A Bayesian Framework for Information-Theoretic Probing

Tiago Pimentel⚽
⚽University of Cambridge
tp472@cam.ac.uk

Ryan Cotterell⚽🏈
🏈ETH Zürich
ryan.cotterell@inf.ethz.ch

Abstract

Pimentel et al. (2020b) recently analysed probing from an information-theoretic perspective. They argue that probing should be seen as approximating a mutual information. This led to the rather unintuitive conclusion that representations encode exactly the same information about a target task as the original sentences. The mutual information, however, assumes the true probability distribution of a pair of random variables is known, leading to unintuitive results in settings where it is not. This paper proposes a new framework to measure what we term **Bayesian mutual information**, which analyses information from the perspective of Bayesian agents—allowing for more intuitive findings in scenarios with finite data. For instance, under Bayesian MI we have that data can add information, processing can help, and information can hurt, which makes it more intuitive for machine learning applications. Finally, we apply our framework to probing where we believe Bayesian mutual information naturally operationalises *ease of extraction* by explicitly limiting the available background knowledge to solve a task.

1 Introduction

Pimentel et al. (2020b) recently undertook an information-theoretic analysis of probing. They argue that probing may be viewed as approximating the mutual information between a linguistic property (e.g., part-of-speech tags) and a contextual representation (e.g., BERT). Counter-intuitively, however, due to the data-processing inequality, contextual representations contain exactly the same information about any task as the original sentence, under mild conditions. When viewed under this lens, the goal of probing is not inherently clear. One limitation of Pimentel et al.'s analysis is that it focuses on the **mutual information** (MI)—to be of practical application, their argument requires that a probe matches the *true distribution* according to which the data were generated in the limit of *finite training data*. In contrast, our paper formulates an information-theoretic framework that is compatible both with model misspecification and the finite data assumption.

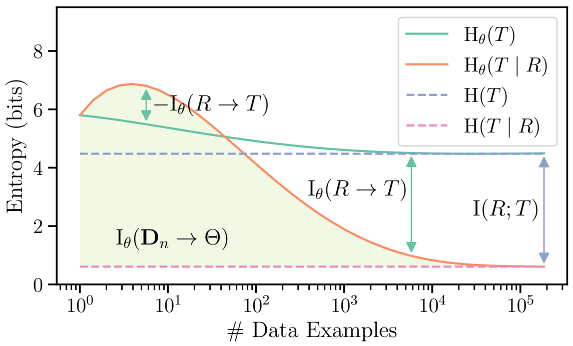

Figure 1: A smoothed example of Bayesian MI on dependency arc labelling in English.

In his seminal work, Shannon (1948) occupied himself with the *limit* of communication. Indeed, mutual information can be described as the theoretical limit (or upper-bound) of how much information can be extracted from one random variable about another. However, this limit is only achievable when one has full knowledge of these random variables, including the true probability distribution according to which they are distributed. In practice, we will not have access to such information and it may be difficult to approximate. It follows that any system with imperfect knowledge of the random variable's true distribution will only be able to extract a subset of this information.

With this in mind, we propose and motivate an agent-based framework for measuring information. We term our quantity **Bayesian mutual information** and show that it generalises Shannon's MI, holistically accounting for uncertainty within the Bayesian paradigm—measuring the amount of information a rational agent could extract from a random variable *under partial knowledge of the true distribution*. In addition to the definition,

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2869–2887
November 7–11, 2021. ©2021 Association for Computational Linguistics

our paper provides many useful theoretical results. For instance, we prove that conditioning does *not* necessarily reduce Bayesian entropy and that Bayesian mutual information does *not* obey the data-processing inequality. We argue that these properties make our Bayesian framework ideal for an analysis of learned representations.

In the empirical portion of our paper, we investigate both part-of-speech tagging and dependency arc labelling. Moreover, our information-theoretic measure holistically captures a notion of ease of extraction, limiting the amount of data available to solve the task. Intuitively, Bayesian MI shows that high dimensional representations, such as BERT, actually hurt performance in the very low-resource scenario, making less information available to a Bayesian agent than a simple categorical distribution. This is because, when little data is available, these agents overfit to the evidence under weak priors. In the high-resource scenario of English, ALBERT dominates the curves, making more information available than other contextualised embedders. In short, Bayesian mutual information reconciles the probing literature with its frequently posed question: *how much information can be extracted from these representations?*

2 Background: Information Theory

Information theory (Shannon, 1948; Cover and Thomas, 2006) provides us with a number of tools to analyse data and their associated probability distributions—among which are the entropy and mutual information. These are traditionally defined according to a "true" probability distribution, i.e. $p(x)$ or $p(x, y)$[1] which may not be known, but dictates the behaviour of random variables X and Y. The atomic unit of information theory is the **surprisal**, which is defined as follows:

$$\mathrm{H}(X = x) = -\log p(x) \tag{1}$$

Arguably, the most important information-theoretic definition is its expected value, termed **entropy**:

$$\mathrm{H}(X) \overset{\text{def}}{=} -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2}$$

Finally, another important concept is the **mutual information** (MI) between two random variables

$$\mathrm{I}(X; Y) \overset{\text{def}}{=} \mathrm{H}(X) - \mathrm{H}(X \mid Y) \tag{3}$$

Unfortunately, information theory has a few properties which do not conform to our intuitions about the mechanics of information in machine learning:

(i) **Data Does Not Add Information:** The entropy is defined according to a source distribution $p(x)$. So, if multiple instances of X are sampled i.i.d. from $p(x)$, access to a set $\mathbf{d}_N = \left\{x^{(1)}, \ldots, x^{(N)}\right\}$ of such instances cannot provide any information, i.e. $\mathrm{H}(X \mid \mathbf{D}_N = \mathbf{d}_N) = \mathrm{H}(X)$.

(ii) **Conditioning Reduces Entropy:** Another basic result from information theory is that conditioning cannot increase entropy, only reduce it, i.e. $\mathrm{H}(X \mid Y) \leq \mathrm{H}(X)$. This implies datapoints can never be misleading, which is not true in practice.

(iii) **Data Processing Does Not Help:** The data processing inequality states that processing some random variable with a function $f(\cdot)$ can never increase how informative it is, but only reduce its information content, i.e. $\mathrm{I}(X; f(Y)) \leq \mathrm{I}(X; Y)$.

2.1 Background: Belief Entropy

A related question that arises is how to estimate information in scenarios where the true distribution is not known. For instance, what is the surprisal of a learning agent with a **belief** $p_{\boldsymbol{\theta}}(x)$ who encounters an instance x? The straightforward answer would be to use eq. (1)—nonetheless, this agent does not know the true distribution $p(x)$. This agent's surprisal is usually taken according to its belief:

$$\mathrm{H}_b(X = x) = -\log p_{\boldsymbol{\theta}}(x) \tag{4}$$

Similarly, this agent's entropy has been historically defined exclusively according to this belief distribution (Gallistel and King, 2011):

$$\mathrm{H}_b(X) \overset{\text{def}}{=} -\sum_{x \in \mathcal{X}} p_{\boldsymbol{\theta}}(x) \log p_{\boldsymbol{\theta}}(x) \tag{5}$$

We term this the **belief-entropy**. We can further extend this to a **belief mutual information**:[2]

$$\mathrm{I}_b(X; Y) \overset{\text{def}}{=} \mathrm{H}_b(X) - \mathrm{H}_b(X \mid Y) \tag{6}$$

We note this definition is not grounded in the true distribution in any form. In fact, about the belief

[1] We use uppercase letters to denote random variables (X, Y), lowercase for their instances (x, y), and calligraphic fonts for their domain space ($x \in \mathcal{X}$, $y \in \mathcal{Y}$).

[2] This definition is found in both the cognitive sciences (Gallistel and King, 2011; Fan, 2014; Sayood, 2018) as well as in active learning (Houlsby et al., 2011; Kirsch et al., 2019).

mutual information, Gallistel and King (2011) state: "the subjectivity that it implies is deeply unsettling [...] the amount of information actually communicated is not an objective function of the signal from which the subject obtained it".

3 A Bayesian Approach to Information

The primary motivation for this paper is developing a series of tools that help us overcome the limitations of traditional information theory as applied to machine learning. Specifically, probing representations requires a data-dependent information theory. We thus formulate analogues of surprisal, entropy and MI in terms of Bayesian agents—using a framework heavily inspired by Bayesian experimental design (Lindley, 1956). We then prove this framework does not suffer the same infelicities as standard information theory in this context.

3.1 An Agent-Based Information Theory

Our discussions will focus on Bayesian agents, so we start by formally defining them.

