Natural Latents: Latent Variables Stable Across Ontologies

John Wentworth* and David Lorell†
(Dated: March 17, 2025)

Suppose two Bayesian agents each learn a generative model of the same environment. We will assume the two have converged on the predictive distribution (i.e. distribution over some observables in the environment), but may have different generative models containing different latent variables. Under what conditions can one agent guarantee that their latents can be faithfully expressed in terms of the other agent's latents?

We give simple conditions under which such translation is guaranteed to be possible: the natural latent conditions. We also show that, absent further constraints, these are the most general conditions under which translatability is guaranteed.

I. BACKGROUND

When is robust translation possible at all, between agents with potentially different internal concepts, like e.g. humans and AI, or humans from different cultures? Under what conditions are scientific concepts guaranteed to carry over to the ontologies of new theories, (e.g. as general relativity reduces to Newtonian gravity in the appropriate limit?) When and how can choices about which concepts to use in creating a scientific model be rigorously justified, like e.g. factor models in psychology? When and why might a wide variety of minds in the same environment converge to use (approximately) the same concept internally?

These sorts of questions all run into a problem of indeterminacy, as popularized by Quine[1]: Different models can make exactly the same falsifiable predictions about the world, yet use radically different internal structures.

On the other hand, in practice we see that

- Between humans: language works at all. Indeed, babies are able to learn new words from only a handful of examples, therefore from an information perspective nearly all the work of identifying potential referents must be done before hearing the word at all.
- Between humans and AI: today's neural nets seem to contain huge amounts of human-interpretable structure, including apparent representations of human-interpretable concepts.[2]
- Between AI systems: the empirical success of grafting and merging[3], as well as the investigation by Huh *et al.* (2024)[4], suggests that different modern neural nets largely converge on common internal representations.

Combining those, we see ample empirical evidence of a high degree of convergence of internal concepts between different humans, between humans and AI, and between different AI systems. So in practice, it seems like convergence of internal concepts is not only possible, but in fact the default outcome to at least a large extent.

Yet despite the ubiquitous convergence of concepts in practice, we lack the mathematical foundations to provide robust *guarantees* of convergence. What properties might a scientist aim for in their models, to ensure that their models are compatible with as-yet-unknown future paradigms? What properties might an AI require in its internal concepts, to guarantee faithful translatability to or from humans' concepts?

In this paper, we'll present a mathematical foundation for addressing such questions.

II. THE MATH

A. Setup & Objective

We'll assume that two Bayesian agents, Alice and Bob, each learn a probabilistic generative model, M^A and M^B respectively. Each model encodes a distribution $P[X, \Lambda^i | M^i]$ over some "observable" random variables X and some "latent" random variables Λ^i. Each model makes the same predictions about observables X, i.e.

$$\forall x : P[X = x | M^A] = P[X = x | M^B] \qquad \text{(Agreement on Observables)}$$

* https://www.lesswrong.com/users/johnswentworth; Contact: johnswentworth@gmail.com
† https://www.lesswrong.com/users/david-lorell; Contact: d.lorell@yahoo.com

However, the two generative models may use completely different latent variables Λ^A and Λ^B in order to model the generation of X (thus the different superscripts for Λ). Note that there might also be additional observables over which the agents disagree; i.e. X need not be all of the observables in agents' full world models.

Crucially, we will assume that the agents can agree (or converge) on *some* way to break up X into individual observables $X_1, ..., X_n$. (We typically picture $X_1, ..., X_n$ as separated in time and/or space, but the math will not require that assumption.)

We require that the latents of each agent's model fully explain the interactions between the individual observables, as one would typically aim for when building a generative model. Mathematically, $X_1, ..., X_n \perp\!\!\!\perp |\Lambda^i, M^i$ (read "$X_1, ..., X_n$ are independent given Λ^i under model M^i"), or fully written out

$$\forall i, x, \lambda^i : P[X = x|\Lambda^i = \lambda^i, M^i] = \prod_j P[X_j = x_j|\Lambda^i = \lambda^i, M^i] \tag{Mediation}$$

Given that Alice' and Bob's generative models satisfy these constraints (Agreement on Observables and Mediation), we'd like necessary and sufficient conditions under which Alice can guarantee that her latent is a stochastic function of Bob's latent. In other words, we'd like necessary and sufficient conditions under which Alice' latent Λ^A is independent of X given Bob's latent Λ^B, for *any* latent which Bob might use (subject to the constraints). Also, we'd like all of our conditions to be robust to approximation.

We will show that:

- Necessity: In order to provide such a guarantee, Alice' latent Λ^A must be a "natural latent", meaning that it satisfies Mediation plus a Redundancy condition to be introduced shortly.
- Sufficiency: If Alice' latent Λ^A is a "natural latent", then it can be used to construct a new latent $\tilde{\Lambda}^A$ (via the Construction equation discussed later) which achieves the guarantee and has the same joint distribution with X as Λ^A.
- Both of the above are robust to approximation.

B. Notation

Throughout the paper, we will use the graphical notation of Bayes nets for equations. While our notation technically matches the standard usage in e.g. Pearl[5], we will rely on some subtleties which can be confusing. We will walk through the interpretation of the graph for the Mediation condition to illustrate.

The Mediation condition is shown graphically in Figure 1 . The graph is interpreted as an equation stating

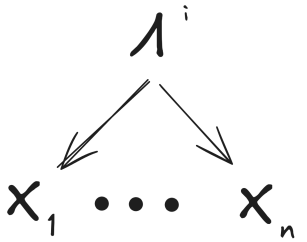

FIG. 1. The Mediation condition under the i^{th} model, graphically. A distribution $P[X_1, ..., X_n, \Lambda^i]$ "satisfies" this graph if and only if it satisfies the factorization $P[X_1, ..., X_n, \Lambda^i] = P[\Lambda^i] \prod_j P[X_j|\Lambda^i]$.

that the distribution over the variables factors according to the graph - in this case, $P[X, \Lambda^i] = P[\Lambda^i] \prod_j P[X_j|\Lambda^i]$. Any distribution which factors this way "satisfies" the graph. Note that **the graph does not assert that the factorization is minimal**; for example, a distribution $P[X_1, ..., X_n, \Lambda^i]$ under which all X_i and Λ^i are independent - i.e. $P[X_1, ..., X_n, \Lambda^i] = P[\Lambda^i] \prod_j P[X_j]$ - satisfies *all* graphs over the variables $X_1, ..., X_n$ and Λ^i, including the graph in Figure 1.

Besides allowing for compact presentation of equations and proofs, the graphical notation also makes it easy to extend our results to the approximate case. When the graph is interpreted as an approximation, we write it with an approximation error ϵ underneath, as in figure 2.

In general, we say that a distribution $P[Y_1, ..., Y_n]$ "satisfies" a graph over variables $Y_1, ..., Y_n$ to within approximation error ϵ if and only if $\epsilon \geq D_{KL}(P[Y_1, ..., Y_n]||| \prod_j P[Y_j|Y_{pa(j)}])$, where D_{KL} is the KL divergence. We will usually avoid writing out these inequalities explicitly.

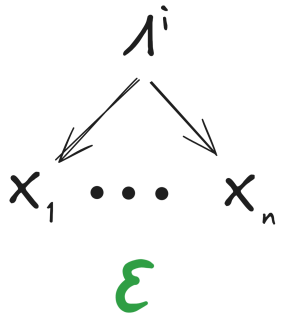

FIG. 2. The Mediation condition under the i^{th} model, graphically with approximation ϵ. A distribution $P[X_1, ..., X_n, \Lambda^i]$ "satisfies" this graph (including the approximation) if and only if it satisfies $\epsilon \geq D_{KL}(P[\Lambda^i, X]||P[\Lambda^i]\prod_j P[X_j|\Lambda^i])$.

C. Foundational Concepts: Mediation, Redundancy & Naturality

Mediation and redundancy are the two main foundational conditions which we'll work with.

