# Contextualized Explanations via Bayesian Model-Extensions

David Reber[1]

[1]*University of Chicago*

**Abstract**

Your abstract here

## 1 Introduction

Suppose we have a model of reality's true data-generating process $M_R$. This could be a model that is only valid in the regime of the training data, and on a particular task: it doesn't need to be comprehensive of reality everywhere.

When we train a generative model to optimality over this context/task, it learns a model $M_L$ which is *compatible* with some constraints induced by $M_R$.

**Definition 1.** Two models $M$ and $M'$ are said to be *$C$-compatible* if they both respect the constraints in $C$. In this case we write $M \sim^C M'$.

**Definition 2.** Constraints can take many forms. For instance:

- Conditional independencies
- symmetries / conservation laws
- bounds on states (nonnegativity of quantities, etc)
- conditional independencies of interventional distributions (if the training data contained a diverse set of environments, then the model should be invariant to interventions on the environment variables)

**Example 3. Bayes Nets** Suppose $M_R$ consists of observable variables $V$ such that $P(V)$ factorizes according to a Bayes net $G$. Then $C$ will contain the conditional independeces induced by $G$. If $M_L$ is trained to optimality, it will inherit these conditional independencies.

However, constraints aren't enough to fully specify $M_L$, so there's a lot of freedom for $M_L$ to deviate from $M_R$. The question is, given two explanations $E_A$ and $E_B$ for the output behavior of a $M_L$, how do we decide which is more compatible with the learned model $M_L$?

Let's associate our informal (natural language, etc) explanations $E_A$ and $E_B$ with formal models $M_A \sim^C M_R$ and $M_B \sim^C M_R$ and ask whether based on observed quantities, $M_L$ is more likely to be $M_A$ or $M_B$.

$M_A$ and $M_B$ are two guesses for $M_L$, and based on a limited observation function $O(M)$ we want to assess a posterior estimate about whether $M_A$ or $M_B$ is the right guess for $M_L$, given that $M_L$, $M_A$, and $M_B$ are all $C$-compatible with $M_R$.

**Definition 4.** Let $O(M)$ be a function that takes a model $M$ and returns a set of observations. Practically, $O(M)$ is whatever information about $M_L$ we can reliably extract using our interpretability tools.

The key is that if we can only infer the values of a few concepts, then we can only use those concepts to distinguish between $M_A$ and $M_B$.

Assume $M_A$ and $M_B$ are both equally likely to be the true model $M_L$. Then we can compute the posterior probability that $M_L$ is $M_A$ given the observations $O(M_L) = x$, by using the likelihoods $P(O(M_L) = x|M_L = M_A)$ and $P(O(M_L) = x|M_L = M_B)$.

If $O(M)$ simply reports the values of a few concepts, then these reduce to $P_A(x)$., that is, $P(x)$ in the model $M_A$.

The more informative $O(M)$, the faster the convergence to $M_A$ or $M_B$; or on the flip side, if $O(M)$ is not informative enough, it may be possible to prove that $M_A$ and $M_B$ can't be distinguished.