Definition 1. *A **Bayesian agent** is a parameterised probability distribution $p_{\boldsymbol{\theta}}(x \mid \boldsymbol{\theta})$ (or set of distributions) and a prior $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.*[3] *Given data $\mathbf{d}_N = \{x^{(1)}, \ldots, x^{(N)}\}$, the **Bayesian posterior** over $\boldsymbol{\theta}$ is*

$$p_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid \mathbf{d}_N) \propto \prod_{n=1}^{N} p_{\boldsymbol{\theta}}(x^{(n)} \mid \boldsymbol{\theta})\, p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{7}$$

*Analogously, the **Bayesian belief** is defined as the following posterior predictive distribution*

$$p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N) = \int p_{\boldsymbol{\theta}}(x \mid \boldsymbol{\theta})\, p_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid \mathbf{d}_N)\, \mathrm{d}\boldsymbol{\theta} \tag{8}$$

Upon encountering an instance x, and after seeing a collection of data $\mathbf{d}_N$, this agent's **posterior-predictive Bayesian surprisal** will be:

$$\mathrm{H}_{\boldsymbol{\theta}}(X = x \mid \mathbf{D}_N = \mathbf{d}_N) = -\log p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N) \tag{9}$$

where $\mathbf{D}_N$ is a data-valued random variable; for notational succinctness, we omit this random variable for the rest of the paper. We further define the **posterior-predictive Bayesian entropy**:

$$\mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) = -\sum_{x \in \mathcal{X}} p(x) \log p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N) \tag{10}$$

As can be readily seen, the Bayesian entropy is the expected value of the Bayesian surprisal—with this expectation taken over the true distribution.[4] In this sense, the Bayesian entropy is a cross-entropy rather than a standard entropy.

3.2 Bayesian Mutual Information

Defining Bayesian mutual information within our framework requires a bit more care. First, in contrast to surprisal and entropy, mutual information is a functional of two random variables. We will name the second random variable Y. To talk about mutual information, we will consider a Bayesian agent with a collection of at least two beliefs, e.g. $\{p_{\boldsymbol{\theta}}(x), p_{\boldsymbol{\theta}}(x \mid y)\}$. The second belief is conditional, but otherwise follows Definition 1.

Definition 2. *Given a collection of data $\mathbf{d}_N = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$, and a Bayesian agent with a pair of beliefs $p_{\boldsymbol{\theta}}(x \mid \boldsymbol{\theta})$ and $p_{\boldsymbol{\theta}}(x \mid y, \boldsymbol{\theta})$ and a prior $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, the **Bayesian mutual information** (Bayesian MI) is defined as*

$$\begin{aligned} &\mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) \overset{\text{def}}{=} \\ &\qquad \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}_{\boldsymbol{\theta}}(X \mid Y, \mathbf{d}_N) \end{aligned} \tag{11}$$

There is an important distinction between the Bayesian and Shannon MI—Bayesian MI decomposes as the difference between two cross-entropies, as opposed to two entropies.[5]

3.3 An Illustrative Example

For the sake of argument, we assume two independent categorical random variables X and Y, both with c classes and uniformly distributed.

$$p(x) = \frac{1}{c}, \quad p(x \mid y) = \frac{1}{c} \tag{12}$$

We further assume a Bayesian agent with two categorical beliefs $\{p_{\boldsymbol{\theta}}(x) = \mathrm{Cat}(\boldsymbol{\theta}), p_{\boldsymbol{\theta}}(x \mid y) = \mathrm{Cat}(\boldsymbol{\theta} + y)\}$—where y is assumed to be encoded as a one hot vector—and a Dirichlet prior

[3] In the case that the Bayesian agent has more than one distribution, we still only have a single prior without loss of generality. Indeed, separate priors for each distribution is a special case where the parameters are partitioned. In the case of $\boldsymbol{\theta} = [\boldsymbol{\phi}; \boldsymbol{\psi}]$, we could define $p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(\boldsymbol{\phi}) \cdot p_{\boldsymbol{\theta}}(\boldsymbol{\psi})$.

[4] We put this in contrast to eq. (5)—which takes this expectation over the belief itself—since the instances are in practice encountered with this true frequency. This distinction has been explicitly noted before, by e.g. Bartlett (1953).

[5] Cross mutual information (XMI) has been used in several previous work such as (Pimentel et al., 2019, 2020b; Bugliarello et al., 2020; McAllester and Stratos, 2020; Torroba Hennigen et al., 2020; Fernandes et al., 2021; O'Connor and Andreas, 2021). In those works, though, it was usually interpreted as a computational approximation to the truth-MI (or to $\mathcal{V}$-information (Xu et al., 2020), which is discussed later in the paper). In this work, we highlight the Bayesian MI's (and XMI's) relevance as a generalisation of Shannon's MI.

$p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathrm{Dir}(\boldsymbol{\alpha})$ with concentration parameters $\boldsymbol{\alpha} = \mathbf{1}$. Note that this (biased) agent believes that y and x are more likely than chance to share a class. Given no data, or given $\mathbf{d}_0$, this agent's prior predictive distributions are:

$$p_{\boldsymbol{\theta}}(x) = \frac{1}{c}, \;\; p_{\boldsymbol{\theta}}(x \mid y) = \begin{cases} \frac{1}{c+1}, & x \neq y \\ \frac{2}{c+1}, & x = y \end{cases} \tag{13}$$

In this example:

(i) **Mutual Information.** We have $\mathrm{I}(X;Y) = 0$ because X and Y are independent by construction.

(ii) **Belief Mutual Information.** The belief-MI is positive, since the agent's uncertainty about X is reduced by knowledge of Y—the prior predictive $p_{\boldsymbol{\theta}}(x)$ is uniform, while the conditional distribution $p_{\boldsymbol{\theta}}(x \mid y)$ is not, which reduces the belief-entropy. This means that $\mathrm{H}_b(X) > \mathrm{H}_b(X \mid Y)$, implying that $\mathrm{I}_b(X;Y) > 0$.

(iii) **Bayesian Mutual Information.** Finally, the Bayesian MI is negative—since x is uniformly distributed, the unconditional Bayesian entropy is tight, i.e. $\mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_0) = \mathrm{H}(X)$, but the conditional one is not, i.e. $\mathrm{H}_{\boldsymbol{\theta}}(X \mid Y, \mathbf{d}_0) > \mathrm{H}(X \mid Y) = \mathrm{H}(X)$. We thus have $\mathrm{I}_{\boldsymbol{\theta}}(X;Y) < 0$. This entails that, on this specific example, an agent's predictive power over X is lower when given Y.

This illustrates an important aspect of Bayesian MI: it is grounded on the true distribution.

3.4 Theoretical Properties

We now prove a few relevant theoretical properties about our framework. We show that Bayesian MI is symmetric if and only if the agent's beliefs respect Bayes' rule. Then, we discuss why it does not respect the data-processing inequality, and its connection to mutual information and to $\mathcal{V}$-information (Xu et al., 2020).

3.4.1 When is Bayesian MI Symmetric?

It is a well known result that Shannon's MI is symmetric, i.e.

$$\begin{aligned} \mathrm{I}(X;Y) &= \mathrm{H}(X) - \mathrm{H}(X \mid Y) \\ &= \mathrm{H}(Y) - \mathrm{H}(Y \mid X) = \mathrm{I}(Y;X) \end{aligned} \tag{14}$$

This means that the knowledge one can extract from random variable Y about X is the same as the knowledge one can extract from X about Y. This is not true in general for Bayesian MI; as we will show, information-theoretic symmetry and Bayes' rule are tightly related. As such, we consider in this section a Bayesian agent with a set of beliefs $\{p_{\boldsymbol{\theta}}(x), p_{\boldsymbol{\theta}}(x \mid y), p_{\boldsymbol{\theta}}(y), p_{\boldsymbol{\theta}}(y \mid x)\}$. We call the agent **consistent** if it respects Bayes' rule, i.e.

$$p_{\boldsymbol{\theta}}(x \mid y) = \frac{p_{\boldsymbol{\theta}}(y \mid x)\, p_{\boldsymbol{\theta}}(x)}{p_{\boldsymbol{\theta}}(y)} \tag{15}$$

the following theorem characterises when we have symmetry.

Theorem 1. *An agent's Bayesian mutual information is symmetric, i.e.*

$$\mathrm{I}_{\boldsymbol{\theta}}(X \rightarrow Y \mid \mathbf{d}_N) = \mathrm{I}_{\boldsymbol{\theta}}(Y \rightarrow X \mid \mathbf{d}_N) \tag{16}$$

for all distributions $p(x,y)$ if and only if the Bayesian agent is consistent.

