condition of Rubenstein et al. (2017), but requires consistency at the counterfactual rather than interventional level, as illustrated by the diagram to the right.

$$\begin{array}{ccc} \bar{\mathcal{N}} & \xrightarrow{\bar{\boldsymbol{s}}^{(\bar{\boldsymbol{i}})}} & \bar{\mathcal{X}} \\ {\scriptstyle \boldsymbol{q} \circ (\boldsymbol{s} \times \mathrm{id})} \uparrow & & \uparrow {\scriptstyle \boldsymbol{p}} \\ \mathcal{N} & \xrightarrow{\boldsymbol{s}^{(\boldsymbol{i})}} & \mathcal{X} \end{array}$$

Interpretation. $\boldsymbol{q}$ must encode all ambiguities of the implementations unseen by the aggregation map. Provided $\boldsymbol{q}$ is preserved, all implementations have consistent macro-level effects. If $\boldsymbol{q}$ is too simple, then ill-defined intervention effects return; if $\boldsymbol{q}$ is too rich (i.e. $\boldsymbol{q} = (\boldsymbol{x}, \boldsymbol{n})$), then no valid implementations achieve a non-zero macro change exists. In practice, we require $\bar{\mathcal{N}}$ to have the same dimensionality as $\bar{\mathcal{X}}$.

5 Linear Case: Unique Macro Noise Form

Set up. Consider a linear micro SCM: $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$, for which we assume solvable (Assumption 1) with solution $\boldsymbol{s}(\boldsymbol{n}) = (I - \boldsymbol{A})^{-1}\boldsymbol{n}$. The aggregation map is a full-rank linear map $\boldsymbol{P} \in \mathbb{R}^{\bar{k} \times k}$. We seek a linear complementary variable $\boldsymbol{q} = (\boldsymbol{q}_x \quad \boldsymbol{q}_n)$ mapping $(\boldsymbol{x}, \boldsymbol{n})$ to $\bar{\boldsymbol{n}}$, such that $\bar{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{x}$ and $\bar{\boldsymbol{n}} = \boldsymbol{q}(\boldsymbol{x}, \boldsymbol{n})$ together satisfies some linear causal model with structural equation $\bar{A}$, that is

$$\bar{\boldsymbol{x}} = \bar{A}\bar{\boldsymbol{x}} + \bar{\boldsymbol{n}} \tag{1}$$

Deriving the Complementary Variable. Noting that $\bar{\boldsymbol{x}}$ can be obtained from the micro-level cause variables by $\bar{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{x} = \boldsymbol{P}(A\boldsymbol{x} + \boldsymbol{n}) = \boldsymbol{P}\begin{pmatrix} A & I_K \end{pmatrix}\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{n} \end{pmatrix}$. We would get the desired form eq. (1) if we can factorise the r.h.s. through $\begin{pmatrix} \boldsymbol{P} & 0 \\ \boldsymbol{q}_x & \boldsymbol{q}_n \end{pmatrix}$. That is, consider the mapping $\boldsymbol{P}\begin{pmatrix} A & I_K \end{pmatrix}\begin{pmatrix} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{q}_x & \boldsymbol{q}_n \end{pmatrix}^{-1}\begin{pmatrix} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{q}_x & \boldsymbol{q}_n \end{pmatrix}\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{n} \end{pmatrix}$. *A priori*, $\begin{pmatrix} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{q}_x & \boldsymbol{q}_n \end{pmatrix}^{-1}$ is not a function; instead it is a mapping from the image to the preimage of $\begin{pmatrix} \boldsymbol{P} & \boldsymbol{0} \\ \boldsymbol{q}_x & \boldsymbol{q}_n \end{pmatrix}$. A necessary condition for attaining $eq.$ (1) is that $\boldsymbol{P}\begin{pmatrix} \boldsymbol{A} & I \end{pmatrix}\begin{pmatrix} \boldsymbol{P} & 0 \\ \boldsymbol{q}_x & \boldsymbol{q}_n \end{pmatrix}^{-1}$ is a well-defined function, which requires:

$$\mathrm{Ker}\begin{pmatrix} \boldsymbol{P} & 0 \\ \boldsymbol{q}_x & \boldsymbol{q}_n \end{pmatrix} \subseteq \mathrm{Ker}\left[\boldsymbol{P}\begin{pmatrix} \boldsymbol{A} & I \end{pmatrix}\right]. \tag{2}$$

Proposition 1 (Unique form for $\boldsymbol{q}$)**.** *Equation* (2) *and the macro model form Equation* (1) *together imply:*

$$\boxed{\boldsymbol{q} = \begin{pmatrix} \boldsymbol{P}\boldsymbol{A} - \bar{\boldsymbol{A}}\boldsymbol{P} & \boldsymbol{P} \end{pmatrix}} \tag{3}$$

for a macro-level causal matrix $\bar{\boldsymbol{A}} \in \mathbb{R}^{\bar{k} \times \bar{k}}$ chosen by the practitioner.

Proof is in C.1.

Remark 1 (Role of $\bar{\boldsymbol{A}}$)**.** *The choice of $\bar{\boldsymbol{A}}$ determines both the macro-level causal mechanism and, through Equation* (3)*, the complementary variable $\boldsymbol{q}$. How $\bar{\boldsymbol{A}}$ should be chosen depends on the practitioner's objective. For control, the practitioner may specify a desired macro-level response, and the resulting $\boldsymbol{q}$ characterises the micro-level implementations required to realise it, when such implementations exist. For auditing, the purpose is to understand the behaviour of the macro model under the distribution of feasible micro-level interventions determined by the set-up of the practitioner. $\bar{\boldsymbol{A}}$ should describe the effect of changing the macro cause variable on the outcome averaged uniformly over choices of micro-level implementations - this choice of $\bar{\boldsymbol{A}}$ is given by setting $\boldsymbol{q}$ to be uncorrelated with values of $\bar{\boldsymbol{A}}$. As pointed out by Zhu et al. (2024a), this choice of $\boldsymbol{q}$ corresponds to a stochastic intervention which leaves the ambiguity of implementations in their natural distribution, that is, we implement a macro intervention $do(\bar{X} := \bar{x})$ by replacing the generating mechanism of $\mathbf{X}$ with $do(\mathbf{X} \sim \rho_{\mathbf{X}|\boldsymbol{p}(\mathbf{X})=\bar{x}})$, where ρ is the observational measure of $\mathbf{X}$.*

Well-defined interventions. We now establish that $\boldsymbol{q}$ renders macro-causal effects well-defined.

Theorem 2 (All macro-interventions are well-defined)**.** *Under Assumptions 1 and 2, given the linear SCM $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$ with a full-rank linear aggregation map $\boldsymbol{P}$ and $\boldsymbol{q}$ as in equation 3, every macro-intervention $\bar{\boldsymbol{i}}$ that has at least one implementation is well-defined. For any implementation $\boldsymbol{i}$ of $\bar{\boldsymbol{i}}$:*

$$\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) = \bar{\boldsymbol{s}}^{(\bar{\boldsymbol{i}})}(\boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}), \boldsymbol{n})), \qquad \textit{with } \bar{\boldsymbol{s}}^{(\bar{\boldsymbol{i}})} = (I - \bar{\boldsymbol{A}})^{-1}$$

Proof is in Appendix C.2.

Motivating example revisited. If we assert a macro-level causal model from total sales volume $\bar{X} = \sum_k X_k$ to total revenue Y, we can express the micro-level structural equation $Y = \sum_k \alpha_k X_k + N_Y$ in the new coordinate system of $\bar{X}$ and $\boldsymbol{q}$: $Y = \beta \underbrace{\sum_k X_k}_{\bar{X}} + \underbrace{\sum_k (\alpha_k - \beta) X_k + N_Y}_{\boldsymbol{q}(\mathbf{X}, N_Y)}$. Note that we have to commit to a macro coefficient β. Now implementing a macro intervention $\text{do}(\bar{X} := \bar{X} + \bar{\delta})$ requires a change on $\mathbf{X}$ to attain the δ change on $\bar{X}$ *as well as* keeping $\boldsymbol{q}(\mathbf{X}, N_Y)$ fixed; for instance, a micro-level shift intervention $\text{do}(\mathbf{X} := \mathbf{X} + \boldsymbol{\delta})$ which implements $\text{do}(\bar{X} := \bar{X} + \bar{\delta})$ in the given macro-level model must satisfy 1. $\sum_k \delta_k = \bar{\delta}$, and 2. $\sum_k (\alpha_k - \beta)\delta_k$. With any valid $\boldsymbol{\delta}$ as implementation, we will clearly always increase Y by $\beta\bar{\delta}$ when $\bar{X}$ increases by $\bar{\delta}$.

