Natural Abstractions: Key claims, Theorems, and Critiques

by LawrenceC, Leon Lang, Erik Jenner

TL;DR: *We distill John Wentworth's Natural Abstractions agenda by summarizing its key claims: the Natural Abstraction Hypothesis—many cognitive systems learn to use similar abstractions—and the Redundant Information Hypothesis—a particular mathematical description of natural abstractions. We also formalize proofs for several of its theoretical results. Finally, we critique the agenda's progress to date, alignment relevance, and current research methodology.*

Author Contributions: *Erik wrote a majority of the post and developed the breakdown into key claims. Leon formally proved the gKPD theorem and wrote most of the mathematical formalization section and appendix. Lawrence formally proved the Telephone theorem and wrote most of the related work section. All of us were involved in conceptual discussions and various small tasks.*

Epistemic Status: We're not John Wentworth, though we did confirm our understanding with him in person and shared a draft of this post with him beforehand.

Appendices: We have an additional appendix post and technical pdf containing further details and mathematical formalizations. We refer to them throughout the post at relevant places.

This post is long, and **for many readers we recommend using the table of contents to skip to only the parts they are most interested in** (e.g. the Key high-level claims to get a better sense for what the Natural Abstraction Hypothesis says, or our Discussion for readers already very familiar with natural abstractions who want to see our views). **Our Conclusion is also a decent 2-min summary of the entire post.**

Introduction

The Natural Abstraction Hypothesis (NAH) says that our universe abstracts well, in the sense that small high-level summaries of low-level systems exist, and that furthermore, these summaries are "natural", in the sense that many different cognitive systems learn to use them. There are also additional claims about how these natural abstractions should be formalized. We thus split up the Natural Abstraction Hypothesis into two main components that are sometimes conflated:

1. **The Universality Hypothesis:** Natural abstractions exist, i.e. many cognitive systems learn similar abstractions.
2. **The Redundant Information Hypothesis:** Natural abstractions are well described mathematically as functions of redundant or conserved information.

Closely connected to the Natural Abstraction Hypothesis are several mathematical results as well as plans to apply natural abstractions to AI alignment. We'll call all of these views together the *natural abstractions agenda*.

The natural abstractions agenda has been developed by John Wentworth over the last few years. The large number of posts on the subject, which often build on each other by each adding small pieces to the puzzle, can make it difficult to get a high-level overview of the key claims and results. Additionally, most of the mathematical definitions, theorems, and proofs are stated only informally, which makes it easy to mix up conjectures, proven claims, and conceptual intuitions if readers aren't careful.

In this post, we

- survey some existing related work, including in the academic literature,
- summarize the key conceptual claims behind the natural abstractions agenda and break them down into specific subclaims,
- formalize some of the key mathematical claims and provide formal proofs for them,
- outline the high-level plan for how the natural abstractions agenda aims to help with AI alignment,
- and critique the agenda by noting gaps in the theory, issues with the relation to alignment, and methodological criticisms.

All except the last of these sections are our attempt to describe John's views, not our own. That said, we attempt to explain things in the way that makes the most sense to us, which may differ from how John would phrase them somewhat. And while John met with us to clarify his thinking, it's still possible we're simply misunderstanding some of his views. The final section discusses our own views: we note some of our agreements but focus on the places where we disagree or see a need for additional work.

In the remainder of this introduction, we provide some high-level intuitions and motivation, and then survey existing distillations and critiques of the natural abstractions agenda. **Readers who are already quite familiar with natural abstractions may wish to skip directly to the next section.**

What do we mean by abstractions?

There are different perspectives on what abstractions are, but one feature is that they throw away a lot of unimportant information, turning a complex system into a smaller representation. This idea of throwing away irrelevant information is the key perspective for the natural

abstractions agenda. Cognitive systems can use these abstractions to make accurate predictions about important aspects of the world.

Let's look at an example (extended from one by John). A computer running a program can be modeled at many different levels of abstraction. On a very low level, lots of electrons are moving through the computer's chips, but this representation is much too complicated to work with. Luckily, it turns out we can throw away almost all the information, and just track voltages at various points on the chips. In most cases, we can predict high-level phenomena with the voltages almost as well as with a model of all the electrons, even though we're tracking vastly fewer variables. This continues to higher levels of abstraction: we can forget the exact voltages and just model the chip as an idealized logical circuit, and so on. Sometimes abstractions are leaky and this fails, but for good abstractions, those cases are rare.

Slightly more formally, an abstraction F is then a *description* or *function* that, when applied to a low-level system X, returns an abstract summary F(X).[1] F(X) can be thought of as throwing away lots of irrelevant information in X while keeping information that is important for making certain predictions.

Why expect abstractions to be *natural*?

Why should we expect abstractions to be *natural*, meaning that most cognitive systems will learn roughly the same abstractions?

First, note that not every abstraction works as well as the computer example we just gave. If we just throw away information in a random way, we will most likely end up with an abstraction that is missing some crucial pieces while also containing lots of useless details. In other words: some abstractions are much better than others.

Of course, which abstractions are useful does depend on which pieces of information are important, i.e. what we need to predict using our abstraction. But the second important idea is that most cognitive systems need to make predictions about similar things. Combined with the first point, that suggests they will use similar abstractions.

Why would different systems need to predict similar things in the environment? The reason is that distant pieces of the environment mostly don't influence each other in ways that can feasibly be predicted. Imagine a mouse fleeing from a cat. The mouse doesn't need to track how each of the cat's hairs move, since these small effects are quickly washed out by noise and never affect the mouse (in a way the mouse could predict). On the other hand, the higher-level abstractions "position and direction of movement of the cat" have more stable effects and thus *are* important. The same would be true for many other goals than surviving by fleeing the cat.

In addition to these conceptual arguments, there is some empirical evidence in favor of natural abstractions. For example, humans often learn a concept used by other humans based on just one or a few examples, suggesting natural abstractions at least among humans. More

interestingly, there are many cases of ML models discovering these human abstractions too (e.g. trees in GANs as John has discussed, or human chess concepts in AlphaZero).