Proof. See App. D. □

3.4.2 No Data-Processing Inequality

Another classical result from information theory is the data processing inequality. This theorem states that processing a random variable can never add information, only reduce it

$$\mathrm{I}(X;Y) \geq \mathrm{I}(X; f(Y)) \tag{17}$$

Although theoretically sound, this theorem is very unintuitive from a practical perspective—effectively, processing noisy data can make it more useful. In fact, representation learning is a subfield of machine learning devoted precisely to finding functions which can extract more "informative" representations from some input. One such example is BERT (Devlin et al., 2019), a large pre-trained language model which produces contextualised representations from sentential inputs. These representations provably contain the exact same information about any task as the original sentence (Pimentel et al., 2020b)—in practice, though, they are much more useful for downstream models.

The data processing inequality does not hold for Bayesian information, making it a more intuitive information-theoretic measure for probing; pre-trained representation extraction functions can increase MI from a Bayesian agent perspective.

Theorem 2. *The data processing inequality does not hold for Bayesian information, i.e.*

$$\mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) \mathrel{?} \mathrm{I}_{\boldsymbol{\theta}}(f(Y) \to X \mid \mathbf{d}_N) \quad (18)$$

Proof. See App. E. □

3.4.3 Relation to Mutual Information

The relationship between Bayesian mutual information and Shannon MI is relevant for our discussion. As mentioned in the introduction, Shannon was concerned with the limits of communication when he defined his measure. We now put forward an intuitive theorem about Bayesian information; it is upper-bounded by the true MI under a weak assumption about the agent's beliefs.

Theorem 3. *Assuming the agent's belief $p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)$ has a smaller Kullback–Leibler (KL) divergence when compared to the true $p(x)$ than the marginal of its beliefs over y, i.e.*

$$\mathrm{KL}\left(p(x) \mid\mid p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)\right) \leq \quad (19)$$
$$\mathrm{KL}\left(p(x) \mid\mid \sum_{y \in \mathcal{Y}} p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N)\, p(y)\right)$$

We show

$$\mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) \leq \mathrm{I}(X; Y) \quad (20)$$

Proof. See App. F. □

In other words, the information any agent can extract from a random variable Y about another variable X is upper-bounded by the true MI. We now define a well-formed belief, which we will use to analyse the Bayesian MI's convergence:

Definition 3. *We say the belief of a Bayesian agent is* ***well-formed*** *if and only if the true distribution is a possible belief, i.e.*

$$\exists\, \boldsymbol{\theta} : \quad p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) > 0 \text{ and } p(x) = p_{\boldsymbol{\theta}}(x \mid \boldsymbol{\theta}) \quad (21)$$

Given this definition, we prove the Bayesian mutual information converges to the true MI under well-defined conditions.

Theorem 4. *If we assume a Bayesian agent's set of beliefs and prior are well-formed and meet the conditions of Bernstein–von Mises Theorem (pg. 339, Bickel and Doksum, 2001).*[6] *Then,*

$$\lim_{N \to \infty} \mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) = \mathrm{I}(X; Y) \quad (22)$$

Proof. See App. G. □

3.4.4 Relation to Variational Information

Variational ($\mathcal{V}$-) information (Xu et al., 2020) is a recent generalisation of mutual information. It extends MI to the case where a fixed family of distributions is considered; in which the true distribution may or not be.

Definition 4. *Suppose random variable X is distributed according to $p(x)$. Let $\mathcal{V}$ be a variational family of distributions. Then, $\mathcal{V}$-entropy is defined*

$$\mathrm{H}_{\mathcal{V}}(X) \overset{\text{def}}{=} \inf_{q \in \mathcal{V}} - \sum_{x \in \mathcal{X}} p(x) \log q(x) \quad (23)$$

and $\mathcal{V}$-information is defined as

$$\mathrm{I}_{\mathcal{V}}(Y \to X) \overset{\text{def}}{=} \mathrm{H}_{\mathcal{V}}(X) - \mathrm{H}_{\mathcal{V}}(X \mid Y) \quad (24)$$

Unlike our Bayesian mutual information, $\mathcal{V}$-information is not a data-dependent measure, i.e. $\mathrm{H}_{\mathcal{V}}(X \mid \mathbf{D}_N) = \mathrm{H}_{\mathcal{V}}(X)$. Thus, it does not meet our desiderata. However, we can prove a straightforward relationship between the Bayesian and $\mathcal{V}$ informations, which we state below.

Theorem 5. *Assume a Bayesian agent's beliefs and prior meet the conditions of Kleijn and van der Vaart (2012), who extend the Bernstein–von Mises Theorem to beliefs which are not well-formed. Further, let $\mathcal{V} = \{p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{\theta}) \mid p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) > 0\}$. Then,*

$$\lim_{N \to \infty} \mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) = \mathrm{I}_{\mathcal{V}}(Y \to X) \quad (25)$$

Proof. See App. H. □

4 A Framework for Incremental Probing

The proposed Bayesian framework for information allows us to take into account the amount of data we have for probing. Crucially, previous work (Pimentel et al., 2020b) failed to adequately account for the observation of data. In doing so, they only analysed the limiting behaviour of information, under which the probing enterprise is not fully sensible—given unlimited data and computation, there is no point in using pre-trained functions. Indeed, the higher-level motivation of this work is to find an information-theoretic framework which serves machine learning, and under which the goal of probing *is* inherently clear. To that end, we propose a relatively simple experimental design. We compute Bayesian mutual information, which is a function of the amount of data, to create several learning curves.

[6] In the case where $\boldsymbol{\theta}$ is discrete and finite, the only requirement is $p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) > 0$, for all values of $\boldsymbol{\theta}$ (Freedman, 1963).

Notation. We define a sentence-level random variable S, with instances $\mathbf{s}$ taken from V^*, the Kleene closure of a potentially infinite vocabulary V. We further define a representation-valued random variable R and a task-valued random variable T, each with instances $r \in \mathbb{R}^d$ and $t \in \mathcal{T}$, where $\mathcal{T}$ is the set of possible values for the analysed task (e.g. the set of parts of speech in a language).

4.1 Probes as Bayesian Agents

The overall trend in NLP is to train supervised probabilistic models on task-specific data. We believe probabilistic probes should analogously be modelled this way—leading to results compatible with our empirical intuitions. We thus define a **probe agent** as a Bayesian agent with the pair of beliefs $\{p_{\boldsymbol{\theta}}(t \mid \boldsymbol{\theta}), p_{\boldsymbol{\theta}}(t \mid r, \boldsymbol{\theta})\}$ and a prior $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. Any prior $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ could be chosen for our probing agents. Nonetheless we have no *a priori* knowledge of how the representations should impact our prediction task. As such, our priors are such that the initial distributions $p_{\boldsymbol{\theta}}(t \mid \mathbf{d}_0)$ and $p_{\boldsymbol{\theta}}(t \mid r, \mathbf{d}_0)$ are identical. A logical conclusion, is that the prior Bayesian MI should be zero:

$$\mathrm{I}_{\boldsymbol{\theta}}(R \to T \mid \mathbf{d}_0) = 0 \qquad (26)$$

On the opposite extreme—i.e. given unlimited data—a well-formed belief will likely converge to the true distribution, yielding the same results as by Pimentel et al. (2020b). Complementarily, an ill-formed belief will converge to the $\mathcal{V}$-information:

$$\begin{aligned}\lim_{N\to\infty} \mathrm{I}_{\boldsymbol{\theta}}(R \to T \mid \mathbf{d}_N) &= \mathrm{I}_{\mathcal{V}}(R \to T) \qquad (27)\\ &\approx \mathrm{I}(S;T)\end{aligned}$$

The novelty of our framework lies in the explicit analysis of information under finite data. Bayesian agents are used here to measure a notion of information directly related to **ease of extraction**—i.e. how much information could be extracted from the representations by a naïve agent with no *a priori* knowledge about the task itself. In other words, we ask the question: *given a specific dataset* $\mathbf{d}_N$*, how much information do the representations yield about this task?* This value is only a subset of the true MI, being upper-bounded by it.

Why Bayesian MI and not Bayesian entropy? We focus our analysis on the amount of information a Bayesian agent can extract from the representations about the task. However, we could as easily analyse the Bayesian entropy instead. We believe, though, that the Bayesian MI is an inherently more intuitive value than the entropy. This is because mutual information puts the Bayesian entropy in perspective to a trivial baseline—how much uncertainty would there be *without the representations*. Furthermore, it has a much more interpretable value: with no data its value is zero, while at the limit it converges to the true mutual information. In this paper, we are concerned with its trajectory, i.e., how fast does the Bayesian MI go up?