Existence of Implementations. To establish when valid implementations exist, we need two additional mild conditions, the first requires we are compressing significantly from high to low dimensions, and the second is a technical condition which is typically satisfied because P and A usually arise independently.

Assumption 3 (High-dimensional micro-state)**.** *The micro-state dimension satisfies $k \geq 2\bar{k}$.*

Assumption 4 (Linear independence)**.** *The rows of $\boldsymbol{P}$ are linearly independent of those of $(\boldsymbol{P}\boldsymbol{A} - \bar{\boldsymbol{A}}\boldsymbol{P})(I - \boldsymbol{A})^{-1}$.*

Theorem 3 (Existence of macro shift interventions)**.** *Under Assumptions 1 to 4, for the linear SCM with $\boldsymbol{q}$ as in equation 3, every macro shift intervention $\bar{\boldsymbol{i}}(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}} + \bar{\boldsymbol{c}}$ has at least one implementation.*

Proof is in Appendix C.3.

Remark 2. *Since shift interventions can map any macro-state to any other, Theorems 2 and 3 together establish that $\boldsymbol{q}$ gives rise to a well-defined macro causal model in the sense of Definition 5.*

6 Nonlinear Case: Numerical scheme for implementation estimation

For a general nonlinear SCM $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n})$,

we seek $\boldsymbol{q}$ such that $\boldsymbol{p} \circ \boldsymbol{f} \circ (\boldsymbol{p}^*, \boldsymbol{q})^{-1}$ is a well-defined function $\bar{\boldsymbol{f}}$, where $\boldsymbol{p}^*(\boldsymbol{x}, \boldsymbol{n}) = \boldsymbol{p}(\boldsymbol{x})$. The complementary variable must satisfy:

$$\boldsymbol{p} \circ \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}) = \bar{\boldsymbol{f}}(\boldsymbol{p}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}, \boldsymbol{n})) \quad \text{for all } (\boldsymbol{x}, \boldsymbol{n}). \tag{4}$$

Well-defined interventions. While a unique form for $\boldsymbol{q}$ is generally unavailable, the well-definedness result extends:

Theorem 4 (Well-defined macro-interventions, nonlinear case)**.** *Suppose $\boldsymbol{q}$ satisfies equation 4 and Assumption 1 holds. Then for every macro-intervention $\bar{\boldsymbol{i}}$ and all implementations $\boldsymbol{i}, \boldsymbol{i}'$ of $\bar{\boldsymbol{i}}$ with respect to $(\boldsymbol{p}, \boldsymbol{q})$ have the same effect, that is: $\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) = \boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i}')}(\boldsymbol{n})$ and we may define $\bar{\boldsymbol{s}}^{(\bar{\boldsymbol{i}})}(\boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}), \boldsymbol{n})) := \boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) = \boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i}')}(\boldsymbol{n})$.*

Proof is in Appendix C.4.

Implementations as ODE solution. We realise that the micro-level solution function under an implementation can be expressed as a solution to an ODE. Denote a macro-shift-intervention direction by a unit vector $\bar{\boldsymbol{i}}$ and let $T\bar{\boldsymbol{i}}$ be the total shift intervention. Furthermore, let $\boldsymbol{I}(\boldsymbol{x}, T)$ denote the implementation of $T\bar{\boldsymbol{i}}$ in the micro-level; $\boldsymbol{I}(\boldsymbol{x}, t)$ parameterises a family of implementations which implement $t\bar{\boldsymbol{i}}$, $t \in [0, T]$.

For a given micro shift intervention $\boldsymbol{I}(\boldsymbol{x}, t)$, let $\boldsymbol{s}(\boldsymbol{n}, t)$ be the solution to the intervened micro causal model, then

$$\boldsymbol{s}(\boldsymbol{n}, t) = \boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n}) + \boldsymbol{I}(\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n}), t) \tag{5}$$

Substitute $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n})$ into condition 1 in Definition 3, $\boldsymbol{p}(\boldsymbol{x} + \boldsymbol{I}(\boldsymbol{x}, t)) = \boldsymbol{p}(\boldsymbol{x}) + t\bar{\boldsymbol{i}}$, and apply Equation (5):

$$\boldsymbol{p}(\boldsymbol{s}(\boldsymbol{n}, t)) - \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n})) = t\bar{\boldsymbol{i}} \tag{6}$$

Differentiating Equation (6) w.r.t. t gives the first condition on $\partial_t \boldsymbol{s}$:

$$\nabla \boldsymbol{p}(\boldsymbol{s}) - \nabla \boldsymbol{p}\big(\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n})\big) \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{n}) \partial_t \boldsymbol{s} = \bar{\boldsymbol{i}} \tag{7}$$

By the noise-preservation condition in Definition 3, we have the second condition on $\partial_t \boldsymbol{s}$:

$$\nabla_{\boldsymbol{x}} \boldsymbol{q}(\boldsymbol{s}, \boldsymbol{n}) \partial_t \boldsymbol{s} = \boldsymbol{0} \tag{8}$$

If solutions $\boldsymbol{s}$ exist, then together with an initial condition $\boldsymbol{s}(\boldsymbol{n}, 0)$, $\boldsymbol{s}$ may be solved for all of $[0, T]$ by integrating along $\partial_t \boldsymbol{s}$. Then $\boldsymbol{I}(\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, T), \boldsymbol{n}), T)$ is given by $\boldsymbol{s}(\boldsymbol{n}, T) - \boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, T), \boldsymbol{n})$.

and is well-defined under the following condition:

Assumption 5 (Consistency of intervention values)**.** *For all* $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{N}$*,*

$$\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, T), \boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}', T), \boldsymbol{n}') \quad \Longrightarrow \quad \boldsymbol{s}(\boldsymbol{n}, T) = \boldsymbol{s}(\boldsymbol{n}', T)$$

We seek an intervention field $\boldsymbol{i} : \mathcal{X} \to \mathcal{X}$ such that at each micro-state, the direction of intervention lies in the null space of the Jacobian of $\boldsymbol{q}$ (ensuring noise preservation) while achieving the macro-target.

The algorithm is summarised in Algorithm 1, Appendix A. Technical conditions for the existence of implementations are in Appendix C.5. Numerical integration approximate $\boldsymbol{s}(\boldsymbol{n}, T)$ to the accuracy of the corresponding solvers.

7 Case study: machine learning model auditing and editing

In mechanistic interpretability, the micro-level causal model is the neural network. There is growing interest in interpreting aggregations of neurons as human-understandable variables. For a network predicting an individual's profession from their biography, safety auditors seek to *understand* whether it encodes a sensitive attribute (e.g. gender) in its latent space and uses it to predict profession. Practitioners may also *control* the model by editing its latent representation, for example to increase the model's truthfulness. We distinguish these two goals of mechanistic interpretability:

- **Auditing** studies how model outputs depend on high-level concepts under input-level interventions actionable by the auditor. For example, a gender audit may examine how predictions change when gender cues change while other signals in the natural-language input are retained (De-Arteaga et al., 2019; Vig et al., 2020).
- **Editing** seeks interventions that achieve a desired high-level effect, such as increasing truthfulness or changing factual associations (Li et al., 2023; Zou et al., 2023; Meng et al., 2022). Unlike auditing, these interventions need not correspond to naturally realisable input changes: they may act directly on latent activations.

Our framework distinguishes auditing from editing through the macro noise $\boldsymbol{q}$. For auditing, the practitioner specifies interventions on network inputs, implicitly defining a distribution over feasible micro-level implementations. The objective is to characterise the average response under this distribution. Least-squares-fitting-with-intercept learns a macro model such that residuals $\boldsymbol{q}_{\text{audit}}$ have zero conditional mean under the fitting distribution (Section 7.1). For editing, implementations are selected to achieve a desired effect, without necessarily respecting the auditing distribution. They may modify internal activations directly rather than network inputs. Hence these implementations are invalid under the auditing model as they change $\boldsymbol{q}_{\text{audit}}$, but may be valid under a different macro model, with a different mechanism and macro noise $\boldsymbol{q}_{\text{edit}}$. This perspective thus resolves paradoxes observed in prior work: causal-tracing scores need not predict factual-editing success (Hase et al., 2023), and subspace patches can activate ordinarily dormant pathways (Makelov et al., 2024), we elaborate with the latter in Section 7.2.