It seems clear that abstractions are natural in *some* sense—that most possible abstractions are just not useful and won't be learned by any reasonable cognitive system. It's less clear just how much we should expect abstractions used by different systems to overlap. We will discuss the claims of the natural abstractions agenda about this more precisely later on.

Why study natural abstractions for alignment?

Why should natural abstractions have anything to do with AI alignment? As motivation for the rest of this post, we'll briefly explain some intuitions for this. We defer a full discussion until a later section.

One conceptualization of the alignment problem is to ensure that AI systems are "trying" to do what we "want" them to do. This raises two large conceptual questions:

- What does it mean to "try" to do "something"? What is this "something"?
- What does it mean for us to "want" "something"? Again, what is this "something"?

One interpretation of "something" is a particular set of physical configurations of the universe. However, this is considerably too complicated to fit into our brain, and we usually care more about high-level structures like our families or status. But what *are* these high-level structures fundamentally, and how can we mathematically talk about them? Intuitively, these structures throw away lots of detailed information about the universe, and thus, they are *abstractions*. So finding a theory of abstractions may be important to make progress on the conceptual question of what we and ML systems care about.

This is admittedly only a vague motivation, and we will later discuss more specific things we might do with a theory of natural abstractions. For example, a definition of abstractions might help find abstractions in neural networks, thus speeding up interpretability, and figuring out whether the universality hypothesis is true has strategic implications.

Existing writing on the natural abstractions agenda

The Natural Abstraction Hypothesis: Implications and Evidence is the largest existing distillation of the natural abstractions agenda. It follows John in dividing the Natural Abstraction Hypothesis into Abstractability, Human-Compatibility, and Convergence, whereas we will propose our own fine-grained subclaims. In addition to summarizing the natural abstractions agenda, the "Implications and Evidence" post mainly discusses possible sources of evidence about the Natural Abstraction Hypothesis. A much shorter summary of John's agenda, also touching on natural abstractions, can be found in What Everyone in Technical Alignment is Doing and Why. Finally, the Hebbian Natural Abstractions sequence aims to motivate the Natural Abstraction Hypothesis from a computational neuroscience perspective.

There have also been a few discussions and critiques related to the natural abstractions agenda. Charlie Steiner has speculated that there may be too many very similar natural abstractions to make them useful for alignment, or that AI systems may not learn enough natural abstractions, essentially questioning claims 1b and 1c in the list we will introduce below. Steve Byrnes has written about why the natural abstractions agenda doesn't focus on the most important alignment bottlenecks. These critiques are largely disjoint from the ones we will discuss later.

John himself has of course written by far the most about the natural abstractions agenda. We give a brief overview of his relevant writing in the appendix to make it easier for newcomers to dive in.

Related work

The universality hypothesis—that many systems will learn convergent abstractions/representations—is a key question in the field of neural network interpretability, and accordingly has been studied a substantial amount. Moreover, the intuitions behind the natural abstractions agenda and the redundant information hypothesis are commonly shared across different fields, of which we can highlight but a few.

Machine learning

Representation Learning

In machine learning, the subfield of representation learning studies how to extract representations of the data that have good downstream performance. Approaches to representation learning include next-frame/next-token prediction, autoencoding, infill/denoising, contrastive learning, predicting important variables of the environment, and many others. It's worth noting that, representations aren't always learned explicitly; for example, it's a standard trick in reinforcement learning to add auxiliary prediction losses or do massive self-supervised pretraining. It's worth noting that work in representation learning generally does not make claims as to universality of learned representations; instead, their focus is on learning representations that are useful for downstream tasks.

In particular, the field of *disentangled* representation learning shares many relevant tools and motivations to the redundant information hypothesis. In disentangled representation learning, we aim to learn representations that separate (that is, disentangle) parts of the world into disjoint parts.

The redundant information hypothesis is also especially related to information bottleneck methods, which aim to learn a good representation T of a variable X for variable Y by solving optimization problems of the form:

$\min_{p(t|x)} I(X;T) - \beta I(T,Y)$

In particular, we think that the deterministic information bottleneck, which tries to find the random variable T with minimum entropy, is quite similar in motivation to the idea of finding abstractions as redundant information.

The universality hypothesis in machine learning

The question of whether different neural networks learn the same representations has been studied in machine learning under the names *convergent learning* and the *universality hypothesis*. Here, the evidence for the universality of representations is more mixed. On one hand, different convolutional neural networks often exhibit similar circuits, have high correlated neurons, often learn similar representations, and learn to classify examples in a similar order. Models at different scales seem to consistently have heads that implement induction-like behavior. In particular, the fact that we can often align the internal representations of neural networks (e.g. see this paper) suggests that the neural networks are in some sense learning the same features of the world.

On the other hand, there are also many papers that argue against strong versions of feature universality. For example, even in the original convergent learning paper (Li et al 2014), the authors find that several features are idiosyncratic and are not shared across different networks. McCoy, Min, and Linzen 2019 find that different training runs of BERT generalize differently on downstream tasks. Recently, Chughtai, Chan, and Nanda 2023 investigated universality on group composition tasks, and found that different networks learn different representations in different orders, *even with the same architecture and data order*.

MCMC and Gibbs sampling

As John mentions in his redundant information post, the resampling-based definition of redundant information he introduces there is equivalent to running a Markov Chain Monte Carlo (MCMC) process. More specifically, this is essentially Gibbs sampling.[2] Redundant information corresponds to long mixing times (at least informally). But the motivation is of course different: in MCMC, we are usually interested in having short mixing times, because that allows efficient sampling from the stationary distribution. In the context of John's post, we're instead interested in mixing times because redundant information is a cause of long (or even infinite) mixing times.

Information Decompositions and Redundancy

John told us that he is now also interested in "*relative*" redundant information: for n random variables $X_1,\ldots,X_n$, what information do they redundantly share *about* a target variable Y?