4.2 Ease of Extraction and Previous Work

Generally speaking, the goal of probing is to test if a set of contextual representations encodes a certain linguistic property (Adi et al., 2017; Belinkov et al., 2017; Tenney et al., 2019; Liu et al., 2021, *inter alia*). Most work in this field claims that, when performing this analysis, we should prefer simple models as probes (Alain and Bengio, 2016; Hewitt and Liang, 2019; Voita and Titov, 2020). This is inline with Pimentel et al.'s results: using a complex probe (complex enough to ensure it is well-formed) with infinite data, we would estimate $\mathrm{I}(S;T)$—a value which does not meaningfully inform us about the representations themselves. Defining model complexity, though, is not trivial (for a longer discussion see Pimentel et al., 2020a). For this reason, many works limit themselves to studying only linearly encoded information (e.g. Alain and Bengio, 2016; Hewitt and Manning, 2019; Hall Maudslay et al., 2020) or a subset of neurons at a time (e.g. Torroba Hennigen et al., 2020; Mu and Andreas, 2020; Durrani et al., 2020). However, restricting our analysis this way seems arbitrary.

A few recent papers have tried to deal with probe complexity in a more nuanced way. Hewitt and Liang (2019) argue for the use of selectivity to control for probe complexity. Voita and Titov (2020) and Whitney et al. (2020) use, respectively, minimum description length (MDL) and surplus description length (SDL) to measure the size (in bits) of the probe model. Pimentel et al. (2020a) argues probe complexity and accuracy should be seen as a Pareto trade-off, and propose new metrics to measure probe complexity. All of these papers define ease of extraction in terms of properties of the probe, e.g., its complexity and size.

We argue here for an opposing view of ease of extraction: Instead of focusing on the complexity of the probes, we should define it according to the **complexity of the task**. We further oper-

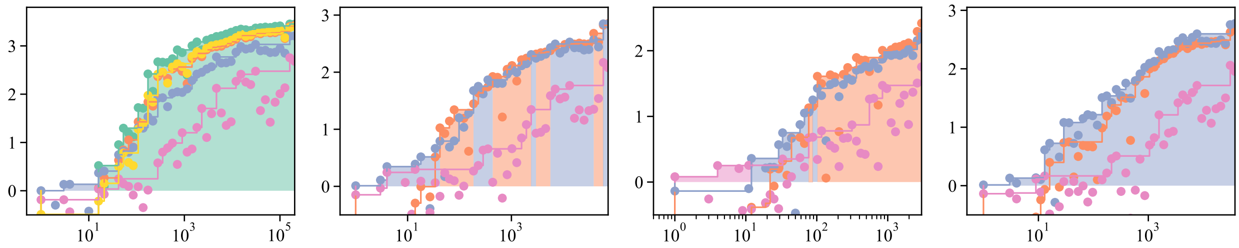

Figure 2: Bayesian MI (bits; y-axis) vs number of data examples (x-axis) in part of speech tagging on (left) English (center-left) Basque, (center-right) Marathi, and (right) Turkish. **ALBERT**, **RoBERTa**, **BERT**, **fastText**, **Random**

ationalise this complexity in a very specific way: how much information a Bayesian agent can extract from the representations, given *limited knowledge* about the task itself—where this limited knowledge is enforced by the size of the observed dataset $\mathbf{d}_N$.[7] With this in mind, we evaluate the Bayesian MI learning curves. In this regard, our analysis is similar to the learning curves used by Talmor et al. (2020) and the complexity–accuracy tradeoffs from Pimentel et al. (2020a).

5 Experiments and Results[8]

5.1 Data and Representations

We focus on part-of-speech (POS) tagging and dependency-arc labelling in our experiments. With this in mind, we make use of the universal dependencies (UD 2.6; Zeman et al., 2020); analysing the treebanks of four typologically diverse languages, namely: Basque, English, Marathi, and Turkish. As our object of analysis, we look at the contextual representations from ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2019),[9] using as a baseline the non-contextual fastText (Bojanowski et al., 2017) and random embeddings. Random embeddings are initialised at the type level and kept fixed during experiments.

5.2 Probe

Our experiments focus on Bayesian agents with multi-layer perceptron (MLP) beliefs:

$$p_{\boldsymbol{\theta}}(t \mid r, \boldsymbol{\theta}) = \mathrm{softmax}(\mathrm{MLP}(r; \boldsymbol{\phi})) \tag{28}$$

$$p_{\boldsymbol{\theta}}(t \mid \boldsymbol{\theta}) = \mathrm{Cat}(\boldsymbol{\psi}) = \frac{\psi_t}{\sum_{t'=1}^{|\mathcal{T}|} \psi_{t'}} \tag{29}$$

where $\boldsymbol{\phi}$ are the MLP parameters, $\boldsymbol{\psi} \in \mathbb{N}^{|\mathcal{T}|}$ is a count vector and $\boldsymbol{\theta} = [\boldsymbol{\phi}; \boldsymbol{\psi}]$ are the agent's parameters. This agent has a Gaussian prior over parameters $\boldsymbol{\phi}$ (with zero mean and standard deviation $\sigma = 10$), and a Dirichlet distribution prior over $\boldsymbol{\psi}$ (with concentration parameter $\alpha = 1$).

As previously discussed, the Gaussian and Dirichlet priors on the parameters will cause these models to initially place a uniform distribution on the output classes—as such, they will have an initial Bayesian MI of zero. We then expose the probe agent to increasingly larger sets of data from the task. Unfortunately, the posterior of eq. (28) has no closed form solution, so we approximate it with the maximum-a-posteriori probability $p_{\boldsymbol{\theta}}(t \mid r, \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\Theta} p_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid \mathbf{d}_n)$. We obtain this MAP estimate using the gradient descent method AdamW (Loshchilov and Hutter, 2019) with a cross-entropy loss and L2 norm regularisation.[10] The posterior predictive belief of eq. (29) has a closed-form solution[11]

$$p_{\boldsymbol{\theta}}(t \mid \mathbf{d}_N) = \frac{\mathrm{count}(\mathbf{d}_N, t) + 1}{N + |\mathcal{T}|} \tag{30}$$

where $\mathrm{count}(\mathbf{d}_N, t)$ is the number of observed instances of class t.

For both analysed tasks, we run 50 experiments with log-linearly increasing data sizes, from 1 instance to the whole language's treebank. For each of these individual experiments, we sample an MLP probe configuration. This probe will have 0, 1, or 2 layers—where 0 layers means a linear probe—dropout between 0 and 0.5, and hidden size from 32 to 1024 (log distributed). We then use the same architecture to train a probe for each of our analysed representations, plotting their Pareto curves.

[7] As we show later in the paper, this background knowledge about the task can also be formally defined as a Bayesian mutual information, i.e. the information the observed data provides about the model parameters $\mathrm{I}_{\boldsymbol{\theta}}(\mathbf{D}_N \to \Theta)$.

[8] Our code is available in `https://www.github.com/rycolab/bayesian-mi`.

[9] We use the pre-trained models made available by the transformers library (Wolf et al., 2019).

[10] L2 weight decay regularisation is equivalent to a Gaussian prior on the parameter space (pg. 350, Bishop, 1995).

[11] This posterior predictive distribution is equivalent to Laplace smoothing (Jeffreys, 1939; Robert et al., 2009).

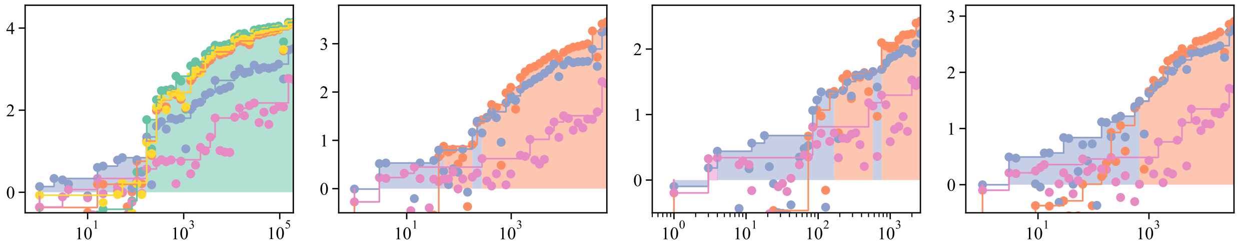

Figure 3: Bayesian MI (bits; y-axis) vs number of examples (x-axis) in dependency arc labelling on (left) English (center-left) Basque (center-right) Marathi, and (right) Turkish. **ALBERT**, **RoBERTa**, **BERT**, **fastText**, **Random**

5.3 Discussion

Fig. 2 presents pareto curves for part-of-speech tagging. These curves convey a few interesting results. The first is the intuitive fact that information is much harder to extract with random embeddings, although with enough training data their results slowly converge to near the fastText ones—this can be seen most clearly in English. This matches our theoretical framework: the true mutual information between the target task and either fastText or random embeddings is the same, thus, if our beliefs are well-formed, the Bayesian MI should converge to this value, although with different speeds. The second result is that ALBERT makes information more easily extractable than either BERT or RoBERTa in English, and that multilingual BERT is roughly equally as informative as fastText under the finite data scenarios of the other analysed languages. Finally, the last result goes against one of the claims of Pimentel et al. (2020a), who in light of their flat Pareto curves for POS tagging claimed that we needed harder tasks for probing. One only needs harder tasks if their measure of complexity is not nuanced enough—as we see, even POS tagging is hard under the low-resource scenarios presented in our learning curves.