Readers are hopefully already familiar with mediation. We say a latent Λ "mediates between" observables $X_1, ..., X_n$ if and only if $X_1, ..., X_n$ are independent given Λ. Intuitively, any information shared across two or more X_j's must "go through" Λ. We call such a Λ a mediator. Canonical example: if $X_1, ..., X_n$ are many rolls of a die of unknown bias Λ, then the bias is a mediator, since the rolls are all independent given the bias. See figure 1 for the graphical representation of mediation.

Redundancy is probably less familiar, especially the definition used here. We say a latent Λ' is a "redund" over observables $X_1, ..., X_n$ if and only if we can remove any one X_j from X and still have the same information about Λ', i.e. $X_j \to X_{\bar{j}} \to \Lambda'$ for all j, where $X_{\bar{j}}$ denotes all variables X_i except for X_j. Though most cases of interest are approximate, the redundancy condition in the exact case means:

$$\forall j, x, \lambda' : P[\Lambda' = \lambda', X = x] = P[\Lambda' = \lambda'|X = x]P[X = x] = P[\Lambda' = \lambda'|X_{\bar{j}} = x_{\bar{j}}]P[X = x] \qquad \text{(Redundancy)}$$

Intuitively, all information about Λ' must be redundantly represented across at least two X_j's, so that any one X_j can be dropped without any loss of information about Λ'. Canonical example: if $X_1, ..., X_n$ are pixel values in a picture of a bike, and Λ' is the color of the bike, then Λ' is a redund, since we can still tell what color the bike is even if any one (or few) pixels of the image are hidden. See 3 for the graphical representation of the redundancy condition.

FIG. 3. The graphical definition of redundancy: Λ' is a "redund" over $X_1, ..., X_n$ if and only if $P[X, \Lambda']$ satisfies the graph for all j. Intuitively, it says that one can drop any one element of X and still have the same information about Λ'

.

We'll be particularly interested in cases where a single latent is both a mediator and a redund over $X_1, ..., X_n$. We call mediation and redundancy together the "naturality conditions", and we call a latent satisfying both mediation and redundancy a "natural latent". Canonical example: if $X_1, ..., X_n$ are low level states of macroscopically separated chunks of a gas at thermal equilibrium, then the temperature is a natural latent over the chunks, since each chunk has the same temperature (thus redundancy) and the chunks' low-level states are independent given that temperature (thus mediation).

Justification of the name "natural latent" is the central purpose of this paper: roughly speaking, we wish to show that natural latents guarantee translatability, and that (absent further constraints) they are the *only* latents which guarantee translatability.

D. Core Theorems

We'll now present our core theorems. The next section will explain how these theorems apply to our motivating problem of translatability of latents across agents; readers more interested in applications and concepts than derivations should skip to the next section. We will state these theorems for generic latents Λ and Λ', which we will tie back to our two agents Alice and Bob later.

Theorem 1 (Mediator Bottlenecks Redund) *Suppose that random variables $X_1, ..., X_n$, Λ, and Λ' satisfy three conditions:*

- *Independent Latents:* $\Lambda \leftarrow X \rightarrow \Lambda'$
- Λ *Mediation:* $X_1, ..., X_n$ *are independent given* Λ
- Λ' *Redundancy:* $\Lambda' \rightarrow X_{\bar{j}} \rightarrow X_j$ *for all* j

Then $\Lambda' \rightarrow \Lambda \rightarrow X$.

The graphical statement of Theorem 1 is shown in figure 4, including approximation errors. The proof is given in Appendix B.

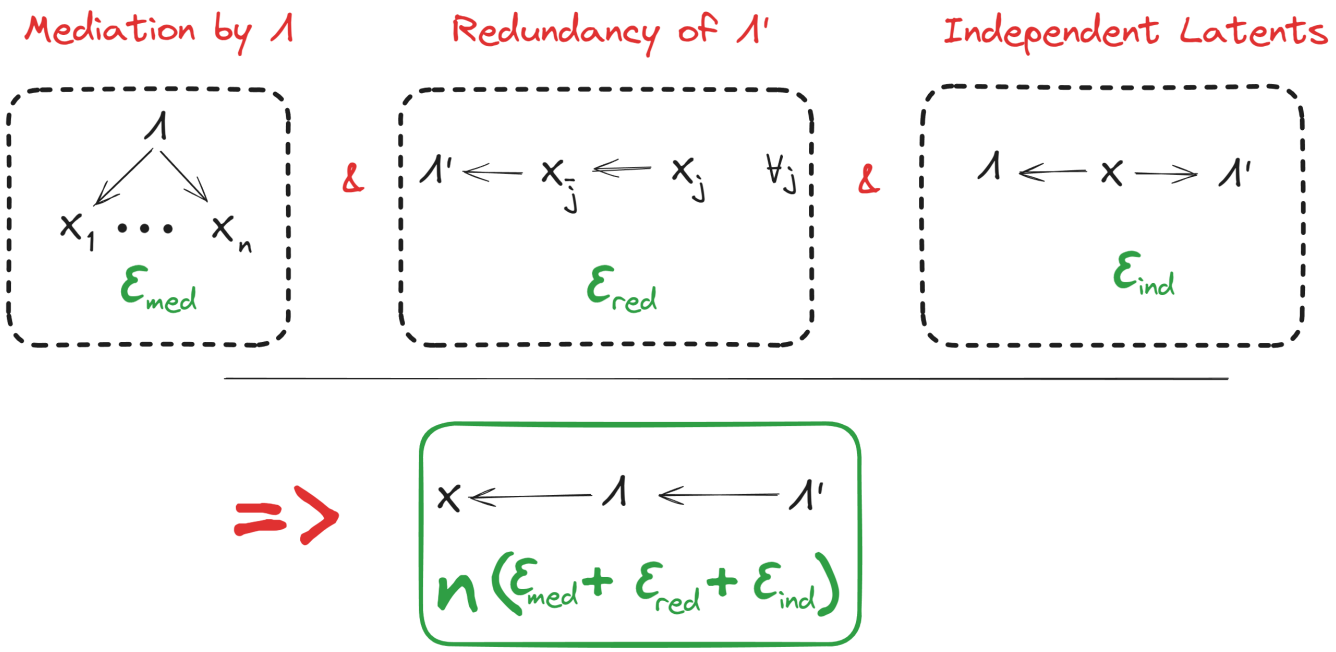

FIG. 4. Graphical statement of Theorem 1

The intuition behind the theorem is easiest to see when X has just two components, X_1 and X_2. The mediation condition says that the only way information can "move between" X_1 and X_2 is by "going through" Λ. The redundancy conditions say that X_1 and X_2 contain the same information about Λ', so intuitively, that information about Λ' must have "gone through" Λ - i.e. all the information which X yields about Λ' must also be present in Λ. Thus, $\Lambda' \rightarrow \Lambda \rightarrow X$; all information relevant to X in the redund must flow through the mediator, so the mediator bottlenecks the redund.

1. *Naturality* $\Longrightarrow$ *Minimality Among Mediators*

We're now ready for the corollaries which we'll apply to translatability in the next section.

Suppose a latent Λ is natural over $X_1, ..., X_n$ - i.e. it satisfies both the mediation and redundancy conditions. Well, Λ is a redund, so by Theorem 1, we can take *any other* mediator Λ'' (subject to the independent latents condition) and find that $\Lambda \rightarrow \Lambda'' \rightarrow X$. So: Λ is a mediator, and any *other* mediator (subject to the independent latents condition) screens off Λ from X. In this sense, Λ is the "minimal" mediator: any other mediator must contain at least the information about X which Λ contains. We sometimes informally call such a latent a "minimal latent".