One well-known approach for this is partial information decomposition. For the special case of two source variables X_1,X_2 and one target variable Y, the idea is to find a decomposition of the mutual information $I(X_1,X_2;Y)$ into:

- Redundant information $RI(X_1,X_2;Y)$ that X_1 and X_2 *both* contain about Y;
- Unique information terms $UI(X_1 \setminus X_2;Y)$ and $UI(X_2 \setminus X_1;Y)$ of information that *only one* one of the variables contains about Y;
- Synergistic information $SI(X_1,X_2;Y)$ that X_1 and X_2 only *together* contain about Y.

The original paper also contains a concrete definition for redundant information, called I_{min}. Later, researchers studied further desirable axioms that a redundancy measure should satisfy. However, it was proven that they can't all be satisfied simultaneously, which led to a development of many more attempts to define redundant information.

John told us that he does not consider partial information decomposition useful for his purposes since it considers small systems (instead of systems in the limit of large n), for which he does not expect there exist formalizations of redundancy that have the properties we want.

Neuroscience

Neuroscience can provide evidence about "how natural" abstractions are between different species of animals. Jan Kirchner has written a short overview of some of the existing work in this field:

> Similarities in structure and function abound in biology; individual neurons that activate exclusively to particular oriented stimuli exist in animals from drosophila (Strother et al. 2017) via pigeons (Li et al. 2007) and turtles (Ammermueller et al. 1995) to macaques (De Valois et al. 1982). The universality of major functional response classes in biology suggests that the neural systems underlying information processing in biology might be highly stereotyped (Van Hooser, 2007, Scholl et al. 2013). In line with this hypothesis, a wide range of neural phenomena emerge as optimal solutions to their respective functional requirements (Poggio 1981, Wolf 2003, Todorov 2004, Gardner 2019). Intriguingly, recent studies on artificial neural networks that approach human-level performance reveal surprising similarity between emerging representations in both artificial and biological brains (Kriegeskorte 2015, Yamins et al. 2016, Zhuang et al. 2020).
>
> Despite the commonalities across different animal species, there is also substantial variability (Van Hooser, 2007). One prominent example of a functional neural structure that is present in some, but absent in other, animals is the orientation pinwheel in the primary visual cortex (Meng et al. 2012), synaptic clustering with respect to orientation selectivity (Kirchner et al. 2021), or the distinct three-layered cortex in reptiles (Tosches et al. 2018). These examples demonstrate that while general organization principles might be universal, the details of how exactly and where in the brain the principles manifest is highly dependent on anatomical factors (Keil et al. 2012, Kirchner et al. 2021), genetic lineage (Tosches et al. 2018), and ecological factors (Roeth et al. 2021). Thus, the universality hypothesis as applied to biological systems does not imply perfect replication of a given feature across all

instances of the system. Rather, it suggests that there are broad principles or abstractions that underlie the function of cognitive systems, which are conserved across different species and contexts.

(Cognitive) Psychology

Similarities of representations between different individuals or cultures is an important topic in psychology (e.g. *psychological universals*—mental properties shared by all humans instead of just specific cultures). Also potentially interesting is research on basic-level categories—concepts at a level of abstraction that appears to be especially natural to humans. Of course similarities between human minds can only provide weak evidence in favor of universally convergent abstractions for *all* minds. Psychology might be more helpful to find evidence *against* the universality of certain abstractions.

Philosophy

Philosophy discusses natural kinds—categories that correspond to real structure in the world, as opposed to being human conventions. Whether natural kinds exist (and if so, which kinds are and are not natural) is a matter of debate.

The universality hypothesis is similar to a naturalist position: natural kinds exist, many of the categories we use are not arbitrary human conventions but rather follow the structure of nature. It's worth noting that in the universality hypothesis, human-made things can form natural abstractions too. For example, cars are probably a natural abstraction in the same way that trees are. Whether artifacts like cars can be natural kinds is disputed among philosophers.

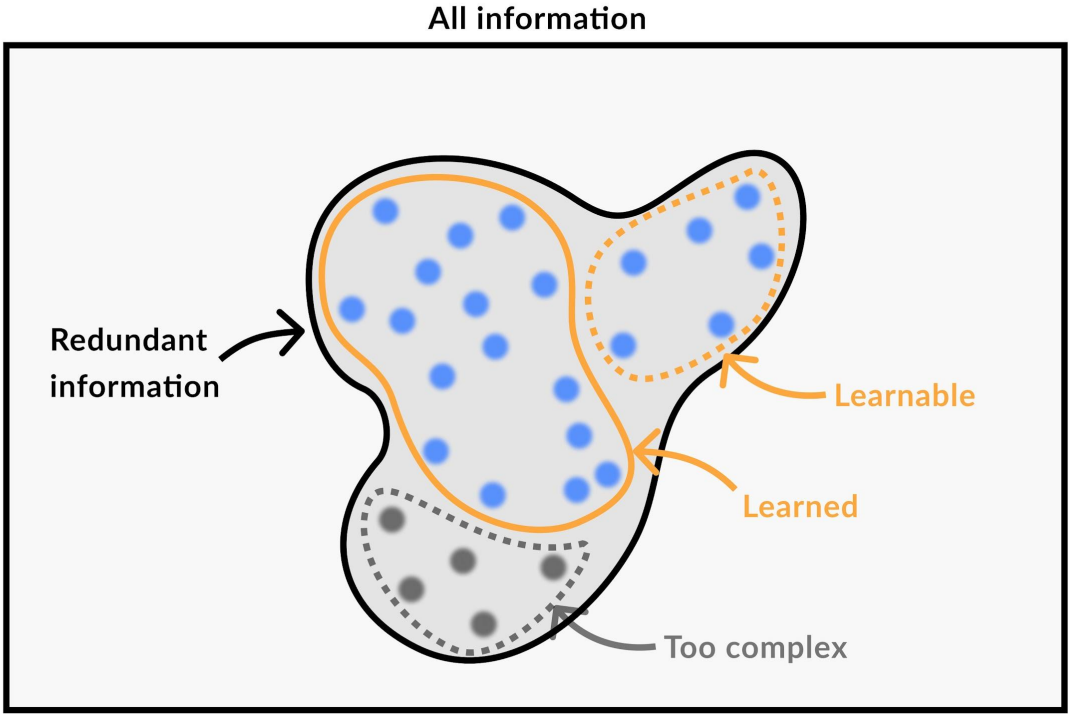
Key high-level claims

Broadly speaking, the natural abstractions agenda makes two main claims that are sometimes conflated:

1. **The Universality Hypothesis:** Natural abstractions exist, i.e. many cognitive systems learn similar abstractions.
2. **The Redundant Information Hypothesis:** Natural abstractions are well described mathematically as functions of redundant or conserved information.