7.1 Auditing Gender Effects in Profession Classification

Background. De-Arteaga et al. (2019); Vig et al. (2020) studied gender bias in a profession classification task where the model is asked to predict the individial's profession from their biography. Here we take a mechanistic interpretability perspective: instead of asking how the model respond to the gender feature in the input data, we audit its response to changes in a *model-internal gender representation*.

Setup. We use a frozen RoBERTa classifier (`Fannyjrd/roberta-bios`) that predicts 28 professions from biography text. Gender is conveyed through textual cues, not supplied as a separate classifier input. Our macro cause variable is gender $G(h)$, where the aggregation map G is learned by predicting the gender label from a layer's activation.[1] To measure the relevance of $G(h)$ with the actual gender label of the biography, we will report the held-out accuracy of G. For simpliciy of exposition, we choose a macro effect variable to be the same as the micro effect variable $\mathbf{Y}$, which is a vector of 28 centered profession logits.

[1] The aggregation map is often called a probe in mechinterp language.

Table 1: Top ten fitted macro coefficients, ordered by magnitude. Each coefficient measures the female-minus-male change in a centered profession logit.

Profession	Nurse	Surgeon	Dietitian	Yoga teacher	Paralegal	Architect	Interior designer	Chiropractor	Software engineer	Psychologist
Coefficient	$+0.62205$	-0.31926	$+0.28134$	$+0.26831$	$+0.24636$	-0.21439	$+0.19793$	-0.18827	-0.18311	$+0.17708$

Experiment 1: auditing the effect of gender. We use the dataset from (De-Arteaga et al., 2019), which contains biographies with their profession and gender labels. As we don't have access to a safety-audit practitioner, we simulate using an LLM to flip gender cues for 5000 biographies while maintaining other aspects of the text, such as length, tone, and contextual information, and we ask the LLM to 'use good judgement' where needed. This protocol defines the implementation distribution used by the audit. We split the data evenly for probe training, probe selection, macro model fitting, and held-out evaluation. For each original and edited biography in the macro-fitting split, we measure model's gender representation G_i and centered-logit vector $\mathbf{Y}_i$, and seek a macro model $\mathbf{Y} = \hat{\mathbf{b}} + \bar{A}G + \boldsymbol{q}$ by fitting

$$(\hat{\mathbf{b}}, \bar{A}) = \arg\min_{\mathbf{b},\bar{A}} \sum_i \left\| \mathbf{Y}_i - \mathbf{b} - \bar{A}G_i \right\|_2^2, \qquad \boldsymbol{q}(h) = \mathbf{Y}(h) - \hat{\mathbf{b}} - \bar{A}G(h).$$

Experiment 2: probability-level effects of gender interventions. To demonstrate the implication of the audit model, we intervene on gender according to the audit model found in experiment 1. We look for implementations $h \mapsto h + i$ of gender interventions $\bar{i}(g) = 1 - g$. Since the macro effect variable is a vector of logits, we transform it to probabilities with $\mathrm{softmax}(\mathbf{Y}(i(h)))$ and examine the resulting female-minus-male probability changes.

Experiment 1 results. We audit the model's output dependence on its representation of gender. A linear probe applied to the masked-mean embedding output (layer 0, the input-embedding representation layer before the transformer blocks) achieves 99.27%. Table 1 reports the ten largest-magnitude coefficients among the 28 professions. Positive coefficients indicate an increase in the centered profession logit when decoded gender changes from male to female. Nurse has the largest positive coefficient, with an increase of 0.622 centered-logit units under the specified audit distribution. The auditor could conclude from this that the model has a mild dependence on its gender representation.

We compare the network's responses to input-level gender edits with the predictions of the macro model. For each original–edited pair, we compare the rate of changes in the highest-scoring profession under input-level gender edits with what would be predicted by the macro model. For reference, we report the residual change $\Delta\mathbf{q}$.

Results are in Table 2. Input-level edits change $\mathbf{q}$ by a small amount, and as a result, profession flip rates are higher for input-level edits than predicted by the macro model fixing $\boldsymbol{q}$. This is because prediction-flip rates are sensitive to noises in the implementations: even if the fitted coefficient were zero, individual residual changes could increase a profession's relative score for some biographies and decrease it for others due to textual edits to flip gender may vary in different natural language contexts. The fitted macro coefficient characterises the average effect under the specified audit distribution, while $\Delta\mathbf{q}$ captures implementation-dependent variation.

Split	Mean absolute $\Delta\mathbf{q}$ per logit	Text profession flips	Fixed-$\mathbf{q}$ macro-predicted flips
Macro fit	0.15801	60/1213	13/1213
Calibration	0.16599	67/1223	13/1223
Held-out	0.16016	65/1220	13/1220

Table 2: Actual textual-edit effects versus fixed-residual macro predictions on retained pairs. Mean absolute residual change is averaged over pairs and all 28 logit coordinates. A profession flip is a change in the highest-scoring class.

Experiment 2 results. Sixteen biographies are selected: eight at random by a deterministic hash-based sample, four with small top-two logit margins, and four whose macro-predicted interventions change the highest-scoring profession. We construct activation edits using Structural Alignment Integration (Algorithm 1). Attainment of gender intervention and preservation of $\boldsymbol{q}$ associated with the binary gender macro model are verified at the final implementation. All 16 examples flip gender while preserving $\mathbf{q}$, with maximum componentwise residual error 1.68×10^{-4}. The maximum absolute discrepancy between actual and macro-predicted class probabilities is 2.75×10^{-5}, demonstrating causal consistency for these constructed interventions. Table 3 reports network predictions before and after implementation,

Table 3: Network predictions with $\mathbf{q}$-preserving activation edits. $G = 0$ is male and $G = 1$ female. Rows with change in the predicted profession are shaded. The probability columns describe the class with the largest absolute probability change, which need not be the highest-scoring class. Δ denotes after-minus-before change in percentage points.

G	Predicted profession before $\to$ after	Largest-change class	Class probability before $\to$ after	Δ (pp)
$1 \to 0$	journalist $\to$ journalist	journalist	$91.028\% \to 90.414\%$	-0.614
$0 \to 1$	professor $\to$ professor	professor	$89.849\% \to 91.564\%$	$+1.715$
$0 \to 1$	professor $\to$ professor	professor	$78.438\% \to 75.370\%$	-3.068
$1 \to 0$	journalist $\to$ journalist	journalist	$95.841\% \to 95.270\%$	-0.570
$1 \to 0$	nurse $\to$ teacher	nurse	$25.157\% \to 15.495\%$	-9.663
$0 \to 1$	physician $\to$ physician	physician	$49.606\% \to 50.581\%$	$+0.975$
$0 \to 1$	professor $\to$ professor	architect	$0.058\% \to 0.047\%$	-0.011
$0 \to 1$	teacher $\to$ teacher	teacher	$98.810\% \to 98.877\%$	$+0.067$
$0 \to 1$	physician $\to$ physician	surgeon	$34.752\% \to 29.194\%$	-5.558
$0 \to 1$	paralegal $\to$ paralegal	paralegal	$33.678\% \to 39.158\%$	$+5.479$
$0 \to 1$	architect $\to$ professor	architect	$49.637\% \to 44.244\%$	-5.393
$0 \to 1$	poet $\to$ poet	composer	$36.702\% \to 33.169\%$	-3.532
$0 \to 1$	physician $\to$ nurse	nurse	$21.821\% \to 35.026\%$	$+13.204$
$1 \to 0$	nurse $\to$ professor	nurse	$36.739\% \to 23.053\%$	-13.685
$1 \to 0$	nurse $\to$ pastor	nurse	$32.864\% \to 20.083\%$	-12.780
$0 \to 1$	rapper $\to$ model	model	$23.135\% \to 26.138\%$	$+3.003$

demonstrating probability changes predicted by the auditting model through valid activation edits. We observe from the results that if a model is very certain on the profession entailed in a biography, the after-edit profession prediction is less likely to change; the sensitive cases are when the model is indecisive about the profession, and gender bias may result in actual profession prediction changes there (e.g. changing from physician to nurse as gender changes from male to female). We include the original and editted biographies of all 16 examples in Appendix B.