Corollary 1.1 (Naturality $\implies$ Minimality Among Mediators) *Corollary is stated graphically; see Figure 5.*

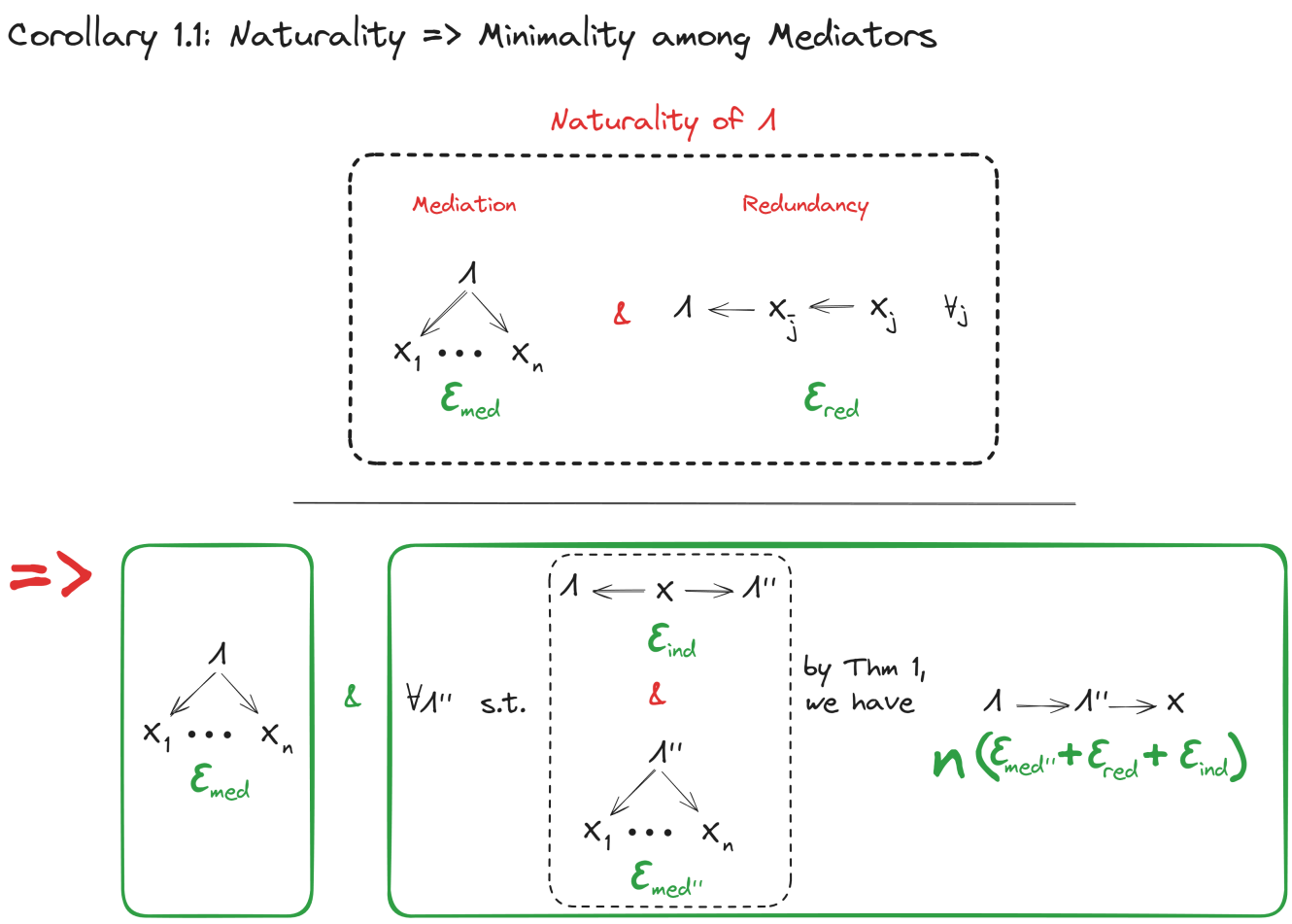

FIG. 5. If Λ is a natural latent, it satisfies the redundancy condition. Given *any* other latent Λ'' which satisfies the mediation condition (and is independent of Λ given X) we have the three conditions necessary to apply Theorem 1, i.e. Λ'' mediates between Λ and X.

2. *Naturality $\implies$ Maximality Among Redunds*

There is also a simple dual to "Naturality $\implies$ Minimality Among Mediators". While the minimal latent conditions describe a *smallest* latent which mediates between $X_1, ..., X_n$ (subject to the independent latents condition), the dual conditions describe a *largest* latent which is redundant across $X_1, ..., X_n$ (subject to the independent latents condition). We sometimes informally call such a latent a "maximal latent".

Corollary 1.2 (Naturality $\implies$ Maximality Among Redunds) *Corollary is stated graphically; see Figure 6.*

3. *Equivalence of Natural Latents*

If two latents Λ, Λ' are both natural latents, then from Theorem 1 we trivially have both $\Lambda' \to \Lambda \to X$ and $\Lambda \to \Lambda' \to X$. In English: the two latents contain the same information about X. In that sense, the two are equivalent.

However, two different natural latents could also contain different independent random noise. For instance, we can take any natural latent over X, append to the latent a single coinflip independent of X, and the result will also be a natural latent.

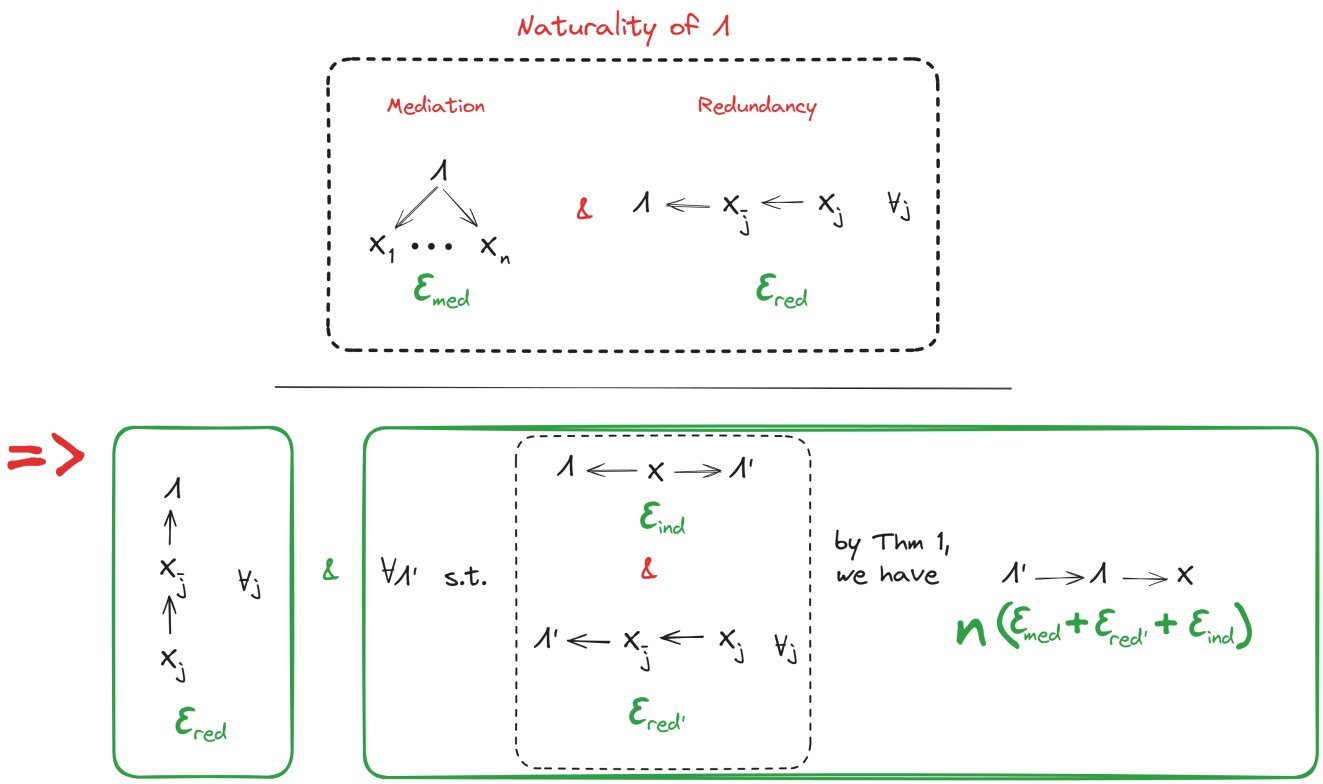

FIG. 6. If Λ is a natural latent, it satisfies the mediation condition. Given *any* other latent Λ' which satisfies the redundancy condition (and is independent of Λ given X) we have the three conditions necessary to apply Theorem 1, i.e. Λ mediates between Λ' and X.