Fig. 3 presents results for dependency arc labelling. These learning curves also present interesting trends. While the POS tagging curves seem to be on the verge of convergence for English, Basque and Turkish, this is not the case for dependency arc labelling. This implies that, as expected, dependency arc labelling is either an inherently harder task, or that the representations encode the necessary information in a harder to extract manner. These results, also highlight the importance of an information-theoretic measure being able to capture negative information—as evinced in Fig. 4. For the low-data scenario, the BERToid models hurt performance, as opposed to helping. This is because high-dimensional representations, together with a weak prior, allow the agent to easily overfit to the little presented evidence. On the other hand, fastText does not present the same problem, having a positive Bayesian MI even in a low-data setting.

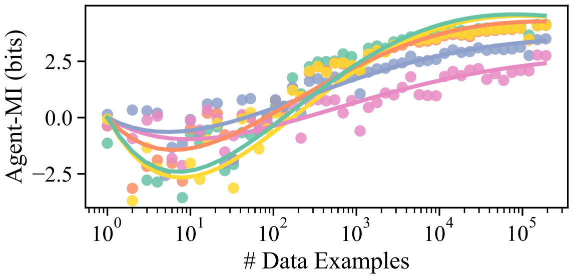

Figure 4: Dependency arc labelling polyfit on English.

6 An Intuitive Decomposition

We now present some basic results about our framework which, although not strictly necessary for the present study, help motivate it. They also serve as a justification for our choice of cross-entropy when formalising Bayesian entropy. With this in mind, we analyse information from the perspective of a fully Bayesian agent with a well-formed belief.[12] A classic decomposition of the cross-entropy is the following:

$$\mathrm{H}_{\boldsymbol{\theta}}(T) = \mathrm{H}(T) + \mathrm{KL}(p \mid\mid p_{\boldsymbol{\theta}}) \tag{31}$$

We posit a new interpretation for this equality.

Theorem 6. *Let Θ be a parameter-valued random variable. The entropy of a consistent Bayesian agent with well-formed beliefs decomposes as*

$$\mathrm{H}_{\boldsymbol{\theta}}(T \mid \mathbf{d}_N) = \underbrace{\mathrm{H}(T)}_{\textit{entropy}} + \underbrace{\mathrm{I}_{\boldsymbol{\theta}}(T \to \Theta \mid \mathbf{d}_N)}_{\textit{information about distribution}}$$

Proof. See App. I. □

In other words, the cross-entropy is composed of the sum between the entropy itself—i.e. the "true" information the data source provides, or its inherent uncertainty—and how much information the data provides about its distribution itself.

[12] We make the same analysis from the perspective of an agent with an ill-formed belief in App. A

Relation to SDL. The minimum description length (MDL; Voita and Titov, 2020) is a probing metric defined as $\mathrm{H}_{\boldsymbol{\theta}}(\mathbf{D}_N)$. In its online coding interpretation, it is rewritten as (Rissanen, 1978; Blier and Ollivier, 2018): $\mathrm{H}_{\boldsymbol{\theta}}(\mathbf{D}_N) = \sum_{n=1}^{N} \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{D}_{n-1})$—where the cross-entropy of each element X in $\mathbf{D}_N$ is computed incrementally because the parameter $\boldsymbol{\theta}$ (which would make them independent) is unknown. The surplus description length (SDL; Whitney et al., 2020) is defined as the difference between a dataset's cross-entropy and its entropy: $\mathrm{H}_{\boldsymbol{\theta}}(\mathbf{D}_N) - \mathrm{H}(\mathbf{D}_N)$. Using Theorem 6, we derive a new interpretation for SDL:

$$\mathrm{I}_{\boldsymbol{\theta}}(\mathbf{D}_N \to \Theta) = \mathrm{H}_{\boldsymbol{\theta}}(\mathbf{D}_N) - \mathrm{H}(\mathbf{D}_N) \qquad (32)$$

where we use prior predictive distributions, as opposed to posterior predictive ones. From this equation, we find that SDL is the information a dataset gives a Bayesian agent about its model parameters.

While closely related to one another, the Bayesian MI, MDL and SDL converge to different values in the limit of infinite dataset sizes:

$$\lim_{N\to\infty} \underbrace{\mathrm{H}_{\boldsymbol{\theta}}(\mathbf{D}_N)}_{\text{MDL}} \to \infty \qquad (33)$$

$$\lim_{N\to\infty} \underbrace{\mathrm{I}_{\boldsymbol{\theta}}(\mathbf{D}_N \to \Theta)}_{\text{SDL}} \to \infty \qquad (34)$$

$$\lim_{N\to\infty} \underbrace{\mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N)}_{\text{Bayesian MI}} \to \mathrm{I}(X;Y) \qquad (35)$$

It is easy to see that MDL goes to infinity as the dataset size grows—$\mathrm{H}_{\boldsymbol{\theta}}(\mathbf{D}_N)$ grows at least linearly with the data size. The reasons behind SDL also exploding as the data increases are less straightforward, though, but become clear from its Bayesian MI interpretation. If the parameter space is continuous, and if the Bayesian belief converges at the limit of infinite data (as per Theorem 4), the Bayesian mutual information in eq. (32) will naturally go to infinity.[13] We thus argue that Bayesian mutual information is a better measure for probing than either MDL or SDL; although all are sensitive to the observed dataset size, Bayesian MI is the only that does not diverge as this size grows.

7 Conclusion

In this paper we proposed an information-theoretic framework to analyse mutual information from the perspective of a Bayesian agent; we term this Bayesian mutual information. This framework has intuitive properties (at least from a machine learning perspective), which traditional information theory does not, for example: data can be informative, processing can help, and information can hurt. In the experimental portion of our paper, we use Bayesian mutual information to probe representations for both part-of-speech tagging and dependency arc labelling. We show that ALBERT is the most informative of the analysed representations in English; and high dimensional representations can provide negative information on low data scenarios.

Acknowledgements

We thank Adina Williams, Vincent Fortuin, Alex Immer, Lucas Torroba Hennigen and Mário Alvim for providing feedback in various stages of this paper. We also thank the anonymous reviewers for their valuable feedback in improving this paper.

Ethical Considerations

The authors foresee no ethical concerns with the research presented in this paper.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Mário S. Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. 2019. *The Science of Quantitative Information Flow*. Springer.

Maurice Bartlett. 1953. The statistical approach to the analysis of time-series. *Transactions of the IRE Professional Group on Information Theory*, 1(1):81–101.

[13] This is a byproduct of the properties of differential entropies (the entropy of continuous random variables). As the distribution $p_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid \mathbf{d}_N)$ converges to a Dirac delta distribution centred on the optimal parameters, which has an entropy of negative infinity, this Bayesian MI goes to positive infinite.

$$\begin{aligned}\lim_{N\to\infty} \mathrm{I}_{\boldsymbol{\theta}}(\mathbf{D}_N \to \Theta) &= \lim_{N\to\infty} (\mathrm{H}_{\boldsymbol{\theta}}(\Theta) - \mathrm{H}_{\boldsymbol{\theta}}(\Theta \mid \mathbf{d}_N)) \\ &= \mathrm{H}_{\boldsymbol{\theta}}(\Theta) - (-\infty) \\ &= \infty \qquad (36)\end{aligned}$$

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Peter J. Bickel and Kjell A. Doksum. 2001. *Mathematical statistics: Basic ideas and selected topics. Volume I*, second edition. Prentice Hall.