Throughout the rest of the piece, we use the term *natural abstraction* to refer to the general concept, and *redundant information abstractions* to refer to the mathematical construct.

In this section, we'll break those two high-level claims down into their subclaims. Many of those subclaims are about various sets of information and how they are related, so we summarize those in the figure below.



Overview of natural abstractions: out of all the information about a system, we are interested in the *redundantly represented* information. Natural abstractions (blue/gray dots) are functions of this redundant information and form a discrete set. A cognitive system might not be able to learn some natural abstractions simply because they are too complex (gray dots). Other than that, general cognitive systems like humans or AGIs can learn the same natural abstractions (blue dots), though in practice they might not learn abstractions that aren't relevant to them. This figure is only meant as a visual overview, see the subsections below for some subtleties (e.g. on discreteness).

0. Abstractability: Our universe abstracts well

An important background motivation for this agenda is that our universe allows good abstractions at all. While almost all abstractions are leaky to some extent, there are many abstractions that work quite well even though they are vastly smaller than reality (recall the example of abstracting a circuit from electrons moving around to idealized logical computations).

Some version of this high-level claim is uncontentious, but it's an important part of the worldview underlying the natural abstractions agenda. Note that John has used the term "abstractability" to mean something a bit more specific, namely that good abstractions are connected to information relevant far away. We will discuss this as a separate claim later (Claim 2d).

1. The Universality Hypothesis: Most cognitive systems learn and use similar abstractions

1a. Most cognitive systems learn subsets of the same abstractions

Cognitive systems are much smaller than the universe, so they can't track all the low-level information anyway—they will certainly have to abstract in *some* way.

A priori, you could imagine that basically "anything goes" when it comes to abstractions: every cognitive system throws away different parts of the available information. Humans abstract CPUs as logical circuits, but other systems use entirely different abstractions.

This claim says that's not what happens: there is some relatively small set of information that a large class of cognitive systems learn a subset of. In other words, the vast majority of information is not represented in *any* of these cognitive systems.

As another example, consider a rotating gear. Different cognitive systems may track different subsets of its high-level properties, such as its angular position and velocity, its mass, or its temperature. But there is a lot of information that none of them track, such as the exact thermal motion of a specific atom inside the gear.

Precisely which cognitive systems are part of this large class is not yet clear. John's current hypothesis is "distributed systems produced by local selection pressures".

1b. The space of abstractions used by most cognitive systems is roughly discrete

The previous claim alone is not enough to give us crisp, "natural" abstractions. As a toy example, you could have a system that tracks a gear's rotational velocity ω and its temperature T, but you could also have one that *only* tracks the combined quantity $\omega^{\alpha} \cdot T^{\beta}$ for some real numbers α,β. Varying α and β smoothly would give a continuous family of abstractions, each keeping slightly different pieces of information.

According to this claim, there is instead a specific, approximately discrete set of abstractions that are actually used by most cognitive systems. These abstractions are what we call "natural abstractions". Rotational velocity and temperature are examples of natural abstractions of a gear, whereas arbitrary combinations of the two are not.

One caveat is that we realistically shouldn't expect natural abstractions to be *perfectly* discrete. Sometimes, slightly different abstractions will be optimal for different cognitive systems, depending on their values and environment. So there will be some ambiguity around some natural abstractions. But the claim is that this ambiguity is very small, in particular small enough that different natural abstractions don't just blend into each other. (See this comment thread for more discussion.)

1c. Most general cognitive systems can learn the same abstractions

The claims so far say that there is a reasonably small, discrete set of "natural abstractions", which a large class of cognitive systems learn a subset of. This would still leave open the possibility that these subsets don't overlap much, e.g. that an AGI might use natural abstractions we simply don't understand.

Clearly, there are cases where an abstraction is learned by one system but not another one. For example, someone who has never seen snow won't have formed the "snow" abstraction. However, if that person does see snow at some later point in their life, they'll learn the concept from only very few examples. So they have the *ability* to learn this natural abstraction as soon as it becomes relevant in their environment.

This claim says that this ability to learn natural abstractions applies more broadly: general-purpose cognitive systems (like humans or AGI) can in principle learn all natural abstractions. If this is true, we should expect abstractions by future AGIs to not be "fundamentally alien" to us. One caveat is that larger cognitive systems may be able to track things in more detail than our cognition can deal with.

1d. Humans and ML models both use natural abstractions

This claim says that humans and ML models are part of the large class of cognitive systems that learn to use natural abstractions. Note that there is no claim to the converse: not all natural abstractions are used by humans. But given claim 1c, once we do encounter the thing described by some natural abstraction we currently don't use, we will pick up that natural abstraction too, unless it is too complex for our brain.

John calls the human part of this hypothesis Human-Compatibility. His writing doesn't mention ML models as much, but the assumption that they will use natural abstractions is important for the connection of this agenda to AI alignment.

2. The Redundant Information Hypothesis: A mathematical description of natural abstractions

2a. Natural abstractions are functions of redundantly encoded information

Claim 1a says there is some small set of information that contains all natural abstractions, and claim 1b says that natural abstractions themselves are a discrete subset of this set of information. This claim describes the set of information from 1a: it is all the information that is encoded in a highly redundant way. Intuitively, this means you can get it from many different parts of a system.

An example (due to John) is the rotational velocity of a gear: you can estimate it based on any small patch of the gear by looking at the average velocity of all the atoms in that patch and the

distance of the patch to the rotational axis. In contrast, the velocity of one single atom is not very redundantly encoded: you can't reconstruct it based on some other far-away patch of the gear.