III. APPLICATION TO TRANSLATABILITY

A. Motivating Question

Our main motivating question is: under what conditions on Alice' model M^A and its latent(s) Λ^A can Alice *guarantee* that Λ^A is a stochastic function of Λ^B (i.e. $\Lambda^A \leftarrow \Lambda^B \leftarrow X$, which is equivalent to $\Lambda^A \rightarrow \Lambda^B \rightarrow X$), for *any* model M^B and latent(s) Λ^B which Bob might have?

Recall that we already have some restrictions on Bob's model and latent(s): Agreement on Observables says $P[X|M^B] = P[X|M^A]$, and Mediation says that $X_1, ..., X_n$ are independent given Λ^B under model M^B.

Since Naturality $\implies$ Minimality Among Mediators, the natural latent conditions seem like a good fit here. If Alice' latent Λ^A satisfies the natural latent conditions, then Minimality Among Mediators says that for *any* latent Λ'' satisfying mediation over $X_1, ..., X_n$ (subject to the independent latents condition), $\Lambda^A \rightarrow \Lambda'' \rightarrow X$. And Bob's latent Λ^B satisfies mediation, so we can take $\Lambda'' = \Lambda^B$ to get the result we want... IF that pesky independent latents condition is satisfied. And for that, we need a digression.

B. Under Which Model?

By assumption, both Alice's and Bob's models agree on the distribution of the observables $P[X] := P[X|M^A] = P[X|M^B]$. The only degree of freedom their models have is therefore the distribution of latent(s) given observables, i.e. the choice of $P[\Lambda^i|X, M^i]$. We can view any choice of $P[\Lambda^i|X, M^i]$ as *defining* the latent variable(s) Λ^i. That definition is the first step toward translating Alice' latent Λ^A into Bob's model (or vice-versa): Bob simply defines a new latent Λ^A in his own model using the same defining distribution as Alice, i.e. $P[\Lambda^A|X, M^B] := P[\Lambda^A|X, M^A]$.

However, this still leaves one important degree of freedom: the distribution of X is locked in, the distribution of Λ^A given X is locked in, the distribution of Λ^B given X is locked in, but there's still (potentially) room for different interactions between Λ^A and Λ^B conditional on X.

To handle that degree of freedom, we will assume that Λ^A and Λ^B are taken to be independent given observables X, i.e. the two agents' latents are related only via the observables. This "independent latents" condition is shown graphically in figure 7.

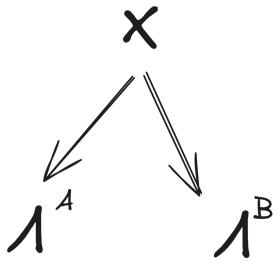

FIG. 7. That pesky independent latents condition: independence of Λ^A and Λ^B, conditional on X.

Note that Alice can always force satisfaction of the independent latents condition by construction: from whatever latent Λ^A she originally uses, she can construct $\tilde{\Lambda}^A$ by sampling from the distribution of Λ^A given X, using new independent bits for any nondeterminism in the sampling function. Mathematically, $\tilde{\Lambda}^A$ is defined as a stochastic function of X, with distribution

$$\forall \lambda^A, \lambda^B, x : P[\tilde{\Lambda}^A = \lambda^A | X = x, \Lambda^B = \lambda^B] = P[\Lambda^A = \lambda^A | X = x] \qquad \text{(Construction)}$$

Alternatively, as a special case, if Λ^A is a deterministic function of X then the independent latents condition is automatically satisfied. And for purposes of our proofs, the approximate version of the independent latents condition is sufficient, so if Λ^A is approximately a deterministic function of X (i.e. Λ^A has small entropy given X) then that also suffices.

With the independent latents condition, there is no longer any "under which model?" ambiguity; we can unambiguously write $P[X, \Lambda^A, \Lambda^B] := P[X]P[\Lambda^A|X, M^A]P[\Lambda^B|X, M^B]$. If the independent latents condition is assumed only approximately, then there is still some ambiguity, but any allowed choice requires a small D_{KL} between $P[X, \Lambda^A, \Lambda^B]$ and $P[X]P[\Lambda^A|X, M^A]P[\Lambda^B|X, M^B]$.

C. Guaranteed Translatability

From here on out, we will assume that Alice' and Bob's latents satisfy the independent latents condition.

With that handled, we can now finally declare half of our main theorem for this paper. If Alice' latent Λ^A is natural, then it's a stochastic function of Bob's latent Λ^B, i.e. $\Lambda^A \to \Lambda^B \to X$ (or equivalently, $\Lambda^A \leftarrow \Lambda^B \leftarrow X$). This is just the Naturality $\implies$ Minimality Among Mediators theorem from earlier.

Now it's time for the other half of our main theorem: the naturality conditions are the *only* way for Alice to achieve this guarantee. In other words, we want to show the converse of Naturality $\implies$ Minimality Among Mediators: if Alice' latent Λ^A satisfies Mediation, and for ANY latent Λ^B Bob could choose (i.e. any other mediator) we have $\Lambda^A \to \Lambda^B \to X$, then Alice' latent must be natural.

The key to the proof is to notice that $X_{\bar{j}}$, i.e. all of X except X_j, is always a mediator. So, Bob could choose $\Lambda^B = X_{\bar{j}}$ for any j. In order to achieve her guarantee, Alice' latent Λ^A must therefore satisfy $\Lambda^A \to X_{\bar{j}} \to X$ for all j, which is equivalent to $\Lambda^A \to X_{\bar{j}} \to X_j$ - i.e. the redundancy condition. Alice' latent already had to satisfy the mediation condition by assumption, it must also satisfy the redundancy condition in order to achieve the desired guarantee, therefore it must be a natural latent.

Theorem 2 (Guaranteed Translatability) *The theorem is stated graphically; see figure 8*

In English, the assumptions required for the theorem are:

- The independent latents condition $\Lambda^A \leftarrow X \to \Lambda^B$ (achievable either by Construction or by Alice' latent being an approximately-deterministic function of X)

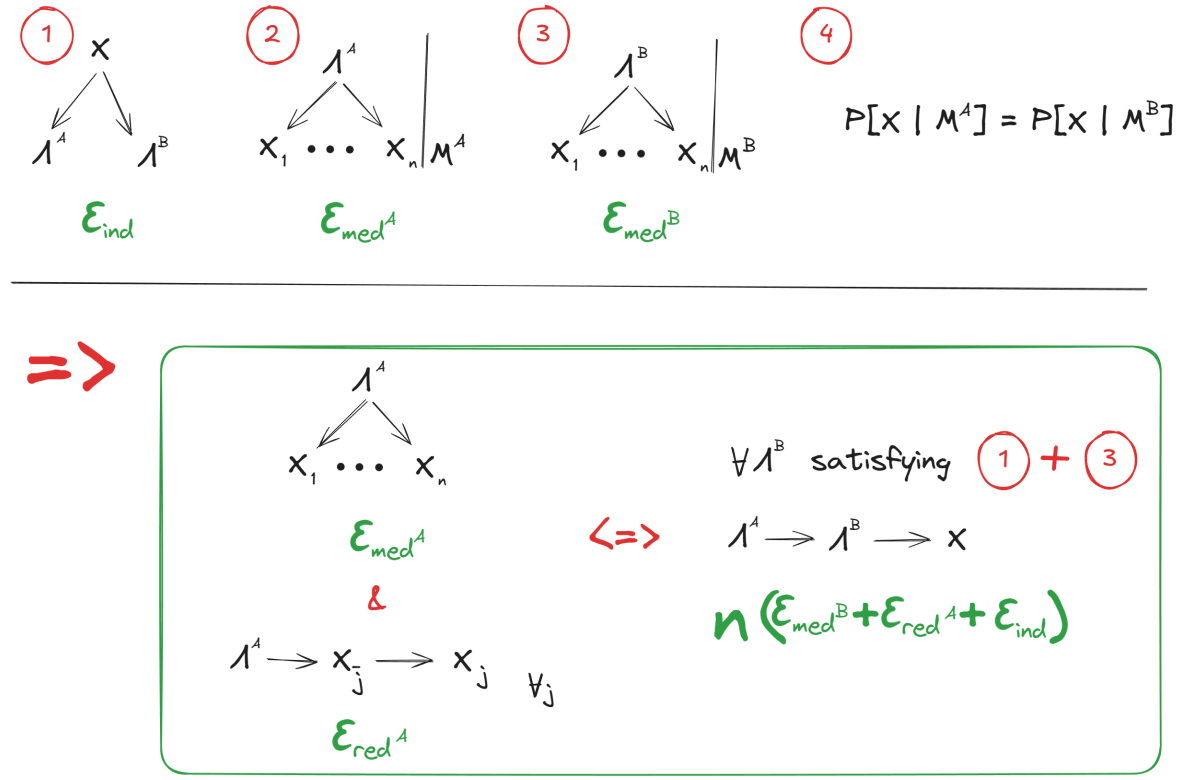

FIG. 8. Graphical statement of Theorem 2

- The Mediation conditions: $X_1, ..., X_n$ are independent given Λ^A under model M^A, and same given Λ^B under model M^B
- The Agreement on Observables condition: $P[X|M^A] = P[X|M^B]$