Christopher M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

Léonard Blier and Yann Ollivier. 2018. The description length of deep learning models. In *Advances in Neural Information Processing Systems*, pages 2216–2226.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Michael R. Clarkson, Andrew C. Myers, and Fred B. Schneider. 2005. Belief in information flow. In *18th IEEE Computer Security Foundations Workshop (CSFW'05)*, pages 31–45. IEEE.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of information theory*, second edition. Wiley-Interscience.

Morris H. DeGroot. 1962. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

Jin Fan. 2014. An information theory account of cognitive control. *Frontiers in Human Neuroscience*, 8:680.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

David A. Freedman. 1963. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403.

Charles R. Gallistel and Adam Philip King. 2011. *Memory and the computational brain: Why cognitive science will transform neuroscience*, volume 6. John Wiley & Sons.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Sardaouna Hamadou, Vladimiro Sassone, and Catuscia Palamidessi. 2010. Reconciling belief and vulnerability in information flow. In *IEEE Symposium on Security and Privacy*, pages 79–92. IEEE.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Laurent Itti and Pierre F. Baldi. 2006. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Laurent Itti and Pierre F. Baldi. 2009. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306.

Harold Jeffreys. 1939. *The Theory of Probability*. The Clarendon Press, Oxford.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bastiaan Jan Korneel Kleijn and Adrianus Willem van der Vaart. 2012. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Yuqing Kong. 2020. Dominantly truthful multi-task peer prediction with a constant number of tasks. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2398–2411. SIAM.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *The 8th International Conference on Learning Representations*.

Ervin K. Lenzi, Renio S. Mendes, and Luciano R. da Silva. 2000. Statistical mechanics based on Rényi entropy. *Physica A: Statistical Mechanics and its Applications*, 280(3-4):337–345.

Dennis V. Lindley. 1956. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005.

Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

David McAllester and Karl Stratos. 2020. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. In *Advances in Neural Information Processing Systems*, volume 33, pages 17153–17163. Curran Associates, Inc.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 851–864, Online. Association for Computational Linguistics.

Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020b. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Christian P. Robert, Nicolas Chopin, and Judith Rousseau. 2009. Harold Jeffreys's theory of probability revisited. *Statistical Science*, 24(2):141–172.

Khalid Sayood. 2018. Information theory and cognition: A review. *Entropy*, 20(9):706.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Jan Storck, Sepp Hochreiter, and Jürgen Schmidhuber. 1995. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the International Conference on Artificial Neural Networks*, volume 2, pages 159–164, Paris, France. Citeseer.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Aad W. van der Vaart. 2000. *Asymptotic statistics*, volume 3. Cambridge University Press.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

William F. Whitney, Min Jae Song, David Brandfonbrener, Jaan Altosaar, and Kyunghyun Cho. 2020. Evaluating representations by the complexity of learning low-loss predictors. *arXiv preprint arXiv:2009.07368*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Angelina Aquino, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Ethan Chi, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Ill-formed Beliefs Loose Information

For the sake of argument, we now assume an agent with an ill-defined belief $p_{\boldsymbol{\theta}}(t \mid \boldsymbol{\theta})$ and a prior $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. We will show that such Bayesian agents loose information, meaning that they will not obtain as much information about their optimal parameters as if they had a well-formed belief.

Theorem 7. *Assume $\boldsymbol{\theta}^*$ are the optimal parameters for a Bayesian agent with ill-formed, but consistent beliefs. The information this agent will receive about its optimal parameters is*

$$\mathrm{I}_{\boldsymbol{\theta}}(T \to \Theta = \boldsymbol{\theta}^* \mid \mathbf{d}_N) < \mathrm{H}_{\boldsymbol{\theta}}(T \mid \mathbf{d}_N) - \mathrm{H}(T) \tag{37}$$

Proof. This proof follows from the Bayesian MI definition, from this Bayesian agent having consistent beliefs, and from the fact that the cross-entropy is an upper-bound to the entropy, with equality only when both probability distributions are the same—which by definition is not possible $p_{\boldsymbol{\theta}}(t \mid \boldsymbol{\theta}^*) \neq p(t)$

$$\begin{aligned}
\mathrm{I}_{\boldsymbol{\theta}}(T \to \Theta = \boldsymbol{\theta}^* \mid \mathbf{d}_N) &= \mathrm{I}_{\boldsymbol{\theta}}(\Theta = \boldsymbol{\theta}^* \to T \mid \mathbf{d}_N) && \text{(symmetry due to belief consistency)} && \text{(38a)}\\
&= \mathrm{H}_{\boldsymbol{\theta}}(T \mid \mathbf{d}_N) - \mathrm{H}_{\boldsymbol{\theta}}(T \mid \Theta = \boldsymbol{\theta}^*, \mathbf{d}_N) && && \text{(38b)}\\
&= \mathrm{H}_{\boldsymbol{\theta}}(T \mid \mathbf{d}_N) - \mathrm{H}_{\boldsymbol{\theta}}(T \mid \Theta = \boldsymbol{\theta}^*) && && \text{(38c)}\\
&< \mathrm{H}_{\boldsymbol{\theta}}(T \mid \mathbf{d}_N) - \mathrm{H}(T) && \text{(strict inequality due to ill-formed beliefs)} && \text{(38d)}
\end{aligned}$$

□

B Measures of Information

Several other measures of information have been proposed, among them are the H entropy (DeGroot, 1962), the Rényi entropy (Rényi, 1961; Lenzi et al., 2000), Bayes vulnerability (Alvim et al., 2019), and the Determinantal Mutual Information (DMI; Kong, 2020). None of these take an agent's belief into consideration, and so our analysis is orthogonal to them. The work most similar to ours, in this respect, is Clarkson et al.'s (2005) investigation of how belief impacts information leakage—and its extension, by Hamadou et al. (2010), to the Rényi min-entropy. Importantly, the results obtained by Clarkson et al. can be similarly derived using our framework.

C A Note on Empirical Limitations

Estimating the true MI between two random variables is known to be a hard problem for which several methods have been proposed (for a detailed review, see McAllester and Stratos, 2020)—estimating the Bayesian MI may be equally challenging. Given knowledge of $p_{\boldsymbol{\theta}}(\cdot)$ and having access to samples from $p(\cdot)$, the Bayesian MI can be trivially estimated using the Bayesian surprisal's sample mean. On the other hand, in a setting such as active learning, where one (by definition) does not have access to the true distribution $p(y \mid x)$—only to the belief—the best approximation to the Bayesian MI may indeed be the belief-MI (used by Houlsby et al. 2011) or the Bayesian surprise (used by Storck et al. 1995 and Itti and Baldi 2006, 2009). Finally, approximating the Bayesian MI in the cognitive sciences may be an even harder problem than estimating the true MI, since it would require approximating both the belief $p_{\boldsymbol{\theta}}(\cdot)$ of a specific agent and the true distribution $p(\cdot)$ of an event.

D Proof of Symmetric Bayesian Mutual Information, Theorem 1

Theorem 1. *An agent's Bayesian mutual information is symmetric, i.e.*

$$\mathrm{I}_{\boldsymbol{\theta}}(X \to Y \mid \mathbf{d}_N) = \mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) \tag{39}$$

for all distributions $p(x, y)$ if and only if the Bayesian agent is consistent.

Proof. We will first prove that if the Bayesian MI is symmetric for all true distributions $p(x, y)$, then the Bayesian agent is consistent (the *if* case). We then prove the inverse proposition (the *only if* case), completing this *if and only if* theorem's proof.

$\Rightarrow$ We show this, by relying on specific distributions where $p(x, y)$ is deterministic, putting all probability mass in a single point, i.e. $p(x_0, y_0) = 1$.

$$\mathrm{I}_{\boldsymbol{\theta}}(X \to Y) = \mathrm{I}_{\boldsymbol{\theta}}(Y \to X) \tag{40a}$$

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log \frac{p_{\boldsymbol{\theta}}(y \mid x)}{p_{\boldsymbol{\theta}}(y)} = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log \frac{p_{\boldsymbol{\theta}}(x \mid y)}{p_{\boldsymbol{\theta}}(x)} \tag{40b}$$

$$p(x_0, y_0) \log \frac{p_{\boldsymbol{\theta}}(y_0 \mid x_0)}{p_{\boldsymbol{\theta}}(y_0)} = p(x_0, y_0) \log \frac{p_{\boldsymbol{\theta}}(x_0 \mid y_0)}{p_{\boldsymbol{\theta}}(x_0)} \tag{40c}$$

$$\frac{p_{\boldsymbol{\theta}}(y_0 \mid x_0)}{p_{\boldsymbol{\theta}}(y_0)} = \frac{p_{\boldsymbol{\theta}}(x_0 \mid y_0)}{p_{\boldsymbol{\theta}}(x_0)} \tag{40d}$$

As we can show this same result for any value of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we conclude the agents must have consistent beliefs, i.e.

$$\forall x \in \mathcal{X}, y \in \mathcal{Y}: \qquad \frac{p(y \mid x)}{p(y)} = \frac{p(x \mid y)}{p(x)} \tag{41}$$

$\Leftarrow$ Bayes' rule can be written as

$$\frac{p(y \mid x)}{p(y)} = \frac{p(x \mid y)}{p(x)} \tag{42}$$

We now show that all consistent agents will have symmetric MI

$$\mathrm{I}_{\boldsymbol{\theta}}(X \to Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log \frac{p_{\boldsymbol{\theta}}(y \mid x)}{p_{\boldsymbol{\theta}}(y)} \tag{43a}$$

$$\overset{(1)}{=} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log \frac{p_{\boldsymbol{\theta}}(x \mid y)}{p_{\boldsymbol{\theta}}(x)} \tag{43b}$$

$$= \mathrm{I}_{\boldsymbol{\theta}}(Y \to X) \tag{43c}$$

where equality (1) relies on Bayes' rule.