7.2 Resolving interpretability illusion: Indirect Object Identification

Background. Wang et al. (2023) study indirect object identification (IOI) in GPT-2 Small: completing "When Mary and John went to the store, John gave a bottle of milk to" with "Mary", the indirect object rather than the subject. They identify "S-Inhibition heads" in layers 7 and 8 that write subject-position information into the residual stream for "Name Mover heads" in later layers. Makelov et al. (2024) find a direction $\mathbf{v}_{\mathrm{MLP}}$ in MLP 8's hidden activations along which patching[2] substantially changes predictions, despite its limited IOI role on the input distribution. This happens since $\mathbf{v}_{\mathrm{MLP}}$ combines a component in the output projection's kernel with an output-sensitive component: patching exploits the former's variation to activate a dormant pathway through the latter. The "illusion" is not causal ineffectiveness, but mistaking successful editing for the MLP's feature mediation under input distribution.

Experiment: auditing versus editing. Our framework distinguishes these interpretations through the implementation distribution and corresponding $\boldsymbol{q}$. Let $\boldsymbol{x}$ denote the selected MLP hidden activation and $\bar{\boldsymbol{x}} = \mathbf{v}_{\mathrm{MLP}}^{\top}\boldsymbol{x}$.

To isolate the MLP pathway, we fix a base prompt c and replace only its final-token MLP 8 hidden activation with $\boldsymbol{x}_i$, extracted from each dataset prompt. The residual stream and other activation sites at the intervention layer retain their base values; downstream computation proceeds normally. For each base prompt, we record the logits $\mathbf{Y}_i = (\mathrm{logit}^{(i)}_{\mathbf{subject}}, \mathrm{logit}^{(i)}_{\mathbf{object}})$ as macro effect variables and fit

$$(\hat{\mathbf{b}}_c, \hat{\mathbf{C}}_c) = \arg\min_{\mathbf{b},\mathbf{C}} \sum_i \|\mathbf{Y}_i - \mathbf{b} - \mathbf{C}\,\boldsymbol{\phi}(\bar{\boldsymbol{x}}_i)\|_2^2, \qquad \bar{\boldsymbol{x}}_i = \mathbf{v}_{\mathrm{MLP}}^{\top}\boldsymbol{x}_i.$$

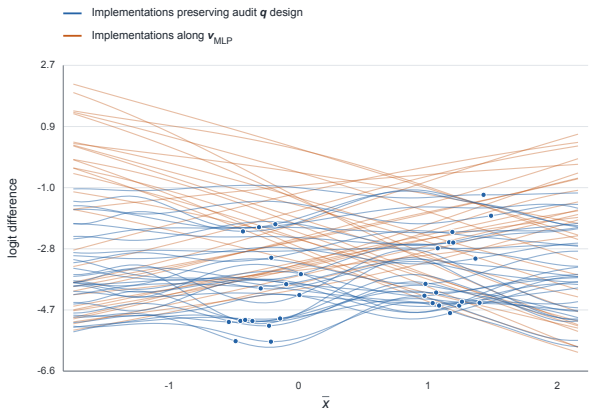
Here $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}^9$ contains nine nonredundant cubic B-spline basis functions, with eight breakpoints placed at equally spaced quantiles of the fitting coordinates, including their minimum and maximum. Write the fitted macro model as

$$\mathbf{Y}_c(\bar{\boldsymbol{x}}, \boldsymbol{q}_c) = \hat{\mathbf{b}}_c + \hat{\mathbf{C}}_c\,\boldsymbol{\phi}(\bar{\boldsymbol{x}}) + \boldsymbol{q}_c, \qquad \boldsymbol{q}_c(\boldsymbol{x}) = Y_c(\boldsymbol{x}) - \hat{\mathbf{b}}_c - \hat{\mathbf{C}}_c\,\boldsymbol{\phi}\big(\mathbf{v}_{\mathrm{MLP}}^{\top}\boldsymbol{x}\big)$$

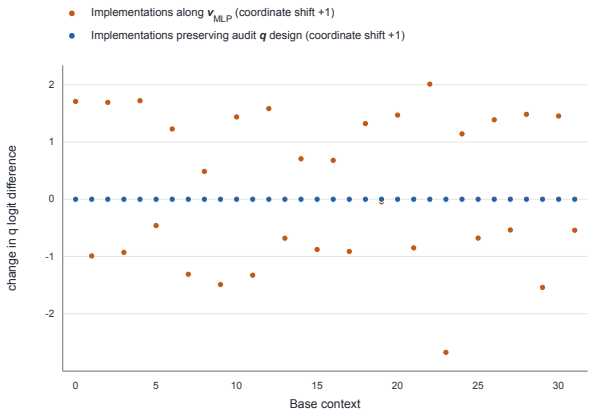
which tells us the average effect of changing the macro variable $\bar{\boldsymbol{x}}$ when the change arises from natural language changes in the prompt distribution. On the other hand, we may also change the macro variable $\bar{\boldsymbol{x}}$ by changing $\boldsymbol{x}$ along $\mathbf{v}_{\mathrm{MLP}}$. Crucially, notice that both implementations may result in the same change on the macro variable $\bar{\boldsymbol{x}}$, but may have different effects on the output logits, since the edit along $\mathbf{v}_{\mathrm{MLP}}$ would not preserve $\boldsymbol{q}$. Successful editing and a negligible audited contribution through the same site therefore need not contradict one another.

[2]Patching changes the activation's component along $\mathbf{v}_{\mathrm{MLP}}$.

Figure 1: **(a)** Subject-minus-indirect-object logit difference against $\bar{x}$. Blue curves are fixed-q macro predictions; blue dots are SAI implementations, one per context, confirming that the resulting effect is identical to macro model prediction. Orange curves show edits along $\mathbf{v}_{\mathrm{MLP}}$. **(b)** Signed changes in the auditing q when both methods implement $\bar{x} \mapsto \bar{x}+1$. Orange points use $x' = x + \mathbf{v}_{\mathrm{MLP}}/\|\mathbf{v}_{\mathrm{MLP}}\|_2^2$. Blue points attain the $\bar{x}$ target while preserving q_c. Orange points flip between pos. and neg. along the y-axis, matching the flipping slopes in (a); this is because $\mathbf{v}_{\mathrm{MLP}}$ encourages the name in the second position of the prompt, which is either the subject or the object depending on the base prompt.



(a) Fixed-q predictions and directional edits.



(b) Signed changes in the auditing complement.

Results. For each of 32 fixed base prompts, we fitted a spline auditing model. Figure 1a compares the predicted logit difference between the base subject and indirect object under interventions preserving the auditing macro noise q_c with the predicted logit difference responding to edits along $\mathbf{v}_{\mathrm{MLP}}$. The fixed-$q_c$ responses show weak dependence on $\bar{x}$: their mean within-context range is 1.03 logits, compared with 4.01 logits for $\mathbf{v}_{\mathrm{MLP}}$-edits over the same coordinate range. Constructed q_c-preserving implementations at the matched source targets agree with the macro predictions in 32 contexts, with maximum two-logit error 9.93×10^{-5}. Figure 1b compares changes in q_c under edits along $\mathbf{v}_{\mathrm{MLP}}$ that increase $\bar{x}$ by $+1$. $\mathbf{v}_{\mathrm{MLP}}$-edits produce large changes in q_c, which account for their departure from the fixed-q_c macro prediction. Their stronger effect therefore does not contradict the weaker causal response prescribed by the auditing model: auditing and editing are different questions and do not require the same implementations.

7.3 Discussion

In Section 7.1, we show a framework for auditing the model's processing mechanism of a sensitive attribute. Requiring q to have zero mean given the macro cause allows the practitioner to square with other audits concluding a different macro model - the implementations from other audits would not leave q in the current macro model with zero conditional mean. Then we illustrated the implications of the macro model by showing the macro coefficients and changes in the predicted professions as we implement gender interventions respecting the auditing model. In Section 7.2, we demonstrate that our framework is robust to a paradox observed in prior works: a valid subspace for model editing does not correspond to the subspace where the model implements the relevant signal. We show that auditing a macro variable constructed using the editing direction still correctly concludes (almost) no effect on the output. Crucially, our framework unifies editting and auditing: intervening along the editing direction changes q under the auditing model.

8 Conclusion

We propose a complementary variable q to remove ambiguous effects in causal abstraction, yielding a well-defined macro causal model. Our framework shifts causal abstraction from a verificational to a constructive paradigm. In the linear case, q admits a unique form and we establish that under mild conditions the existence of implementations for any macro shift interventions as well as the well-definedness of causal effects. In the nonlinear case, the same uniqueness holds, and we provide an ODE-based algorithm to construct valid implementations, and demonstrate it in experiments. In the case study, we audit machine learning models with our framework and resolves a well-known paradox in mechanistic interpretability; our framework can be readily extended to explain similar paradoxes such as the mismatch between causal localisation and editing identified in (Hase et al., 2023). [3]

[3]Code used for the case study will be released after the manuscript is published.