Under those constraints, Alice can guarantee that her latent Λ^A is a stochastic function of Bob's latent Λ^B (i.e. $\Lambda^A \leftarrow \Lambda^B \leftarrow X$) if and only if Alice' latent is a natural latent over X, meaning that it satisfies both the mediation condition (already assumed) and the redundancy condition $\Lambda^A \rightarrow X_{\bar{j}} \rightarrow X_j$ for all j.

Proof: the "if" part is just Naturality $\implies$ Minimality Among Mediators; the "only if" part follows trivially from considering $\Lambda^B = X_{\bar{j}}$ (which is always an allowed choice of Λ^B).

IV. NATURAL LATENTS: INTUITION & EXAMPLES

Having motivated the natural latent conditions as exactly those conditions which guarantee translatability, we move on to building some intuition for what natural latents look like and when they exist.

A. When Do Natural Latents Exist? Some Intuition From The Exact Case

For a given distribution $P[X_1, ..., X_n]$, a natural latent over X does not always exist, whether exact or approximate to within some error ϵ. In practice, the cases where interesting natural latents *do* exist usually involve approximate natural latents (as opposed to exact), and we'll see some examples in the next section. But first, we'll look at the exact case, in order to build some intuition.

Let's suppose that there are just two observables X_1, X_2. If Λ is natural over those two observables, then the redundancy conditions say $P[\Lambda, X] = P[\Lambda|X_1]P[X] = P[\Lambda|X_2]P[X]$, i.e. $P[\Lambda|X_1] = P[\Lambda|X_2]$ for all X_1, X_2 such that $P[X_1, X_2] > 0$. This condition can be somewhat confusing at first, but we can express it more intuitively by defining two functions:

- $f_1(x_1) := (\lambda \mapsto P[\Lambda = \lambda|X_1 = x_1])$
- $f_2(x_2) := (\lambda \mapsto P[\Lambda = \lambda|X_2 = x_2])$

In other words, f_j takes in a value of X_j, and returns the entire distribution of Λ given X_j. In terms of f_1, f_2, the redundancy condition then says

$$f_1(X_1) = f_2(X_2) \text{ with probability 1} \tag{1}$$

This is a deterministic constraint between X_1 and X_2.

While we won't prove it here, it turns out that we can take $\Lambda' := f_1(X_1) = f_2(X_2)$, and Λ' will also be a natural latent (assuming the Λ which we used to construct f_1, f_2 is an exact natural latent, as opposed to an approximate natural latent). We'll use Λ' instead of Λ as our natural latent in the next step.

Next, the mediation condition. The mediation condition says that X_1 and X_2 are independent given Λ' - i.e. they're independent given the value of the deterministic constraint. So: assuming existence of a natural latent Λ, X_1 and X_2 must be independent given the value of a deterministic constraint.

On the other hand, if X_1 and X_2 are independent given the value of a deterministic constraint, then the value of the constraint clearly satisfies the natural latent conditions.

That gives us an intuitive characterization of the existence conditions for exact natural latents: an exact natural latent between two variables exists if-and-only-if the variables are independent given the value of a deterministic constraint. More generally, an exact natural latent over many variables exists if-and-only-if the variables are independent given the value of some deterministic constraints over subsets of the variables.

B. Worked Quantitative Example of Theorem 1

Consider a biased coin with bias θ. Two individuals, Alice and Bob, each flip the coin 1000 times and compute the median of their respective flips. For simplicity, we assume a uniform prior on θ over the interval $[0, 1]$.

Intuitively, if the bias θ is unlikely to be very close to $\frac{1}{2}$, Alice and Bob will find the same median with high probability. Let X_1 and X_2 denote Alice's and Bob's 1000 flips, respectively, and let Λ represent the bias θ. Note that the flips are independent given θ, satisfying the mediation condition of Theorem 1 exactly. Let Λ' be the median computed by either Alice or Bob (assuming they are the same with high probability). Since the same median can be computed with high probability from either X_1 or X_2, the redundancy condition is approximately satisfied. Finally, since the median is a deterministic function of X, the independent latents condition is satisfied exactly.

Theorem 1 then implies that the bias approximately mediates between the median (either Alice's or Bob's) and the coinflips X. To quantify the approximation, we first quantify the approximation on the redundancy condition (the other two conditions hold exactly, so their ϵ's are 0). Taking Λ' to be Alice's median, Alice's flips mediate between Bob's flips and the median exactly (i.e., $\Lambda' \to X_1 \to X_2$), but Bob's flips mediate between Alice's flips and the median (i.e., $\Lambda' \to X_2 \to X_1$) only approximately. The corresponding D_{KL} is given by:

$$\begin{aligned} D_{KL}(P[X_1, X_2, \Lambda']||P[X_1]P[X_2|X_1]P[\Lambda'|X_2]) &= -\sum_{X_1, X_2} P[X_1, X_2] \ln P[\Lambda'(X_1)|X_2] \\ &= \mathbb{E}[H(\Lambda'(X_1)|X_2)] \end{aligned}$$

This is a Dirichlet-multinomial distribution, so it is cleaner to rewrite in terms of $N_1 := \sum X_1$, $N_2 := \sum X_2$, and $n := 1000$. Since Λ' is a function of N_1, the D_{KL} becomes:

$$= \mathbb{E}[H(\Lambda'(N_1)|N_2)]$$

Writing out the distribution and simplifying the gamma functions, we obtain:

$$P[N_2] = \frac{1}{n+1} \text{ (i.e., uniform over 0, ..., n)}$$
$$P[N_1|N_2] = \frac{\Gamma(n+2)\Gamma(N_2+1)\Gamma(n-N_2+1)}{\Gamma(n+1)\Gamma(N_1+1)\Gamma(n-N_1+1)} \frac{\Gamma(N_1+N_2+1)\Gamma(2n-N_1-N_2+1)}{\Gamma(2n+2)}$$
$$P[\Lambda'(N_1)=0|N_2] = \sum_{n_1<500} P[N_1|N_2]$$
$$P[\Lambda'(N_1)=1|N_2] = \sum_{n_1>500} P[N_1|N_2]$$

There are only 1001^2 values of (N_1, N_2), so these expressions can be combined and evaluated using a Python script (see Appendix C for code). The script yields $H = 0.058$ bits. As a sanity check, the main contribution to the entropy should be when θ is near 0.5, in which case the median should have roughly 1 bit of entropy. With n data points, the posterior uncertainty should be of order $\frac{1}{\sqrt{n}}$, so the estimate of θ should be precise to roughly $\frac{1}{30} \approx .03$ in either direction. Since θ is initially uniform on $[0, 1]$, a distance of 0.03 in either direction around 0.5 covers about 0.06 in prior probability, and the entropy should be roughly 0.06 bits, which is consistent with the computed value.

Returning to Theorem 1, we have $\epsilon_1 = 0$, $\epsilon_3 = 0$, and $\epsilon_2 \approx 0.058$ bits. Thus, the theorem states that the coin's true bias approximately mediates between the coinflips and Alice's median, to within $2(\epsilon_1 + \epsilon_2 + \epsilon_3) \approx 0.12$ bits.