□

E Proof of No Data Processing Inequality, Theorem 2

Theorem 2. *The data processing inequality does not hold for Bayesian information, i.e.*

$$\mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) \;?\; \mathrm{I}_{\boldsymbol{\theta}}(f(Y) \to X \mid \mathbf{d}_N) \tag{44}$$

Proof. We prove this theorem with a counter example. Let Y be a Poisson distributed random variable with unknown mean, i.e. $p(y) = \mathrm{Pois}(\widehat{\theta})$, where $\widehat{\theta}$ is this distributions true mean. We further define a second random variable Z, as:

$$f(y) = y - \widehat{\theta}, \qquad p(z \mid y) = \mathbb{1}\{z = f(y)\} \tag{45}$$

Z is, thus, a deterministic function of Y, where the function $f(y)$ mean-centres random variable Y. Finally, we also define X as a Bernoulli distributed random variable:

$$g(z) = \frac{1}{1 + e^{-z}}, \qquad p(x \mid z) = \begin{cases} g(z) & x = 1 \\ 1 - g(z) & x = 0 \end{cases} \tag{46}$$

where $g(z)$ is a sigmoid function. We can further define the distribution

$$p(x \mid y) = \begin{cases} (g \circ f)(y) & x = 1 \\ 1 - (g \circ f)(y) & x = 0 \end{cases} \tag{47}$$

From these distributions, we see that $\mathrm{H}(Z \mid Y) = 0$, as $f(y)$ is a deterministic function. Further, in this specific example $\mathrm{H}(X \mid Z) = \mathrm{H}(X \mid Y)$—this can be proved from the fact that $f(y)$ is a bijective function, and that the data processing inequality is tight for bijections

$$\mathrm{H}(X \mid Y) \leq \mathrm{H}(X \mid f(Y)) = \mathrm{H}(X \mid Z) \leq \mathrm{H}(X \mid f^{-1}(Z)) = \mathrm{H}(X \mid Y) \tag{48}$$

We now define a Bayesian agent, which correctly knows the relationship between Y, Z and X, i.e. with well-formed beliefs $p_{\boldsymbol{\theta}}(x \mid y, \theta)$ and $p_{\boldsymbol{\theta}}(x \mid z, \theta)$, and with a prior $p_{\boldsymbol{\theta}}(\theta)$—this agent does not know the true value of parameter θ though. For this Bayesian agent

$$\mathrm{H}_{\boldsymbol{\theta}}(X \mid Y, \mathbf{d}_N) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \int p_{\boldsymbol{\theta}}(x \mid y, \theta) p_{\boldsymbol{\theta}}(\theta \mid \mathbf{d}_N) \mathrm{d}\theta \tag{49}$$

When given $Z = f(Y)$, however, this agent does not need to know θ, since the data is already mean-centred (there are no unknown parameters in $p_{\boldsymbol{\theta}}(x \mid z)$). This Bayesian agent's conditional entropy given Z is

$$\begin{aligned}
\mathrm{H}_{\boldsymbol{\theta}}(X \mid Z, \mathbf{d}_N) &= \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} p(x, z) \log p_{\boldsymbol{\theta}}(x \mid z, \mathbf{d}_N) && \text{(50a)}\\
&= \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} p(x, z) \log p(x \mid z, \mathbf{d}_N) && \text{(50b)}\\
&= \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} p(x, z) \log p(x \mid z) && \text{(50c)}\\
&= \mathrm{H}(X \mid Z) && \text{(50d)}
\end{aligned}$$

This concludes the example that a deterministic (mean-centring) function can help this Bayesian agent.

$$\begin{aligned}
\mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) &= \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}_{\boldsymbol{\theta}}(X \mid Y, \mathbf{d}_N) && \text{(51a)}\\
&\overset{(1)}{\leq} \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}(X \mid Y) && \text{(51b)}\\
&= \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}(X \mid Z) && \text{(51c)}\\
&= \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}_{\boldsymbol{\theta}}(X \mid Z, \mathbf{d}_N) && \text{(51d)}\\
&= \mathrm{I}_{\boldsymbol{\theta}}(Z \to X \mid \mathbf{d}_N) && \text{(51e)}\\
&= \mathrm{I}_{\boldsymbol{\theta}}(f(Y) \to X \mid \mathbf{d}_N) && \text{(51f)}
\end{aligned}$$

where (1) becomes a strict inequality if the belief $p_{\boldsymbol{\theta}}(\theta) \neq \delta(\theta - \widehat{\theta})$, i.e. if the prior does not place all probability mass in the true parameters $\widehat{\theta}$. □

F Proof of Bayesian MI is Upper-bounded by the True MI, Theorem 3

Theorem 3. *Assuming the agent's belief $p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)$ is tighter than the marginal of its beliefs over y, i.e. than $\sum_{y \in \mathcal{Y}} p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N) p(y)$. We show*

$$\mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) \leq \mathrm{I}(X; Y) \tag{52}$$

Proof. We start by noting that the difference between the true MI, and its Bayesian counterpart is equal to the difference between two KL-divergences

$$\begin{aligned}
\mathrm{I}(X;Y) - \mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) &= \mathrm{H}(X) - \mathrm{H}(X \mid Y) - \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) + \mathrm{H}_{\boldsymbol{\theta}}(X \mid Y, \mathbf{d}_N) && \text{(53a)}\\
&= \mathrm{KL}(p(x \mid y) \mid\mid p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N)) - \mathrm{KL}(p(x) \mid\mid p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)) && \text{(53b)}
\end{aligned}$$

To prove this theorem, we need to show that one KL-divergence is smaller than the other, i.e.

$$\mathrm{KL}(p(x) \mid\mid p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)) \leq \mathrm{KL}(p(x \mid y) \mid\mid p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N)) \tag{54}$$

We can show this with a bit of algebraic manipulation and an assumption about our Bayesian agent

$$\begin{aligned}
\mathrm{KL}(p(x \mid y) \,||\, p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N)) &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y \mid x) \log \frac{p(x \mid y)}{p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N)} && \text{(55a)} \\
&= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y \mid x) \log \frac{p(x \mid y) p(y)}{p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N) p(y)} && \text{(55b)} \\
&\overset{(1)}{\geq} \sum_{x \in \mathcal{X}} p(x) \left(\sum_{y \in \mathcal{Y}} p(y \mid x) \right) \log \frac{\sum_{y \in \mathcal{Y}} p(x \mid y) p(y)}{\sum_{y \in \mathcal{Y}} p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N) p(y)} && \text{(55c)} \\
&= \sum_{x \in \mathcal{X}} p(x) \cdot 1 \cdot \log \frac{p(x)}{\sum_{y \in \mathcal{Y}} p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N) p(y)} && \text{(55d)} \\
&\overset{(2)}{\geq} \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)} && \text{(55e)} \\
&= \mathrm{KL}(p(x) \,||\, p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)) && \text{(55f)}
\end{aligned}$$

In this equations, (1) relies on the log sum inequality, while (2) assumes the following inequality

$$\begin{aligned}
\mathrm{H}_{\boldsymbol{\theta}}(X) &= - \sum_{x \in \mathcal{X}} p(x) \log p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N) && \text{(56a)} \\
&\leq - \sum_{x \in \mathcal{X}} p(x) \log \sum_{y \in \mathcal{Y}} p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N) p(y) && \text{(56b)}
\end{aligned}$$