AI USE STATEMENT

We used generative AI tools for constructing gender edited biographies in Section 7.1. AI drafted the supplementary technical conditions included in Appendix C.5, and we reviewed the conditions and proofs. The experiment code was written by AI and we reviewed the code. We asked AI assistants to flesh out and simplify Algorithm 1 from our ODE solution idea and under our supervision. We also used AI for discovering related works in mechanistic interpretability and we studied the papers AI found ourselves. Finally, we used AI for drafting parts of the paper from our plans and for standard text editing and polishing; we have reviewed and take responsibility of the final version of the paper. We have not used AI for developing the main ideas and framework in the paper, for designing experiments, nor for proving any of the theoretical results included in the main paper.

REFERENCES

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021. doi: 10.1214/21-AOS2064.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Multi-level cause-effect systems. In *Proceedings of AISTATS*, 2016.

Maria De-Arteaga, Alexis Romber, Hanna Briggs, Alexandra Chouldechova, Deborah Raji, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *NeurIPS*, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2301.04213.

Erik P. Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013.

Armin Kekic, Bernhard Schölkopf, and Michel Besserve. Targeted causal abstraction. In *CLeaR*, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.03341.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.17030.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 68–87. PMLR, 2023. URL https://proceedings.mlr.press/v213/massidda23a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. ROME. arXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. MEMIT. arXiv:2210.07229.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2407.08734.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. INLP.

Eigil F. Rischel. The category theory of causal models. Master's thesis, University of Copenhagen, 2020.

Eigil F. Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1013–1023, 2021.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *UAI*, 2017.

Cosma R. Shalizi and Cristopher Moore. What is a macrostate? Subjective observations and objective dynamics. *arXiv:cond-mat/0303625*, 2003.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.12404.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-8328-0. URL https://www.mat.univie.ac.at/~gerald/ftp/book-ode/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. In *arXiv preprint arXiv:2308.10248*, 2023. Also known as Activation Addition (ActAdd). Verify final venue: a later version may have appeared at ICLR; cite the version you actually read.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.00593.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to makelov et al. (2023)'s "interpretability illusion" arguments. *arXiv preprint arXiv:2401.12631*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16042.

Yuchen Zhu, Kailash Budhathoki, Jonas M. Kübler, and Dominik Janzing. Meaningful causal aggregation and paradoxical confounding. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 1192–1217. PMLR, 2024a. URL https://proceedings.mlr.press/v236/zhu24a.html.

Yuchen Zhu, Sergio Hernan Garrido Mejia, Bernhard Schölkopf, and Michel Besserve. Unsupervised causal abstraction. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024b. URL https://openreview.net/forum?id=BP0e8RvFwd.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A NONLINEAR IMPLEMENTATION ALGORITHM

Algorithm 1 Structural Alignment Integration via the Solution Map

Require: Micro mechanism $\boldsymbol{f}$, macro noise $\boldsymbol{q}$, aggregation map $\boldsymbol{p}$, initial solution $\boldsymbol{s}_0$, macro-shift $\bar{\boldsymbol{i}}$, and strength T.
Ensure: An intervention $\boldsymbol{I}_T$ attaining the macro mechanism shift $T\bar{\boldsymbol{i}}$ while preserving $\boldsymbol{q}$.

1: **The velocity field.** For each state–noise pair $(\boldsymbol{x}, \boldsymbol{n})$, form

$$H(\boldsymbol{x}, \boldsymbol{n}) = \begin{bmatrix} \nabla \boldsymbol{p}(\boldsymbol{x}) - \nabla \boldsymbol{p}\big(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n})\big) \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}) \\ \nabla_{\boldsymbol{x}} \boldsymbol{q}(\boldsymbol{x}, \boldsymbol{n}) \end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix} \bar{\boldsymbol{i}} \\ \mathbf{0} \end{bmatrix}.$$

Select the minimum-norm feasible velocity:

$$V(\boldsymbol{x}, \boldsymbol{n}) = \arg\min_{\mathbf{v}} \left\{ \frac{1}{2} \|\mathbf{v}\|^2 : H(\boldsymbol{x}, \boldsymbol{n})\mathbf{v} = \mathbf{c} \right\}.$$

2: **Estimate the final solution map.** For a micro noise $\boldsymbol{n}$, numerically solve the ODE $\partial_t \boldsymbol{s}(\boldsymbol{n}, t) = V(\boldsymbol{x}(t), \boldsymbol{n})$, with initial value $\boldsymbol{s}_0(\boldsymbol{n})$ for $t \in [0, T]$.
3: **Return the micro-implementation.** $\boldsymbol{I}(\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, T), \boldsymbol{n}), T) = \boldsymbol{s}(\boldsymbol{n}, T) - \boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, T), \boldsymbol{n})$.

B BIOGRAPHIES AND CONSTRUCTED GENDER INTERVENTIONS

We report the original biographies for all 16 selected examples. Each "after" prediction is obtained by applying the saved embedding intervention to the original biography, not by substituting a text-edited biography. The decoded gender coordinate is $G = 0$ for male and $G = 1$ for female. Each row reports the logit of that row's predicted profession; when the prediction changes, the before and after scores therefore refer to different classes. Raw logits are denoted by ℓ_j. Centered logits are $Y_j = \ell_j - \frac{1}{28}\sum_{k=1}^{28} \ell_k$ and are the output coordinates used in $q = Y - b - \bar{A}G$. Centering leaves softmax probabilities unchanged. A prediction change is marked explicitly below. These examples were deliberately selected and do not estimate a population prediction-flip rate.

Example 1 (`gender-27200`). $G : 1 \to 0$. Profession prediction unchanged.

> She lives in Etna, New Hampshire. Edie Thys Morgan is a two-time Olympic skier, a former senior editor at SKI Magazine, and a freelance journalist. Her articles have appeared in O, The Oprah Magazine, Outside for Women, and Sports Illustrated for Women. She lives in Hanover, New Hampshire.

State	Predicted profession	Raw logit	Centered logit
Before	journalist	6.60161	7.02656
After	journalist	6.51683	6.94218

Example 2 (`gender-64323`). $G : 0 \to 1$. Profession prediction unchanged.

> He is also an Honorary Senior Lecturer at the Bartlett School of Architecture, University College London. He received a bachelor's degree in Biology from Pennsylvania State University, a doctorate in Genetics from Yale University and was a postdoctorate fellow under Jack Szostak at Harvard University. He has published in the area of protocells, complex systems, evolution and the origin of life in various journals including JACS and Science. He is developing novel synthetic chemical systems based on the properties of living systems. These synthetic systems are termed 'protocells' as they are model systems of primitive living cells and chemical examples of 'artificial' life.

State	Predicted profession	Raw logit	Centered logit
Before	professor	7.82825	8.26664
After	professor	7.82278	8.26971

Example 3 (`gender-10599`). $G : 0 \to 1$. Profession prediction unchanged.

Dr. Schielke specializes in the assessment and treatment of trauma-related and dissociative disorders, and his research focuses on increasing understanding of the treatment of these disorders and the relational components of psychotherapy. He completed a two-year postdoctoral fellowship in the assessment, differential diagnosis, and treatment of patients with complex trauma and dissociative disorders at The Trauma Disorders Program at Sheppard Pratt Health System in Baltimore, MD, and is the Co-Investigator for the Treatment of Patients with Dissociative Disorders (TOP DD) Network Study.

State	Predicted profession	Raw logit	Centered logit
Before	professor	7.72440	8.23771
After	professor	7.72115	8.24077

Example 4 (`gender-1075`). $G : 1 \to 0$. Profession prediction unchanged.

She recently joined Thirteen/WNET.org "WIDE ANGLE" as a production assistant on the documentary series "Women, War & Peace." She previously worked as an assistant producer at Voice of America and interned at 60 Minutes and WNYC's "The Leonard Lopate Show." Janisch graduated from the Columbia University Graduate School of Journalism. This piece was adapted from her master's thesis.