Exercise for the Reader: By tracking the ϵ's more carefully through the proof, show that, for this example, the coin's true bias approximately mediates between the coinflips and Alice's median to within ϵ_2, i.e., roughly 0.058 bits.

C. Intuitive Examples of Natural Latents

This section will contain no formalism, but will instead walk through a few examples in which one would *intuitively* expect to find a nontrivial natural latent, in order to help build some intuition for the reader. The When Do Natural Latents Exist? section provides the foundations for the intuitions of this section.

1. Ideal Gas

Consider an equilibrium ideal gas in a fixed container, through a Bayesian lens. Prior to observing the gas, we might have some uncertainty over temperature. But we can obtain a very precise estimate of the temperature by measuring any one mesoscopic chunk of the gas. That's an approximate deterministic constraint between the low-level states of all the mesoscopic chunks of the gas: with probability close to 1, they approximately all yield approximately the same temperature estimate.

Due to chaos, we also expect that the low-level state of mesoscopic chunks which are not too close together spatially are approximately independent given the temperature.

So, we have a system in which the low-level states of lots of different mesoscopic chunks are approximately independent given the value of an approximate deterministic constraint (temperature) between them. Intuitively, those are the conditions under which we expect to find a nontrivial natural latent. In this case, we expect the natural latent to be approximately (isomorphic to) temperature.

2. Biased Die

Consider 1000 rolls of a die of unknown bias. Any 999 of the rolls will yield approximately the same estimate of the bias. That's (approximately) the redundancy condition for the bias.

We also expect that the 1000 rolls are independent given the bias. That's the mediation condition. So, we expect the bias is an approximate natural latent over the rolls.

However, the approximation error bound in this case is quite poor, since our proven error bound scales with the number n of observables. We can easily do better by viewing the first 500 and second 500 rolls as two observables. We expect that the first 500 rolls and the second 500 rolls will yield approximately the same estimate of the bias, and that the first 500 and second 500 rolls are independent given the bias, so the bias is a natural latent between the first and second 500 rolls of the die. This view of the problem will likely yield much better error bounds. More generally, chunking together many observables this way typically provides much better error bounds than applying the theorems directly to many observables.

3. Timescale Separation In A Markov Chain

In a Markov Chain, timescale separation occurs when there is some timescale T such that, if the chain is run for T steps, then the state can be split into a component which is almost-certainly conserved over the T steps and a component which is approximately ergodic over the T steps. In that case, we expect both the initial state and T^{th} state to almost-certainly yield the same estimate of the conserved component, and we expect that the initial state and T^{th} state are approximately independent given the conserved component, so the conserved component should be an approximate natural latent between the initial and T^{th} state.

V. DISCUSSION & CONCLUSION

We began by asking when one agent's latent can be *guaranteed* to be expressible in terms of another agent's latent(s), given that the two agree on predictions about observables. We've shown that:

- The natural latent conditions are necessary for such a guarantee.
- The natural latent conditions ensure that such a guarantee can be achieved via a simple Construction.
- Both of the above are robust to approximation.

... for a specific broad class of possibilities for the other agent's latent(s). In particular, the other agent can use any latent(s) which fully explain the interactions between observables. So long as the other agent's latent(s) are in that class, and the first agent uses a natural latent constructed appropriately, the first agent's latent can be expressed in terms of the second for purposes of modeling the observables. Furthermore, for this particular class of other agent's latents, a natural latent is the *only* way to achieve such a guarantee.

These results provide a potentially powerful tool for many of the questions posed at the beginning.

When is robust translation possible at all, between agents with potentially different internal concepts, like e.g. humans and AI, or humans from different cultures? Insofar as the agents make the same predictions about some parts of the world, and both their latent concepts induce independence between those parts of the world, either agent can ensure robust translatability into the other agent's ontology by using a natural latent. In particular, if the agents are trying to communicate, they can look for parts of the world over which natural latents exist, and use words to denote those natural latents; the equivalence of natural latents will ensure translatability in principle, though the agents still need to do the hard work of figuring out which words refer to natural latents over which parts of the world.

Under what conditions are scientific concepts guaranteed to carry over to the ontologies of new theories, like how e.g. general relativity has to reduce to Newtonian gravity in the appropriate limit? Insofar as the old theory correctly predicted at least some parts of the world, and the new theory introduces latents to explain all the interactions between those parts of the world, the old theorist can guarantee forward-compatibility by working with natural latents over the relevant parts of the world. This allows scientists a potential way to check that their work is likely to carry forward into as-yet-unknown future paradigms.

When and why might a wide variety of minds in the same environment converge to use (approximately) the same concept internally? While this question wasn't the main focus of this paper, both the minimality and maximality conditions suggest that natural latents (when they exist) will often be convergently used by a variety of optimized systems. For minimality: the natural latent is the minimal variable which mediates between observables, so we should intuitively expect that systems which need to predict some observables from others and are bandwidth-limited somewhere in that process will often tend to represent natural latents as intermediates. For maximality: the natural latent is the maximal variable which is redundantly represented, so we should intuitively expect that systems which need to reason in ways robust to individual inputs will often tend to track natural latents.

The natural latent conditions are a first step toward all these threads. Most importantly, they offer any mathematical foothold at all on such conceptually-fraught problems. We hope that foothold will both provide a foundation for others to build upon in tackling such challenges both theoretically and empirically, and inspire others to find their own footholds, having seen that it can be done at all.

ACKNOWLEDGMENTS

We thank the Long Term Future Fund for funding this work.

Appendix A: Graphical Notation and Some Rules for Graphical Proofs

In this paper, we use the diagrammatic notation of Bayesian networks (Bayes nets) to concisely state properties of probability distributions. Unlike the typical use of Bayes nets, where the diagrams are used to define a distribution, we assume that the joint distribution is given and use the diagrams to express properties of the distribution. Specifically, we say that a distribution $P[Y]$ "satisfies" a Bayes net diagram if and only if the distribution factorizes according to the diagram's structure. In the case of approximation, we say that $P[Y]$ "approximately satisfies" the diagram, up to some $\epsilon \geq 0$, if and only if the Kullback-Leibler divergence (D_{KL}) between the true distribution and the distribution implied by the diagram is less than or equal to ϵ.

1. Frankenstein Rule

a. *Statement*

Let $P[X_1, \ldots, X_n]$ be a probability distribution that satisfies two different Bayesian networks, represented by directed acyclic graphs G_1 and G_2. If there exists an ordering of the variables $X_1, \ldots, X_n$ that respects the topological order of both G_1 and G_2 simultaneously, then $P[X_1, \ldots, X_n]$ also satisfies any "Frankenstein" Bayesian network constructed by taking the incoming edges of each variable X_i from either G_1 or G_2. More generally, if $P[X_1, \ldots, X_n]$ satisfies m different Bayesian networks $G_1, \ldots, G_m$, and there exists an ordering of the variables that respects the topological order of all m networks simultaneously, then $P[X_1, \ldots, X_n]$ satisfies any "Frankenstein" Bayesian network constructed by taking the incoming edges of each variable X_i from any of the m original networks.

We'll prove the approximate version, then the exact version follows trivially.

b. *Proof*

Without loss of generality, assume the order of variables respected by all original diagrams is $X_1, \ldots, X_n$. Let $P[X] = \prod_i P[X_i|X_{pa_j(i)}]$ be the factorization expressed by diagram j, and let $\sigma(i)$ be the diagram from which the parents of X_i are taken to form the Frankenstein diagram. (The factorization expressed by the Frankenstein diagram is then $P[X] = \prod_i P[X_i|X_{pa_{\sigma(i)}(i)}]$.)