This claim says that all natural abstractions are functions of redundant information, but it does *not* say that all functions of redundant information are natural abstractions. For example, since both angular velocity ω and temperature T of a gear are redundantly encoded, mixed quantities such as $\omega^{\alpha} \cdot T^{\beta}$ are functions of redundant information, but this does not make them natural abstractions.

2b. Redundant information can be formalized via resampling or minimal latents

The concept of redundant information as “information that can be obtained from many different pieces of the system” is a good intuitive starting point, but John has also given more specific definitions. Later, we will formalize these definitions a bit more, for now we only mean to give a high-level overview. Note that John told us that his confidence in this claim specifically is lower than in most of the other claims.

Originally, John defined redundant information as information that is conserved under a certain resampling process (essentially Gibbs sampling): given initial samples of variables $X_1,\dots,X_n$, you repeatedly pick one of the variables at random and resample it conditioned on the samples of all the other variables. The information that you still have about the original variable values after resampling many times must have been redundant, i.e. contained in at least two variables. In practice, we probably don’t want such a loose definition of redundancy: what we care about is information that is *highly* redundant, i.e. present in many variables. This means we would resample several variables at a time.

In a later post, John proposed another potential formalization for natural abstractions, namely the *minimal latent variable* conditioned on which $X_1,\dots,X_n$ are all independent. He argues that these minimal latent variables only depend on the information conserved by resampling (see below for our summary of the argument).

2c. In our universe, most information is not redundant

If most of the information in our universe was encoded highly redundantly, then claim 2a (natural abstractions are functions of redundant information) wouldn't be surprising. The additional claim that most information is *not* redundant is what makes 2a interesting. This is a more formal version of the background claim 0 that “our universe abstracts well”.

2d. Locality, noise, and chaos are the key mechanisms for most information not being redundant

Claim 2c raises a question: why should most information be non-redundant? This claim says the reason is roughly as follows:

- Interactions in our universe are local. For a piece of information to be redundantly represented in many places, it needs to be mediated by many layers in between.
- Transmission of most information is noisy: at each step, some information is lost due to influences from other variables that we aren't tracking. So over long distances, most information is lost. Due to chaos, this happens quite quickly (or equivalently, the “long” distances only need to be moderately long).

A closely related claim is that the information which *is* redundantly represented must have been transmitted very faithfully, i.e. close to deterministically. Conversely, information that is transmitted faithfully is redundant, since it is contained in every layer.

Key Mathematical Developments and Proofs

(This section is more mathematically involved than the rest of the post. If you like, you can skip to the next section and still follow most of the remaining content.)

In this section, we describe the key mathematical developments from the natural abstractions program and describe how they all relate to redundant information. We start by formulating the telephone theorem, which is related to abstractions as information "relevant at a distance". Afterward, we explain in more detail how redundant information can be defined as resampling-invariant information, and describe why information at a distance is expected to be a function of redundant information. We continue with the definition of abstraction as minimal latent variables and why they are *also* expected to be functions of redundant information. All of this together supports claims 2a and 2b from earlier.

Finally, we discuss the generalized Koopman-Pitman-Darmois theorem (KPD) and how it was originally conjectured to be connected to redundant information. Note that based on private communication with John, it is currently unclear how relevant generalized KPD is to abstractions.

This section is meant to strike a balance between formalization and ease of exposition, so we only give proof sketches here. The full definitions and proofs for the telephone theorem and generalized KPD can be found in our accompanying pdf. We will discuss on a more conceptual level how the results here fit together later.

Epistemic status: *We have carefully formalized the proofs of the telephone theorem and the generalized KPD theorem, with only some regularity conditions to be further clarified for the latter. For the connection between redundant information and the telephone theorem, and also the minimal latents approach, we present our understanding of the original arguments but believe that there is more work to be done to have precisely formalized theorems and proofs. We note some of that work in the appendix.*

The Telephone Theorem

An early result in the natural abstractions agenda was the telephone theorem, which was proven before the framework settled on redundant information. In this theorem, the abstractions are defined as limits of minimal sufficient statistics along a Markov chain, which we now explain in more detail:

A sufficient statistic of a random variable Y for the purpose of predicting X is, roughly speaking, a function f(Y) that contains all the available information for predicting X:

$P(X|Y)=P(X|f(Y))$.

If X and Y are variables in the universe and very "distant" from each other, then there is usually not much predictable information available, which means that f(Y) can be "small" and might be thought of as an "abstraction".

Now, the telephone theorem describes how these summary statistics behave along a Markov chain when chosen to be "minimal". For more details, especially about the proof, see the accompanying pdf.

Theorem (The telephone theorem**).** *For any Markov chain* $X_0 \to X_1 \to \dots$ *of random variables* $X_t:\Omega \to X_i$ *that are either discrete or absolutely continuous, there exists a sequence of measurable functions* $f_1, f_2, \dots$*, where* $f_t:X_i \to R_{X_0(\Omega)}$*, such that:*

- $f_t(X_t)$ *converges in probability to some random variable* f_∞*, and*
- *for all* $t, P(X_0|X_t)=P(X_0|f_t(X_t))$ *pointwise on* Ω *(so* $f_t(X_t)$ *is a sufficient statistic of* X_t *for the purpose of predicting* X_0*).*

Concretely, we can pick $f_t(X_t):=P(X_0|X_t)$ *as the* minimal *sufficient statistic.*

Proof sketch. $f_t(X_t):=P(X_0|X_t)$ can be viewed as a random variable on Ω mapping $\omega \in \Omega$ to the conditional probability distribution

$P(X_0|X_t=X_t(\omega)) \in R_{X_0(\Omega)}$.

Then clearly, this satisfies the second property: if you know how to predict X_0 from the (unknown) $X_t(\omega)$, then you do just as well in predicting X_0 as if you know $X_t(\omega)$ itself:

$P(X_0|X_t(\omega))=P(X_0|P(X_0|X_t=X_t(\omega)))=P(X_0|f_t(X_t)=f_t(X_t(\omega)))$

For the first property, note that the mutual information $I(X_0;X_t)$ decreases across the Markov chain, but is also bounded from below by 0 and thus eventually converges to a limit information I_∞. Thus, for any $\epsilon>0$, we can find a T such that for all $t \geq T$ and $k \geq 0$ the differences in mutual information are bounded by ϵ:

$\epsilon > |I(X_0;X_t)-I(X_0;X_{t+k})|=|I(X_0;X_t,X_{t+k})-I(X_0;X_{t+k})|=|I(X_0;X_t|X_{t+k})|$.