This is equivalent to our assumption that this agent's estimate of $p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)$ is tighter than if the agent marginalised its beliefs over y. While not necessarily true, in practice, if X is discrete and has a small cardinality $|\mathcal{X}|$, a simple Laplace smoothed estimate of $p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)$ is likely to result in this inequality. One could instead assume an agent which uses a Monte Carlo sampling approximation for estimating $p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)$ from $p_{\boldsymbol{\theta}}(x \mid y, \mathbf{d}_N)$. This would switch the inequality (2) for an approximation, and result in an expected lower bound instead

$$\mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) \lesssim \mathrm{I}(X; Y) \tag{57}$$

□

G Proof of the Convergence to Mutual Information, Theorem 4

Theorem 4 *If we assume a Bayesian agent's set of beliefs and prior are well-formed and meet the conditions of Bernstein–von Mises Theorem (pg. 339, Bickel and Doksum, 2001). Then,*

$$\lim_{N \to \infty} \mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) = \mathrm{I}(X; Y) \tag{58}$$

Proof. The Bernstein–von Mises Theorem only applies to well-formed beliefs, i.e. beliefs which can model the true probability distribution—a condition which is satisfied by our assumptions to this theorem. By this theorem—and under a number of other specified conditions, e.g. absolute continuity of the prior in a neighbourhood around $\widehat{\boldsymbol{\theta}}$ and continuous positive density at $\widehat{\boldsymbol{\theta}}$ (see pg. 141 in van der Vaart 2000 for the full set of conditions)—we have

$$p(x) = \lim_{N \to \infty} p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N) \tag{59}$$

Now, we apply the continuous mapping theorem to analyse the convergence of the Bayesian entropy

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) &= -\lim_{N\to\infty} \sum_{x\in\mathcal{X}} p(x) \log p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N) && \text{(60a)} \\
&\overset{(1)}{=} -\sum_{x\in\mathcal{X}} p(x) \log \left(\lim_{N\to\infty} p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N) \right) && \text{(60b)} \\
&= -\sum_{x\in\mathcal{X}} p(x) \log p(x) && \text{(60c)} \\
&= \mathrm{H}(X) && \text{(60d)}
\end{aligned}$$

where (1) relies on the continuous mapping theorem. A similar convergence applies to $\mathrm{H}_{\boldsymbol{\theta}}(X \mid Y, \mathbf{d}_N)$. Finally, we can complete the proof

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) &= \lim_{N\to\infty} (\mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}_{\boldsymbol{\theta}}(X \mid Y, \mathbf{d}_N)) && \text{(61a)} \\
&= \mathrm{H}(X) - \mathrm{H}(X \mid Y) && \text{(61b)} \\
&= \mathrm{I}(X; Y) && \text{(61c)}
\end{aligned}$$

□

H Proof of the Convergence to $\mathcal{V}$-information, Theorem 5

Theorem 5 *Assume a Bayesian agent's beliefs and prior meet the conditions of Kleijn and van der Vaart (2012), who extend the Bernstein–von Mises Theorem to beliefs which are not well-formed. Further, let $\mathcal{V} = \{p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{\theta}) \mid p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) > 0\}$. Then,*

$$\lim_{N\to\infty} \mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) = \mathrm{I}_{\mathcal{V}}(Y \to X) \tag{62}$$

Proof. Kleijn and van der Vaart (2012) extend the Bernstein–von Mises Theorem to ill-formed beliefs, showing that, under specific conditions for the Bayesian belief and priors, the predictive posterior distribution converges to

$$\lim_{N\to\infty} p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N) = p_{\boldsymbol{\theta}}(x \mid \boldsymbol{\theta}^*) \tag{63}$$

where $\boldsymbol{\theta}^*$ is a unique set of parameters which minimises the KL-divergence between $p_{\boldsymbol{\theta}}(x \mid \boldsymbol{\theta})$ and the true distribution $p(x)$, i.e.

$$p_{\boldsymbol{\theta}}(x \mid \boldsymbol{\theta}^*) = \arg \inf_{q\in\mathcal{V}} \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{q(x)} \tag{64}$$

Given this convergence property, we can finish the proof similarly to the one for the well-formed belief:

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) &= \lim_{N\to\infty} \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)} && \text{(65a)} \\
&\overset{(1)}{=} \sum_{x\in\mathcal{X}} p(x) \log \left(\lim_{N\to\infty} \frac{1}{p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)} \right) && \text{(65b)} \\
&= \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p_{\boldsymbol{\theta}}(x \mid \boldsymbol{\theta}^*)} && \text{(65c)} \\
&= \mathrm{H}_{\mathcal{V}}(X) && \text{(65d)}
\end{aligned}$$

where $\mathcal{V}$ is defined as $\{p(x \mid \boldsymbol{\theta}) \mid p_{\boldsymbol{\theta}}(\boldsymbol{\theta}) > 0\}$, and (1) relies on the continuous mapping theorem. We now conclude this proof:

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{I}_{\boldsymbol{\theta}}(Y \to X \mid \mathbf{d}_N) &= \lim_{N\to\infty} (\mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}_{\boldsymbol{\theta}}(X \mid Y, \mathbf{d}_N)) && \text{(66a)} \\
&= \mathrm{H}_{\mathcal{V}}(X) - \mathrm{H}_{\mathcal{V}}(X \mid Y) && \text{(66b)} \\
&= \mathrm{I}_{\mathcal{V}}(Y \to X) && \text{(66c)}
\end{aligned}$$

□

I Proof of the Intuitive Decomposition, Theorem 6

Theorem 6 *Let Θ be a parameter-valued random variable. The entropy of a consistent Bayesian agent with well-formed beliefs decomposes as*

$$\mathrm{H}_{\boldsymbol{\theta}}(T \mid \mathbf{d}_N) = \underbrace{\mathrm{H}(T)}_{\text{entropy}} + \underbrace{\mathrm{I}_{\boldsymbol{\theta}}(T \to \Theta \mid \mathbf{d}_N)}_{\text{information about distribution}} \tag{67}$$

Proof. Note that under the Bayesian MI only the information about the true model parameters, i.e. $\widehat{\boldsymbol{\theta}}$, matters

$$\mathrm{I}_{\boldsymbol{\theta}}(X \to \Theta \mid \mathbf{d}_N) = \sum_{x \in \mathcal{X}} \int p(x, \boldsymbol{\theta}) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid x, \mathbf{d}_N)}{p_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid \mathbf{d}_N)} \mathrm{d}\boldsymbol{\theta} \tag{68a}$$

$$\overset{(1)}{=} \sum_{x \in \mathcal{X}} \int p(x)\, \delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid x, \mathbf{d}_N)}{p_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid \mathbf{d}_N)} \mathrm{d}\boldsymbol{\theta} \tag{68b}$$

$$= \sum_{x \in \mathcal{X}} p(x) \log \frac{p_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} \mid x, \mathbf{d}_N)}{p_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} \mid \mathbf{d}_N)} \tag{68c}$$

$$= \mathrm{I}_{\boldsymbol{\theta}}(X \to \Theta = \widehat{\boldsymbol{\theta}} \mid \mathbf{d}_N) \tag{68d}$$

where (1) relies on the fact that the true $p(\boldsymbol{\theta})$ places all probability mass on the value $\widehat{\boldsymbol{\theta}}$. Using this result, we can show the Bayesian mutual information in eq. (67) is the same as the KL-divergence.

$$\mathrm{I}_{\boldsymbol{\theta}}(X \to \Theta \mid \mathbf{d}_N) = \mathrm{I}_{\boldsymbol{\theta}}(X \to \Theta = \widehat{\boldsymbol{\theta}} \mid \mathbf{d}_N) \tag{69a}$$

$$\overset{(2)}{=} \mathrm{I}_{\boldsymbol{\theta}}(\Theta = \widehat{\boldsymbol{\theta}} \to X \mid \mathbf{d}_N) \tag{69b}$$

$$= \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}_{\boldsymbol{\theta}}(X \mid \Theta = \widehat{\boldsymbol{\theta}}, \mathbf{d}_N) \tag{69c}$$

$$= \mathrm{H}_{\boldsymbol{\theta}}(X \mid \mathbf{d}_N) - \mathrm{H}(X) \tag{69d}$$

$$= \mathrm{KL}(p(x) \mid\mid p_{\boldsymbol{\theta}}(x \mid \mathbf{d}_N)) \tag{69e}$$

where (2) relies on the assumption that this agent's beliefs are consistent, and by definition $\mathrm{H}(X) = \mathrm{H}_{\boldsymbol{\theta}}(X \mid \Theta = \widehat{\boldsymbol{\theta}}, \mathbf{d}_N)$. □