The proof starts by applying the chain rule to the D_{KL} of the Frankenstein diagram:

$$\begin{aligned} D_{KL}\left(P[X]||\prod_i P[X_i|X_{pa_{\sigma(i)}(i)}]\right) &= D_{KL}\left(\prod_i P[X_i|X_{<i}]||\prod_i P[X_i|X_{pa_{\sigma(i)}(i)}]\right) \\ &= \sum_i \mathbb{E}\left[D_{KL}\left(P[X_i|X_{<i}]||P[X_i|X_{pa_{\sigma(i)}(i)}]\right)\right] \end{aligned}$$

Then, we add a few more expected KL-divergences (i.e., add some non-negative numbers) to get:

$$\begin{aligned} &\leq \sum_i \sum_j \mathbb{E}\left[D_{KL}\left(P[X_i|X_{<i}]||P[X_i|X_{pa_j(i)}]\right)\right] \\ &= \sum_j D_{KL}\left(P[X]||\prod_i P[X_i|X_{pa_j(i)}]\right) \\ &\leq \sum_j \epsilon_j \end{aligned}$$

Thus, we have

$$\begin{aligned} D_{KL}\left(P[X]||\prod_i P[X_i|X_{pa_{\sigma(i)}(i)}]\right) &\leq \sum_j D_{KL}\left(P[X]||\prod_i P[X_i|X_{pa_j(i)}]\right) \\ &\leq \sum_j \epsilon_j \end{aligned}$$

2. Stitching Rule

a. *Statement*

In the exact case, the Stitching Rule says:
Let $P[X, Y, Z]$ be a probability distribution, and suppose that:

1. $P[X, Y]$ satisfies a Bayesian network G_{XY} over variables X and Y.
2. $P[Z, Y]$ satisfies a Bayesian network G_{ZY} over variables Z and Y.
3. Y is a Markov blanket between X and Z in $P[X, Y, Z]$, i.e., $X \perp Z|Y$.
4. Each variable $Y_i \in Y$ is a child of variables in X in G_{XY}, or a child of variables in Z in G_{ZY}, but not both.
5. There exists an ordering of all variables in X, Y, and Z that respects the topological order of both G_{XY} and G_{ZY} simultaneously.

Then, $P[X, Y, Z]$ satisfies a "stitched" Bayesian network G_{XYZ} constructed as follows:

- Each variable $X_i \in X$ takes its parents from G_{XY}.
- Each variable $Z_i \in Z$ takes its parents from G_{ZY}.
- Each variable $Y_i \in Y$ with a parent in X takes its parents from G_{XY}.
- Each variable $Y_i \in Y$ with a parent in Z takes its parents from G_{ZY}.

In the approximate case, we have:
Let $P[X, Y, Z]$ be a probability distribution that approximately satisfies the conditions of the Stitching Rule for a Markov Blanket, with:

- G_{XY} satisfied up to ϵ_{XY}.
- G_{ZY} satisfied up to ϵ_{ZY}.
- Y being a Markov blanket between X and Z up to $\epsilon_{blanket}$.

Then, the "stitched" Bayesian network G_{XYZ} constructed according to the Stitching Rule is satisfied by $P[X, Y, Z]$ up to $\epsilon_{XY} + \epsilon_{ZY} + \epsilon_{blanket}$. Furthermore, if we have fine-grained bounds on the D_{KL} for individual variables in each diagram, we can obtain a tighter bound on the D_{KL} of the "stitched" Bayesian network, similar to the more general Approximate Frankenstein Rule.

b. *Proof*

We begin the proof with the $X \leftarrow Y \rightarrow Z$ condition:

$$\begin{aligned}\epsilon_{\text{blanket}} &\geq D_{KL}(P[X, Y, Z]||P[X|Y]P[Y]P[Z|Y]) \\ &= D_{KL}(P[X, Y, Z]||P[X, Y]P[Z, Y]/P[Y])\end{aligned}$$

At a cost of at most ϵ_{XY}, we can replace $P[X, Y]$ with $\prod_i P[(X, Y)_i|(X, Y)_{pa_{XY}(i)}]$ in that expression, and likewise for the $P[Z, Y]$ term. (You can verify this by writing out the D_{KL}'s as expected log probabilities.)

$$\epsilon_{\text{blanket}} + \epsilon_{XY} + \epsilon_{ZY} \geq D_{KL}\left(P[X, Y, Z]||\frac{\prod_i P[(X, Y)_i|(X, Y)_{pa_{XY}(i)}]\prod_i P[(Z, Y)_i|(Z, Y)_{pa_{ZY}(i)}]}{P[Y]}\right)$$

Notation:

- Y_{XY} denotes the components of Y whose parents are taken from the XY-diagram; Y_{ZY} denotes the components of Y whose parents are taken from the ZY-diagram. (Together, these should include all components of Y.)

- All products are implicitly over components of whatever variables they're indexing - e.g., $\prod_i P[Y_{XY_i}|(X,Y)_{pa_{XY}(i)}]$ (which will appear shortly) is over components of Y_{XY}.
- $(X,Y)_{pa_{XY}(i)}$ denotes the parents of i in the XY-diagram. Each such parent will be a component of X or Y, which is why we're subscripting the pair (X,Y). Likewise for similar expressions.

Recall that each component of Y_{ZY} must have no X-parents in the XY-diagram, and each component of Y_{XY} must have no Z-parents in the ZY-diagram. Let's pull those terms out of the products above so we can simplify them:

$$\epsilon_{\text{blanket}} + \epsilon_{XY} + \epsilon_{ZY} \geq D_{KL}\left(P[X,Y,Z]||\frac{\prod_i P[(X,Y_{XY})_i|(X,Y)_{pa_{XY}(i)}]\prod_i P[Y_{ZY_i}|Y_{pa_{XY}(i)}]}{\prod_i P[(Z,Y_{ZY})_i|(Z,Y)_{pa_{ZY}(i)}]\prod_i P[Y_{XY_i}|Y_{pa_{ZY}(i)}]/P[Y]}\right)$$

Those simplified terms in combination with $1/P[Y]$ are themselves a D_{KL}, which we can separate out:

$$\begin{aligned}
&= D_{KL}(P[X,Y,Z]||\prod_i P[(X,Y_{XY})_i|(X,Y)_{pa_{XY}(i)}]\prod_i P[(Z,Y_{ZY})_i|(Z,Y)_{pa_{ZY}(i)}]) \\
&\quad + D_{KL}(P[Y]||\prod_i P[Y_{XY_i}|Y_{pa_{ZY}(i)}]\prod_i P[Y_{ZY_i}|Y_{pa_{XY}(i)}]) \\
&\geq D_{KL}(P[X,Y,Z]||\prod_i P[(X,Y_{XY})_i|(X,Y)_{pa_{XY}(i)}]\prod_i P[(Z,Y_{ZY})_i|(Z,Y)_{pa_{ZY}(i)}])
\end{aligned}$$

The last line is the D_{KL} for the stitched diagram.

3. Factorization Transfer

a. *Statement*

Let $P[X_1,\ldots,X_n]$ and $Q[X_1,\ldots,X_n]$ be two probability distributions over the same set of variables. If Q satisfies a given factorization (represented by a diagram) and Q approximates P with an error of at most ϵ, i.e.,

$$\epsilon \geq D_{KL}(P||Q),$$

then P also approximately satisfies the same factorization, with an error of at most ϵ:

$$\epsilon \geq D_{KL}\left(P[X_1,\ldots,X_n]||\prod_i P[X_i|X_{pa(i)}]\right),$$

where $X_{pa(i)}$ denotes the parents of X_i in the diagram representing the factorization.

b. *Proof*

As with the Frankenstein rule, we start by splitting our D_{KL} into a term for each variable:

$$D_{KL}(P[X]||Q[X]) = \sum_i \mathbb{E}\left[D_{KL}(P[X_i|X_{<i}]||Q[X_i|X_{pa(i)}])\right]$$

Next, we subtract some more D_{KL}'s (i.e., subtract some non-negative numbers) to get:

$$\begin{aligned}
&\geq \sum_i \left(\mathbb{E}\left[D_{KL}(P[X_i|X_{<i}]||Q[X_i|X_{pa(i)}])\right] - \mathbb{E}\left[D_{KL}(P[X_i|X_{pa(i)}]||Q[X_i|X_{pa(i)}])\right]\right) \\
&= \sum_i \mathbb{E}\left[D_{KL}(P[X_i|X_{<i}]||P[X_i|X_{pa(i)}])\right] \\
&= D_{KL}\left(P[X]||\prod_i P[X_i|X_{pa(i)}]\right)
\end{aligned}$$

Thus, we have

$$D_{KL}(P[X]||Q[X]) \geq D_{KL}\left(P[X]||\prod_i P[X_i|X_{pa(i)}]\right)$$

4. Bookkeeping Rule(s)

a. *General Statement*

If all distributions which exactly factor over Bayes net G_1 also exactly factor over Bayes net G_2, then:

$$D_{\mathrm{KL}}\left(P[X]\middle\|\prod_i P(X_i \mid X_{\mathrm{pa}_{G_1}(i)})\right) \geq D_{\mathrm{KL}}\left(P[X]\middle\|\prod_i P(X_i \mid X_{\mathrm{pa}_{G_2}(i)})\right)$$

b. *Proof*

Let $Q[X] := \prod_i P(X_i \mid X_{\mathrm{pa}_{G_1}(i)})$. By definition, Q factors over G_1. Since all distributions which factor over G_1 also factor over G_2, it follows that Q also factors over G_2.