In the second step, we used that $X_0 \to X_t \to X_{t+k}$ forms a Markov chain, and the final step is the chain rule of mutual information. Now, the latter mutual information is just a KL divergence:

$$D_{KL}(P(X_0, X_t \mid X_{t+k}) \;//\; P(X_0 \mid X_{t+k}) \cdot P(X_t \mid X_{t+k})) < \epsilon.$$

Thus, "approximately" (with the detailed arguments involving the correspondence between KL divergence and total variation distance) we have the following independence:

$$P(X_0, X_t \mid X_{t+k}) \approx P(X_0 \mid X_{t+k}) \cdot P(X_t \mid X_{t+k}).$$

By the chain rule, we can also decompose the left conditional in a different way:

$$P(X_0, X_t \mid X_{t+k}) = P(X_0 \mid X_t, X_{t+k}) \cdot P(X_t \mid X_{t+k}) = P(X_0 \mid X_t) \cdot P(X_t \mid X_{t+k}),$$

where we have again used the Markov chain $X_0 \to X_t \to X_{t+k}$ in the last step. Equating the two expansions of the conditional and dividing by $P(X_t \mid X_{t+k})$, we obtain

$$f_t(X_t) = P(X_0 \mid X_t) \approx P(X_0 \mid X_{t+k}) = f_{t+k}(X_{t+k}).$$

By being careful about the precise meaning of these approximations, one can then show that the sequence $f_t(X_t)$ indeed converges in probability. □

Abstractions as Redundant Information

The following is a semiformal summary of Abstractions as Redundant Information. We explain how to define redundant information as resampling-invariant information and why the abstractions f_∞ from the telephone theorem are expected to be a function of redundant information.

More Details on Redundant information as resampling-invariant information

The setting is a collection $X_1, \dots, X_N$ of random variables. The idea is that redundantly encoded information should be recoverable even when repeatedly resampling individual variables. This is, roughly, formalized as follows:

Let $X_0 = X_1, \dots, X_N$ be the original collection of variables and denote by $X_1, X_2, \dots, X_t, \dots$ collections of variables $X_{t1}, \dots, X_{tN}$ that iteratively emerge from the previous time step $t-1$ as follows: choose a resampling index $i \in \{1, \dots, N\}$, keep the $N-1$ variables $X_{t-1 \neq i}$ fixed and resample the remaining variable $X_{t-1 i}$ conditioned on the fixed variables. The index i of the variable to be resampled is thereby (possibly randomly) changed for each time step t. As discussed in the related work section, this is essentially Gibbs sampling.

Let X_∞ be the random variable this process converges to.[3] Then the *amount* of redundant information in X_0 is defined to be the mutual information between X_0 and X_∞:

$$\text{RedInfo}(X_0) := \text{MI}(X_0; X_\infty).$$

Ideally, one would also be able to mathematically construct an object that *contains* the redundant information. One option is to let F be a sufficient statistic of X_0 for the purpose of predicting X_∞:

$P(X_\infty|X_0)=P(X_\infty \mid F(X_0))$.

Then one indeed obtains $\text{RedInfo}(X_0)=\text{MI}(F(X_0);X_\infty)$. Concretely, one can choose $F(X_0):=P(X_\infty|X_0)$, which is a minimal sufficient statistic as explained in the above proof-sketch of the telephone theorem.

Telephone Abstractions are a Function of Redundant Information

Imagine that we "cluster together" some of the variables X_{0i} into variables $B_1,B_2,\dots$ that together form a Markov chain $B_1\to B_2\to\dots$. Each B_j contains possibly several of the variables X_{0i} in a non-overlapping way and such that the Markov chain property holds. [One example often used by John] is that the variables B_j form a sequence of growing Markov blankets in a causal model of variables X_{0i}. For all $j<k$, all the information in B_j then has to pass through all intermediate blankets to reach B_k, which results in the Markov chain property. Then from the telephone theorem one obtains an "abstract summary" of B_1 given by a limit variable f_∞.

Now, let $F(X_0)$ be the variable containing all the redundant information from earlier. Then the claim is that this contains f_∞ for *any* choice of a Markov chain $B_1\to B_2\to\dots$ above, i.e., $f_\infty=G(F(X_0))$ for some suitable function G.

Theorem (Informal). *We have* $f_\infty=G(F(X_0))$ *for some function* G *that depends on the choice of the Markov chain* $B_1\to B_2\to\dots$

Proof Sketch. Note that we did not formalize this proof sketch and thus can't be quite sure that this claim can be proven (see appendix for some initial notes). The original proof does not contain many more details than our sketch.

The idea is that $F(X_0)$ contains all information that is invariant under resampling. Thus, it is enough to show that f_∞ is invariant under resampling as well. Crucially, if you resample a variable X_i, then this will either not be contained in any of the variables $B_1,B_2,\dots$ at all, which leaves f_∞ invariant, or it will be contained in only one variable B_j. But for $T>j$, the variable B_T is kept fixed in the resampling and we have $\lim_{T\to\infty} f_T(B_T)=f_\infty$ by the construction of f_∞ detailed in the telephone theorem. Thus, f_∞ remains invariant in this process. □

Minimal Latents as a Function of Redundant Information

Another approach is to define abstractions by a minimal latent variable, i.e., the "smallest" function $\Lambda_*(X_0)$ that makes all the variables in X_0 conditionally independent:

$P(X_0|\Lambda_*)=\prod_{i=1}^{n} P(X_{0i}|\Lambda_*)$.