State	Predicted profession	Raw logit	Centered logit
Before	journalist	7.38996	7.68091
After	journalist	7.29712	7.59654

Example 5 (`gender-23204`). $G : 1 \to 0$. **Profession prediction changed.**

After giving birth to her daughter she soon realized that motherhood is not as easy as it looks in the movies and spent most of her days getting covered in both baby and puppy poop, pee, and puke. Shannon started a blog as a way to document her daughter's growth and vent about the daily challenges of being a mom. Luckily, she discovered many other women share the same experiences as her and now enjoys sharing her journey with moms all over the world on her appropriately named blog PoopPeePuke.com. Shannon is originally from Kansas but now resides in New York City with her awesome husband, daughter, and dog. In her free time she likes to sleep, eat pancakes, and wait Shannon has no free time because she is a mom! That's alright though because motherhood rocks!

State	Predicted profession	Raw logit	Centered logit
Before	nurse	3.01095	3.15919
After	teacher	2.46487	2.60749

Example 6 (`gender-24260`). $G : 0 \to 1$. Profession prediction unchanged.

He accepts Blue Cross Blue Shield EPO, Blue Cross Blue Shield Bronze, and Blue Cross Blue Shield HMO, as well as other insurance carriers. Dr. Golden is conversant in Spanish.

State	Predicted profession	Raw logit	Centered logit
Before	physician	6.45077	6.57199
After	physician	6.30170	6.42219

Example 7 (`gender-51121`). $G : 0 \to 1$. Profession prediction unchanged.

He received his PhD from the Hong Kong University of Science and Technology. His research interests include cement chemistry, properties, structures, ecological building materials, new phosphate bonded cements, water-resistant magnesia chloride cement, and durability of concrete.

State	Predicted profession	Raw logit	Centered logit
Before	professor	9.28036	9.73203
After	professor	9.27139	9.73508

Example 8 (`gender-3058`). $G : 0 \to 1$. Profession prediction unchanged.

> He is known for his ability to get a reluctant learner to buy into education. His lessons have stayed with students long after graduation, and he's often cited as students' favorite LHS teacher. He, too, will retire this year. "Greg is a very loyal, dedicated and hardworking teacher who is committed to science education," Glaze said. "He connects well with high school students and has motivated many students to learn more science that they would have without his inspiration. His professionalism will be missed!"

State	Predicted profession	Raw logit	Centered logit
Before	teacher	8.03097	8.26698
After	teacher	8.11296	8.35774

Example 9 (`gender-94131`). $G : 0 \to 1$. Profession prediction unchanged.

> He is also the super-hero known as Stars & Garters. Stars & Garters first encountered the Venture family when he and some other members of the CAL arrived at VenTech Tower to recruit Dr. Venture into the CAL's protection scheme. Tosh was unaware that the Ventures were already under the protection of the OSI, and he left, embarrassed.

State	Predicted profession	Raw logit	Centered logit
Before	physician	3.87696	4.22555
After	physician	3.72806	4.07575

Example 10 (`gender-32675`). $G : 0 \to 1$. Profession prediction unchanged.

> He has worked with prominent lawyers in raising public awareness to the dangers of asbestos. He holds a Bachelor's degree in Finance. In his free time, he enjoys attending concerts and spending time with friends.

State	Predicted profession	Raw logit	Centered logit
Before	paralegal	4.07271	4.14003
After	paralegal	4.31922	4.38627

Example 11 (`gender-39465`). $G : 0 \to 1$. **Profession prediction changed.**

> He earned his Bachelor of Science in Architecture and Master of Architecture degrees from The Ohio State University and is currently a PhD student in Architecture with concentration in Design Computation at the Georgia Institute of Technology. Jeff worked as a designer in the office of Mack Scogin Merrill Elam Architects from 2002 to 2009, working in various phases of design from concept through construction and in roles from model builder to senior project manager. This semester, he is teaching second year architectural design studio at Kennesaw State University and has previously taught architecture courses at Auburn University, Georgia Institute of Technology and Southern Polytechnic State University.

State	Predicted profession	Raw logit	Centered logit
Before	architect	6.71517	7.05107
After	professor	6.71094	7.04219

Example 12 (`gender-11100`). $G : 0 \to 1$. Profession prediction unchanged.

He makes his Young Masters debut in New York City at in the 'Travellers in Time/ Focus on New Work Exhbibition' at Site 109, LES. Adam Mysock was born in Cincinnati, Ohio in 1983 – the son of … Continue reading →

State	Predicted profession	Raw logit	Centered logit
Before	poet	4.51052	4.65102
After	poet	4.54399	4.68889

Example 13 (`gender-13492`). $G : 0 \to 1$. **Profession prediction changed.**

He is the co-founder of the Abstinence Clearinghouse, and served on the South Dakota Task Force to Study Abortion. He is a member of the Vote Yes For Life Campaign.

State	Predicted profession	Raw logit	Centered logit
Before	physician	3.72353	4.10760
After	nurse	4.22029	4.60673

Example 14 (`gender-84032`). $G : 1 \to 0$. **Profession prediction changed.**

She is also the co-founder of Suriname Indigenous Health Fund. Sarah led a team of Indigenous and church leaders to draft the World Council of Churches (WCC) Statement on the Doctrine of Discovery and its enduring impact on Indigenous Peoples, which was adopted in 2012, and organized an international coalition to deepen institutional commitment to an Indigenous-led Program at the WCC. Sarah is working within an international ecumenical movement to dismantle the Doctrine of Discovery and to build a solidarity movement promoting dialogue with resource extractive industries among people of faith. She is a member of Seattle Mennonite Church.

State	Predicted profession	Raw logit	Centered logit
Before	nurse	4.38941	4.78116
After	professor	4.21909	4.60897

Example 15 (`gender-18038`). $G : 1 \to 0$. **Profession prediction changed.**

Melissa's passions include worship, spiritual formation, outdoor adventures, health and fitness, wilderness medicine, and serving as mom to her two vivacious children.

State	Predicted profession	Raw logit	Centered logit
Before	nurse	3.66052	3.79442
After	pastor	3.63947	3.76770

Example 16 (`gender-11542`). $G : 0 \to 1$. **Profession prediction changed.**

In response to a hateful comment made about him, Ross recorded an empowering video encouraging youth to unapologetically be true to who they are, and to respond to messages of hate with messages of love. It Gets Better Project

State	Predicted profession	Raw logit	Centered logit
Before	rapper	3.23067	3.29494
After	model	3.17316	3.24885

C THEORETICAL RESULTS

C.1 LINEAR CLOSED FORM OF MACRO NOISE

Proof of Theorem 1. The kernel condition implies that the rows of $(\boldsymbol{q}_x \quad \boldsymbol{q}_n)$ lie in the row space of $(\boldsymbol{P} \quad 0)$ and $\boldsymbol{P}(\boldsymbol{A} \quad I)$. Hence $(\boldsymbol{q}_x \quad \boldsymbol{q}_n) = B(\boldsymbol{P} \quad 0) + C\boldsymbol{P}(\boldsymbol{A} \quad I)$ for matrices B, C. Comparing the noise block gives $\boldsymbol{q}_n = C\boldsymbol{P}$; since $\boldsymbol{P}$ is full-rank, C must be invertible. Requiring the macro model to have the form $\left(\bar{\boldsymbol{A}} \quad I\right)$ fixes $C = I$ and $B = -\bar{\boldsymbol{A}}$, yielding equation 3. □

C.2 WELL-DEFINEDNESS OF LINEAR MACRO-INTERVENTIONS

Proof of Theorem 2. Let $\boldsymbol{i}$ be any implementation of $\bar{\boldsymbol{i}}$ with respect to $(\boldsymbol{p}, \boldsymbol{q})$. By definition:

$$\begin{aligned}
\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) &= \boldsymbol{p} \circ \boldsymbol{i}(\boldsymbol{A}\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) + \boldsymbol{n}) = \bar{\boldsymbol{i}} \circ \boldsymbol{p}(\boldsymbol{A}\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) + \boldsymbol{n}) \\
&= \bar{\boldsymbol{i}}\left(\boldsymbol{P}\boldsymbol{A}\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) + \boldsymbol{P}\boldsymbol{n}\right) \\
&= \bar{\boldsymbol{i}}\left(\boldsymbol{q}(\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}), \boldsymbol{n}) + \bar{\boldsymbol{A}}\boldsymbol{P}\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n})\right) \\
&= \bar{\boldsymbol{i}}\left(\bar{\boldsymbol{A}}\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) + \boldsymbol{q}(\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}), \boldsymbol{n})\right).
\end{aligned}$$

The fourth equality uses the construction of $\boldsymbol{q}$: $\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{n}) = (\boldsymbol{P}\boldsymbol{A} - \bar{\boldsymbol{A}}\boldsymbol{P})\boldsymbol{x} + \boldsymbol{P}\boldsymbol{n} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{P}\boldsymbol{n} - \bar{\boldsymbol{A}}\boldsymbol{P}\boldsymbol{x}$.