Now, we have:

$$Q[X] = \prod_i Q(X_i \mid X_{\mathrm{pa}_{G_2}(i)})$$

Thus:

$$D_{\mathrm{KL}}\left(P[X]\middle\|\prod_i P(X_i \mid X_{\mathrm{pa}_{G_1}(i)})\right) = D_{\mathrm{KL}}\left(P[X]\middle\|\prod_i Q(X_i \mid X_{\mathrm{pa}_{G_2}(i)})\right)$$

By the Factorization Transfer Theorem, we have:

$$D_{\mathrm{KL}}\left(P[X]\middle\|\prod_i Q(X_i \mid X_{\mathrm{pa}_{G_2}(i)})\right) \geq D_{\mathrm{KL}}\left(P[X]\middle\|\prod_i P(X_i \mid X_{\mathrm{pa}_{G_2}(i)})\right)$$

Which completes the proof.

Appendix B: Graphical Proofs

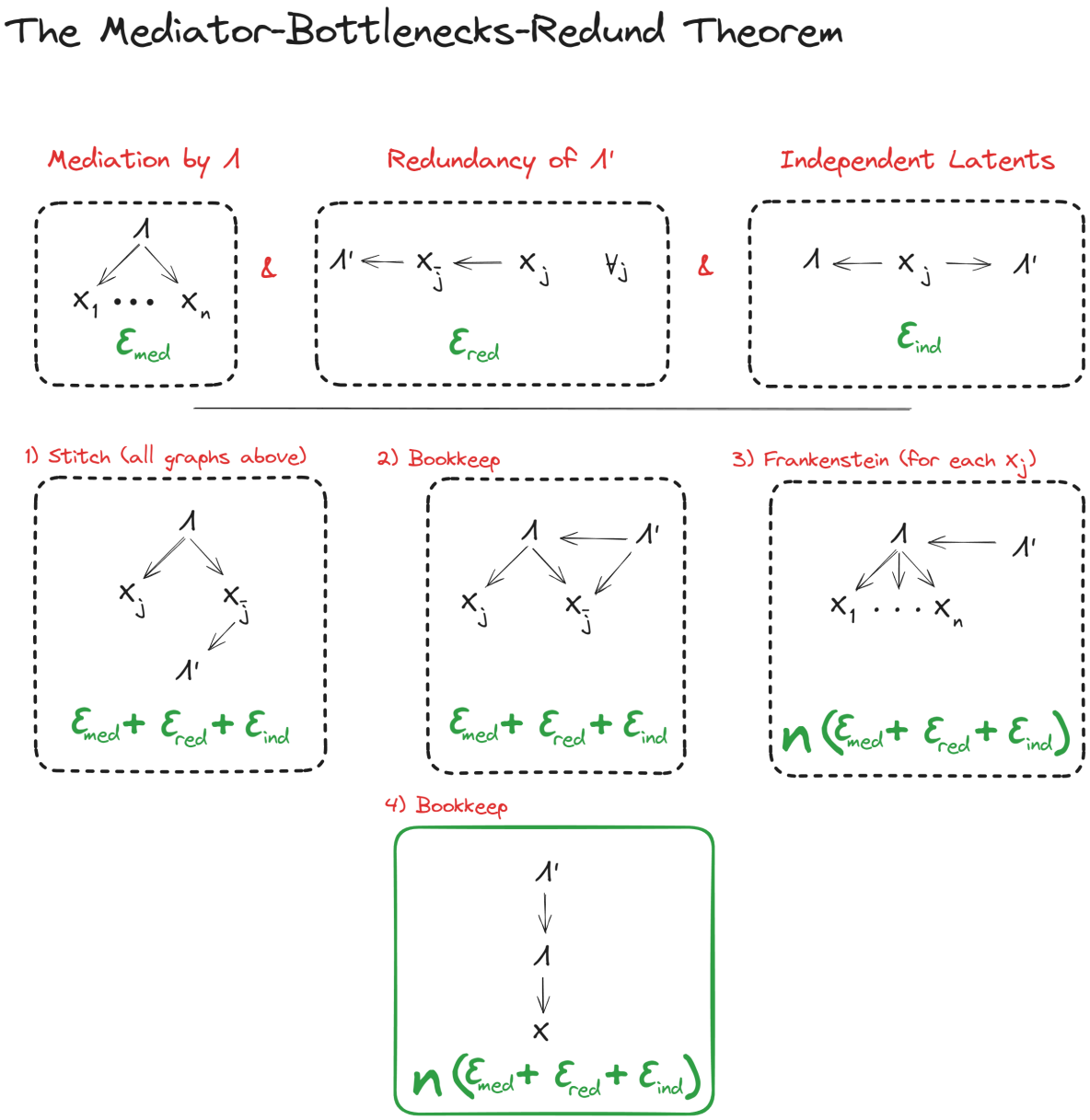

FIG. 9. Proof of the Mediator Bottlenecks Redund Theorem

Appendix C: Python Script for Computing D_{KL} in Worked Example

```
import numpy as np
from scipy.special import gammaln, logsumexp, xlogy

n = 1000
p_N2 = np.ones(n+1)/(n+1)
N1 = np.outer(np.arange(n + 1), np.ones(n + 1))
N2 = np.outer(np.ones(n + 1), np.arange(n + 1))
# logP[N1|N2]; we're tracking log probs for numerical stability
lp_N1_N2 = (gammaln(n + 2) - gammaln(N2 + 1) - gammaln(n - N2 + 1) +
            gammaln(n + 1) - gammaln(N1 + 1) - gammaln(n - N1 + 1) +
```

```
            gammaln(N1 + N2 + 1) + gammaln(2*n - N1 - N2 + 1) - gammaln(2*n + 2))

# logP[\Lambda' = 0|N2] and logP[\Lambda' = 1|N2]
lp_lam0_N2 = logsumexp(lp_N1_N2[:500], axis=0)
lp_lam1_N2 = logsumexp(lp_N1_N2[500:], axis=0)

p_lam0_N2 = np.exp(lp_lam0_N2)
p_lam1_N2 = np.exp(lp_lam1_N2)

print(p_lam0_N2 + p_lam1_N2)  # Check: these should all be 1.0

# ... aaaand then it's just the ol' -p * logp to get the expected entropy E[H(\Lambda')|N2]
H = - np.sum(p_lam0_N2 * lp_lam0_N2 * p_N2) - np.sum(p_lam1_N2 * lp_lam1_N2 * p_N2)
print(H / np.log(2))  # Convert to bits
```

---

[1] W. V. O. Quine, On empirically equivalent systems of the world, Erkenntnis **9**, 313 (1975).

[2] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey, Sparse autoencoders find highly interpretable features in language models, (2023), arXiv:2309.08600 [cs.LG].

[3] Marvik, Model merging: Combining different fine-tuned LLMs, Blog post (2024), retrieved from `https://marvik.com/model-merging`.

[4] M. Huh, B. Cheung, T. Wang, and P. Isola, The platonic representation hypothesis, (2024), arXiv:2405.07987 [cs.LG].

[5] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. (Cambridge University Press, USA, 2009).