To be the "smallest" of these functions means that for *any other* random variable Λ with the independence property, Λ_* only contains information about X_0 that is also in Λ, meaning one has the following Markov chain:

$$\Lambda_* \to \Lambda \to X_0.$$

How is Λ_* connected to redundant information? Note that $X_{0\neq i}$ is, for each i, *also* a variable making all the variables in X_0 conditionally independent, and so Λ_* fits due to its minimality (by definition) in a Markov chain as follows:

$$\Lambda_* \to X_{0\neq i} \to X_0.$$

But this means that Λ_* will be preserved when resampling any one variable in X_0, and thus, Λ_* contains only redundant information of X_0. Since $F(X_0)$ contains *all* redundant information of X_0, we obtain that $\Lambda_*=G(F(X_0))$ for some function G. This is an informal argument and we would like to see a more precise formalization of it.

The Generalized Koopman-Pitman-Darmois Theorem

This section describes the generalized Koopman-Pitman-Darmois theorem (gKPD) on a high level. The one-sentence summary is that *if* there is a low-dimensional sufficient statistic of a sparsely connected system $X=X_1,\dots,X_n,$, then "most" of the variables in the distribution P(X) should be of the exponential family form. This would be nice since the exponential family has many desirable properties.

We will first formulate an almost formalized version of the theorem. The accompanying pdf contains more details on regularity conditions and the spaces the parameters and values "live" in. Afterward, we explain what the hope was for how this connects to redundant information, as described in more detail in Maxent and Abstractions. John has recently told us that the proof for this maxent connection that he hoped to work out according to his 2022 plan update is incorrect and that he currently has no further evidence for it to be true in the stated form.

An almost formal formulation of generalized KPD

We formulate this theorem in slightly more generality than in the original post to reveal the relevant underlying structure. This makes it clear that it applies to both Bayesian networks (already done by John) and Markov random fields (not written down by John, but an easy consequence of his proof strategy).

Let $X=X_1,\dots,X_n$ be a collection of continuous random variables. Assume that its joint probability distribution factorizes when conditioning on the model parameters Θ, e.g. as a Bayesian network or Markov random field. Formally, we assume there is a finite index set I and neighbor sets $N_i\subseteq\{1,\dots,n\}$ for $i\in I$, together with potential functions $\psi_i>0$, such that

$$P(X\mid\Theta)=\prod_{i\in I}\psi_i(X_{N_i}\mid\Theta).$$

Here, $X_{N_i} := (X_j)_{j \in N_i}$.

This covers both the case of Bayesian networks and Markov random fields:

- If X forms a Bayesian network according to a directed acyclic graph G, then $I = \{1,\dots,n\}$ and $N_i = \{i, pa_i\}$, where pa_i are the indices of parents of the variable X_i in the graph G.
- If X forms a Markov random field according to a (non-directed) graph G, then the Hammersley-Clifford Theorem shows that I can be chosen to be the set of maximal cliques C in the graph, and $N_C = C$ for all maximal cliques C.

Assume that we also have a prior $P(\Theta)$ on model parameters. Using Bayes rule, we can then also define the posterior $P(\Theta \mid X)$.

Now, assume that there is a sufficient statistic G of X with values in $\mathbb{R}^D$ for $D \ll n$. As before, to be a sufficient statistic means that it summarizes all the information contained in the data that is useful for predicting the model parameters:

$P(\Theta \mid X) = P(\Theta \mid G(X))$.

The generalized KPD theorem says the following:

Theorem (generalized KPD (almost formal version)**).** *There is:*

- *a dimension* $K \leq D$*;*
- *a set* $E \subseteq I$ *of "exceptions" that is reasonably "small";*
- *functions* $g_i, i \in I \setminus E$ *mapping to* $\mathbb{R}^K$*;*
- *a function* U *mapping to* $\mathbb{R}^K$*;*
- *and a function* h *mapping to* $\mathbb{R}_{\geq 0}$*;*

such that the distribution $P(X \mid \Theta)$ *factorizes as follows:*

$$P(X \mid \Theta) = \frac{1}{Z(\Theta)} \cdot e^{[U(\Theta)^T \sum_{i \notin E} g_i(X_{N_i})]} \cdot h(X_{\overline{N_E}}) \cdot \prod_{i \in E} \psi_i(X_{N_i} \mid \Theta).$$

Thereby, $\overline{E} := I \setminus E$ *and* $N_{\overline{E}} := \bigcup_{i \in \overline{E}} N_i$*.* $Z(\Theta)$ *is thereby a normalization constant determined by the requirement that the distribution integrates to 1.*

Proof: see our pdf appendix.

The upshot of this theorem is as follows: from the existence of the low-dimensional sufficient statistic, one can deduce that $P(X \mid \Theta)$ is roughly of exponential family form, with the factors ψ_i with $i \in E$ being the "exceptions" that cannot be expressed in simpler form. If $D \ll n$ and if each N_i is also small, then it turns out that the number of exception variables $|N_E|$ is overall small compared to n, meaning the distribution may be easy to work with.

The Speculative Connection between gKPD and Redundancy

As stated earlier, Maxent and Abstractions tries to connect the generalized KPD theorem to redundancy, and the plan update 2022 is hopeful about a proof. According to a private conversation with John, the proof turned out to be wrong. Let us briefly summarize this:

Let X factorize according to a sparse Bayesian network. Then, by replacing X with X_∞, Θ with X_0 and $G(X_\infty)$ with the resampling-invariant information $F(X_\infty)$ in the setting of the generalized KPD theorem, one can hope that:

- $F(X_\infty)$ is low-dimensional;
- $P(X_\infty \mid X_0)$ is also a sparse Bayesian network.

With these properties, one could apply generalized KPD. The second property relies on the proposed factorization theorem whose proof is, according to John, incorrect. He told us that he currently believes that not only the proof of the maxent form is incorrect, but that there is an 80% chance of the whole statement being wrong.

How is the natural abstractions agenda relevant to alignment?

We've discussed the key claims of the natural abstractions agenda and the existing theoretical results. Now, we turn to the bigger picture and attempt to connect the claims and results we discussed to the overall research plan. This section represents our understanding of John's views and there are places where we disagree—we will discuss those in the next section.