This shows that $\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n})$ is a solution to the intervened macro SCM $\bar{\boldsymbol{x}} = \bar{\boldsymbol{i}}(\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{n}})$ with noise $\bar{\boldsymbol{n}} = \boldsymbol{q}(\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}), \boldsymbol{n}) = \boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}), \boldsymbol{n})$ (the last equality by criterion 2 of implementation). By Assumption 1, this solution is unique, so $\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n})$ is the same for all implementations $\boldsymbol{i}$. □

C.3 EXISTENCE OF MACRO SHIFT INTERVENTIONS

Proof of Theorem 3. By Definition 3, a micro-level shift $\boldsymbol{i}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\delta}$ implements $\bar{\boldsymbol{i}}$ if and only if :

$$\boldsymbol{P}\boldsymbol{\delta} = \bar{\boldsymbol{c}}, \tag{9}$$

$$(\boldsymbol{P}\boldsymbol{A} - \bar{\boldsymbol{A}}\boldsymbol{P})(I - \boldsymbol{A})^{-1}\boldsymbol{\delta} = \boldsymbol{0}. \tag{10}$$

Together these impose $2\bar{k}$ linear constraints on $\boldsymbol{\delta} \in \mathbb{R}^k$. Under Assumption 3, $k \geq 2\bar{k}$, and Assumption 4 ensures the $2\bar{k}$ constraints are independent. Hence the system has a solution space of dimension $k - 2\bar{k} > 0$. □

C.4 WELL-DEFINEDNESS OF NONLINEAR MACRO-INTERVENTIONS

Proof of Theorem 4. The proof follows the same structure as Theorem 2:

$$\begin{aligned}
\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}) &= \bar{\boldsymbol{i}} \circ \boldsymbol{p} \circ \boldsymbol{f}(\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}), \boldsymbol{n}) \\
&= \bar{\boldsymbol{i}} \circ \bar{\boldsymbol{f}}(\boldsymbol{p}(\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n})), \boldsymbol{q}(\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n}), \boldsymbol{n})) \\
&= \bar{\boldsymbol{i}} \circ \bar{\boldsymbol{f}}(\boldsymbol{p}(\boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n})), \boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}), \boldsymbol{n})),
\end{aligned}$$

where the last equality uses the noise-preservation criterion. This shows $\boldsymbol{p} \circ \boldsymbol{s}^{(\boldsymbol{i})}(\boldsymbol{n})$ solves the intervened macro SCM with noise $\boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}), \boldsymbol{n})$; uniqueness (Assumption 1) gives invariance over implementations. □

C.5 EXISTENCE AND APPROXIMATION OF IMPLEMENTATIONS IN THE NONLINEAR REGIME

Assumption 6 (Injectivity of the mechanism in noise)**.** *For every fixed $\boldsymbol{x} \in \mathbb{R}^k$, the map $\boldsymbol{n} \mapsto \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n})$ is injective. That is, for all $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{N}$,*

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}') \quad \Longrightarrow \quad \boldsymbol{n} = \boldsymbol{n}'.$$

Theorem 5 (Existence of valid implementations)**.** *Let $\mathcal{N} \subseteq \mathbb{R}^d$ be open, and let the micro-state space be $\mathbb{R}^k$. Suppose that $\boldsymbol{f}$, $\boldsymbol{p}$, and $\boldsymbol{q}$ are twice continuously differentiable, and that $\boldsymbol{s}_0 : \mathcal{N} \to \mathbb{R}^k$ is continuously differentiable and satisfies*

$$\boldsymbol{s}_0(\boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{s}_0(\boldsymbol{n}), \boldsymbol{n}).$$

Fix a macro-shift direction $\bar{\boldsymbol{i}}$ and strength $T > 0$. Let H and $\mathbf{c}$ be as defined in Algorithm 1. Assume:

1. ***Regularity and feasibility.*** *The matrix $H(\boldsymbol{x}, \boldsymbol{n})$ has constant rank, and*
$$\mathbf{c} \in \operatorname{range} H(\boldsymbol{x}, \boldsymbol{n})$$
for every $(\boldsymbol{x}, \boldsymbol{n}) \in \mathbb{R}^k \times \mathcal{N}$.

2. ***Controlled growth.*** *The minimum-norm velocity*
$$V(\boldsymbol{x}, \boldsymbol{n}) = H(\boldsymbol{x}, \boldsymbol{n})^\dagger \mathbf{c}$$
satisfies, for every compact $K \subseteq \mathcal{N}$,
$$\|V(\boldsymbol{x}, \boldsymbol{n})\| \le a_K + b_K \|\boldsymbol{x}\|$$
for constants $a_K, b_K \ge 0$, uniformly over $\boldsymbol{x} \in \mathbb{R}^k$ and $\boldsymbol{n} \in K$.

Then the initial-value problem
$$\partial_t \boldsymbol{s}(\boldsymbol{n}, t) = V(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n}), \qquad \boldsymbol{s}(\boldsymbol{n}, 0) = \boldsymbol{s}_0(\boldsymbol{n}),$$
has a unique solution on $[0, T]$ for each $\boldsymbol{n}$. It satisfies
$$\boldsymbol{p}(\boldsymbol{s}(\boldsymbol{n}, t)) - \boldsymbol{p}\big(\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n})\big) = t\bar{\boldsymbol{i}},$$
and
$$\boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n}) = \boldsymbol{q}(\boldsymbol{s}_0(\boldsymbol{n}), \boldsymbol{n}).$$
Define
$$\boldsymbol{s}_T(\boldsymbol{n}) = \boldsymbol{s}(\boldsymbol{n}, T), \qquad g_T(\boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{s}_T(\boldsymbol{n}), \boldsymbol{n}), \qquad \mathcal{U}_T = g_T(\mathcal{N}).$$
If Assumption 5 holds, then
$$\boldsymbol{I}_T(g_T(\boldsymbol{n})) = \boldsymbol{s}_T(\boldsymbol{n}) - g_T(\boldsymbol{n})$$
defines a single-valued intervention $\boldsymbol{I}_T : \mathcal{U}_T \to \mathbb{R}^k$ satisfying
$$\boldsymbol{p}\big(\mathbf{u} + \boldsymbol{I}_T(\mathbf{u})\big) = \boldsymbol{p}(\mathbf{u}) + T\bar{\boldsymbol{i}} \qquad \text{for all } \mathbf{u} \in \mathcal{U}_T,$$
and
$$\boldsymbol{s}_T(\boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{s}_T(\boldsymbol{n}), \boldsymbol{n}) + \boldsymbol{I}_T\big(\boldsymbol{f}(\boldsymbol{s}_T(\boldsymbol{n}), \boldsymbol{n})\big).$$
Under Appendix 6, $\boldsymbol{s}_T$ is the intervened solution map. Thus $\boldsymbol{I}_T$ is a valid implementation on $\mathcal{U}_T$.

If implementations are required on the entire mechanism-output domain, additionally require that $\mathcal{U}_T$ equals that domain, or that $\boldsymbol{I}_T$ admits an extension satisfying the same alignment condition there.

Proof. The constant-rank assumption implies that $H^\dagger$ is continuously differentiable. Consequently, $V = H^\dagger \mathbf{c}$ is locally Lipschitz in $\boldsymbol{x}$, giving local existence and uniqueness of the ODE solution.

For fixed $\boldsymbol{n} \in K$, the growth bound gives
$$\|\boldsymbol{s}(\boldsymbol{n}, t)\| \le \|\boldsymbol{s}_0(\boldsymbol{n})\| + \int_0^t (a_K + b_K \|\boldsymbol{s}(\boldsymbol{n}, \tau)\|)\, d\tau.$$
Gronwall's inequality implies, when $b_K > 0$,
$$\|\boldsymbol{s}(\boldsymbol{n}, t)\| \le e^{b_K t}\|\boldsymbol{s}_0(\boldsymbol{n})\| + \frac{a_K}{b_K}(e^{b_K t} - 1).$$
When $b_K = 0$, the bound is $\|\boldsymbol{s}_0(\boldsymbol{n})\| + a_K t$. Thus trajectories remain bounded on finite time intervals and extend to $[0, T]$ (Teschl, 2012, Theorem 2.17).

Feasibility gives $HV = \mathbf{c}$. By the chain rule,
$$\frac{d}{dt}\left[\boldsymbol{p}(\boldsymbol{s}(\boldsymbol{n}, t)) - \boldsymbol{p}\big(\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n})\big)\right] = \bar{\boldsymbol{i}},$$
and
$$\frac{d}{dt}\boldsymbol{q}(\boldsymbol{s}(\boldsymbol{n}, t), \boldsymbol{n}) = \mathbf{0}.$$

The initial structural equation makes the bracketed quantity zero at $t = 0$. Integration therefore gives

$$\boldsymbol{p}(\boldsymbol{s}_T(\boldsymbol{n})) - \boldsymbol{p}\big(\boldsymbol{f}(\boldsymbol{s}_T(\boldsymbol{n}), \boldsymbol{n})\big) = T\bar{\boldsymbol{i}},$$

and

$$\boldsymbol{q}(\boldsymbol{s}_T(\boldsymbol{n}), \boldsymbol{n}) = \boldsymbol{q}(\boldsymbol{s}_0(\boldsymbol{n}), \boldsymbol{n}).$$

Write

$$g_T(\boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{s}_T(\boldsymbol{n}), \boldsymbol{n}), \qquad \mathcal{U}_T = g_T(\mathcal{N}).$$

For $\mathbf{u} \in \mathcal{U}_T$, choose any $\boldsymbol{n}$ such that $g_T(\boldsymbol{n}) = \mathbf{u}$ and define

$$\boldsymbol{I}_T(\mathbf{u}) = \boldsymbol{s}_T(\boldsymbol{n}) - \mathbf{u}.$$

By Assumption 5, this value is independent of the chosen representative. The preceding identities imply

$$\boldsymbol{p}\big(\mathbf{u} + \boldsymbol{I}_T(\mathbf{u})\big) = \boldsymbol{p}(\mathbf{u}) + T\bar{\boldsymbol{i}},$$

and

$$\boldsymbol{s}_T(\boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{s}_T(\boldsymbol{n}), \boldsymbol{n}) + \boldsymbol{I}_T\big(\boldsymbol{f}(\boldsymbol{s}_T(\boldsymbol{n}), \boldsymbol{n})\big).$$

Thus $\boldsymbol{s}_T(\boldsymbol{n})$ is an intervened solution.

It remains to establish uniqueness. Fix $\boldsymbol{n}$ and suppose $\boldsymbol{x}'$ is another solution, with $\mathbf{u}' = \boldsymbol{f}(\boldsymbol{x}', \boldsymbol{n}) \in \mathcal{U}_T$. Choose $\boldsymbol{n}' \in \mathcal{N}$ such that $g_T(\boldsymbol{n}') = \mathbf{u}'$. By the definition of $\boldsymbol{I}_T$,

$$\boldsymbol{x}' = \mathbf{u}' + \boldsymbol{I}_T(\mathbf{u}') = \boldsymbol{s}_T(\boldsymbol{n}').$$

Consequently,

$$\boldsymbol{f}(\boldsymbol{x}', \boldsymbol{n}) = \mathbf{u}' = g_T(\boldsymbol{n}') = \boldsymbol{f}(\boldsymbol{s}_T(\boldsymbol{n}'), \boldsymbol{n}') = \boldsymbol{f}(\boldsymbol{x}', \boldsymbol{n}').$$

By Appendix 6, $\boldsymbol{n} = \boldsymbol{n}'$, and hence $\boldsymbol{x}' = \boldsymbol{s}_T(\boldsymbol{n})$. Therefore $\boldsymbol{s}_T$ is the unique solution map on the constructed mechanism-output domain. Together with alignment and noise preservation, this establishes validity of $\boldsymbol{I}_T$. □

Theorem 6 (Approximation accuracy)**.** *Suppose the conditions of Appendix 5 hold, and fix a compact noise domain $K \subseteq \mathcal{N}$.*

Let $\widehat{\boldsymbol{s}}_T$ approximate the terminal solution map with uniform error

$$\sup_{\boldsymbol{n} \in K} \|\widehat{\boldsymbol{s}}_T(\boldsymbol{n}) - \boldsymbol{s}_T(\boldsymbol{n})\| \le \varepsilon_s.$$

Assume that, on the relevant domains,

$$\|\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}) - \boldsymbol{f}(\boldsymbol{x}', \boldsymbol{n})\| \le L_f \|\boldsymbol{x} - \boldsymbol{x}'\|,$$
$$\|\boldsymbol{s}_T(\boldsymbol{n}) - \boldsymbol{s}_T(\boldsymbol{n}')\| \le L_s \|\boldsymbol{n} - \boldsymbol{n}'\|,$$

and

$$\|\boldsymbol{n} - \boldsymbol{n}'\| \le \kappa \|g_T(\boldsymbol{n}) - g_T(\boldsymbol{n}')\| \qquad \text{for all } \boldsymbol{n}, \boldsymbol{n}' \in K.$$

Define

$$\widehat{g}_T(\boldsymbol{n}) = \boldsymbol{f}(\widehat{\boldsymbol{s}}_T(\boldsymbol{n}), \boldsymbol{n}).$$

For an input $\mathbf{u} \in g_T(K)$, suppose the numerical inverse solver returns $\widehat{\boldsymbol{n}} \in K$ satisfying

$$\|\widehat{g}_T(\widehat{\boldsymbol{n}}) - \mathbf{u}\| \le \varepsilon_{\mathrm{inv}},$$

and define

$$\widehat{\boldsymbol{I}}_T(\mathbf{u}) = \widehat{\boldsymbol{s}}_T(\widehat{\boldsymbol{n}}) - \mathbf{u}.$$

Then

$$\boxed{\|\widehat{\boldsymbol{I}}_T(\mathbf{u}) - \boldsymbol{I}_T(\mathbf{u})\| \le (1 + L_s \kappa L_f)\varepsilon_s + L_s \kappa \varepsilon_{\mathrm{inv}}.}$$

The bound is uniform over inputs for which the stated inverse-solver condition holds.

In particular, explicit Euler integration with $h = T/N$ and exact initialization satisfies, for sufficiently small h,

$$\max_{0 \le j \le N} \sup_{\boldsymbol{n} \in K} \|\widehat{\boldsymbol{s}}_{jh}(\boldsymbol{n}) - \boldsymbol{s}_{jh}(\boldsymbol{n})\| \le C_s h,$$

where C_s is independent of h. Consequently,

$$\sup_{\mathbf{u}} \|\widehat{\boldsymbol{I}}_T(\mathbf{u}) - \boldsymbol{I}_T(\mathbf{u})\| = O(h + \varepsilon_{\mathrm{inv}}).$$

More generally, a uniform solution-map error $O(h^\ell)$ yields intervention error $O(h^\ell + \varepsilon_{\mathrm{inv}})$.

Proof. Write $\boldsymbol{n} = g_T^{-1}(\mathbf{u})$. The forward approximation satisfies

$$\|\widehat{g}_T(\widehat{\boldsymbol{n}}) - g_T(\widehat{\boldsymbol{n}})\| \le L_f \varepsilon_s.$$

Therefore,

$$\begin{aligned}\|\widehat{\boldsymbol{n}} - \boldsymbol{n}\| &\le \kappa \|g_T(\widehat{\boldsymbol{n}}) - g_T(\boldsymbol{n})\| \\ &\le \kappa\big(L_f \varepsilon_s + \varepsilon_{\mathrm{inv}}\big).\end{aligned}$$

Using the definitions of the two interventions,

$$\begin{aligned}\|\widehat{\boldsymbol{I}}_T(\mathbf{u}) - \boldsymbol{I}_T(\mathbf{u})\| &= \|\widehat{\boldsymbol{s}}_T(\widehat{\boldsymbol{n}}) - \boldsymbol{s}_T(\boldsymbol{n})\| \\ &\le \varepsilon_s + L_s \|\widehat{\boldsymbol{n}} - \boldsymbol{n}\| \\ &\le (1 + L_s \kappa L_f)\varepsilon_s + L_s \kappa \varepsilon_{\mathrm{inv}}.\end{aligned}$$

For the Euler estimate, continuous dependence on the noise and compactness of $K \times [0, T]$ imply that the exact trajectories lie in a common compact set. The velocity field is continuously differentiable on a neighbourhood of this set. Thus its spatial Lipschitz constant and the local truncation-error constant can be chosen uniformly over K. The Euler error recursion and a discrete Gronwall bound give the uniform $O(h)$ estimate. Substitution into the intervention bound completes the proof. □