Four reasons to work on natural abstractions

We briefly discussed why natural abstractions might be important for alignment research in the Introduction. In this section, we will describe the connection in more detail and break it down into four components.

An important caveat: part of John's motivation is simply that abstractions seem to be a core bottleneck to various problems in alignment, and that connections beyond the four we list could appear in the future. So you can view the motivations we describe as the current key *examples* for the centrality of abstractions to alignment.

1. The Universality Hypothesis being true or false has strategic implications for alignment

If the Universality Hypothesis is true, and in particular if humans and AI systems both learn similar abstractions, this would make alignment easier in important ways. It would also have implications about which problems should be the focus of alignment research.

In an especially fortunate world, *human values* could themselves be natural abstractions learned by most AI systems, which would mean that even very simple hacky alignment schemes might work. More generally, if human values are represented in a simple way in most advanced AI systems, alignment mainly means pointing the AI at these values (for example by retargeting the search). On the other hand, if human values aren't part of the AI's ontology by default, viewing alignment as just "pointing" the AI at the right concept is a less appropriate framing.

Even if human values themselves turn out not to be natural abstractions, the Universality Hypothesis being true would still be useful for alignment. AIs would at least have simple internal representations of many human concepts, which should make approaches like interpretability much more likely to succeed. Conversely, if the Universality Hypothesis is false and we don't expect AI systems to share human concepts by default, then we may for example want to put more effort into *making* AI use human concepts.

2. Defining abstractions is a bottleneck for agent foundations

When trying to define what it means for an "agent" to have "values", we quickly run into questions involving abstractions. John has written a fictional dialogue about this: we might for example try to formalize "having values" via utility functions—but then what are the inputs to these utility functions? Clearly, human values are not directly a function of quantum wavefunctions—we value higher-level things like apples or music. So to formally talk about values, we need some account of what "higher-level things" are, i.e. we need to think about abstractions.

3. A formalization of abstractions would accelerate alignment research

For many central concepts in alignment, we currently don't have robust definitions ("agency", "search", "modularity", …). It seems plausible these concepts are themselves natural abstractions. If so, a formalization of natural abstractions could speed up the process of finding good formalizations for these elusive concepts. If we had a clear notion of what counts as a "good definition", we could easily check any proposed definition of "agency" etc.—this would give us a clear and generally agreed upon paradigm for evaluating research.

This could be helpful to both agent foundations research (e.g. defining agency) and to more empirical approaches (e.g. a good definition of modularity could help understand neural networks).

Many of these abstractions in alignment seem closer to *mathematical abstractions*. These are not directly covered by the current work on natural abstractions. However, we might hope that ideas will transfer. Additionally, if mathematical abstractions are instantiated, they might become ("physical") natural abstractions. For example, the Fibonacci sequence is clearly a mathematical concept, but it also occurs very often in nature so you might use it simply to compactly describe our world. Similarly, perhaps modularity is a natural abstraction when describing different neural networks.

4. Interpretability

In John's view, the main challenge in interpretability is robustly identifying which things in the real world the internals of a network correspond to (for example that a given neuron *robustly* detects trees and nothing else). Current mechanistic interpretability research tries to find readable "pseudocode" for a network but doesn't have the right approach to find these correspondences according to John:

> I think a lot of the interpretability crowd hasn't yet fully internalized the framing of "interpretability is primarily about mapping net-internal structures to corresponding high-level interpretable structures in the environment". In particular I think a lot of interpretability researchers have not yet internalized that mathematically understanding what kinds of high-level interpretable structures appear in the environment is a core part of the problem of interpretability. You have to interpret the stuff-in-the-net as something, and it's approximately-useless if the thing-you-interpret-stuff-in-the-net-as is e.g. a natural-language string without any legible mathematical structure attached, or an ad-hoc mathematical structure which doesn't particularly cut reality at the joints.

A theory of abstractions would address this problem: natural abstractions are exactly about figuring out a good mathematical description for high-level interpretable structures in the environment. Additionally, knowing the "type signature" of abstractions would make it easier to find crisp abstractions inside neural networks: we would know more precisely what we are looking for.

We don't have a good understanding of parts of this perspective (or disagree with our understanding of it)—we will discuss that more in the Discussion section.

How existing results fit into the larger plan

John developed the theoretical results we discussed above, such as the Telephone theorem, in the context of his plan to empirically test the natural abstraction hypothesis. Quoting him:

> The natural abstraction hypothesis is mainly an empirical claim, which needs to be tested in the real world.

In this section, we'll mainly explain how the plan to do these empirical tests led to all the theoretical work John has done on abstractions. But we also briefly want to note that a lot of this work could alternatively be motivated as simply trying to formalize and better understand natural abstractions, which is connected to all of the four motivations we just described. We focus on the angle of empirical tests (i.e. motivation 1) because this was the reasoning John originally gave, and because it is perhaps least obvious how it is connected to his work.

To run empirical tests of the natural abstraction hypothesis, it would be nice to have a tool that can find the abstractions in a given system. For example, we could use this tool to check

whether different ML systems learn the same abstractions and whether those abstractions are the same ones humans use. “Abstractions” in this context refer to redundant information or information at a distance. Overall, these experiments could test aspects of both the Universality Hypothesis and the Redundant Information Hypothesis.

There is a problem: naively computing the information at a distance or redundant information is computationally intractable. Example by John:

> Even just representing abstractions efficiently is hard - we’re talking about e.g. the state-distribution of a bunch of little patches of wood in some chunk of a chair given the state-distribution of some other little patches of wood in some other chunk of the chair. Explicitly writing out that whole distribution would take an amount of space exponential in the number of variables involved; that would be a data structure of size roughly O((# of states for a patch of wood)^(# of patches)).

The theoretical work John did can be understood as trying to develop *efficient representations* of information-at-a-distance-abstractions. The initial attempt was based on linear approximations, but that did not pan out as John himself has explained, so we won’t discuss it further.

In this context, the point of the Telephone theorem is that it narrows down the form abstractions can take and gets us closer to tractability. As John summarizes it:

> All information is either perfectly conserved or completely lost in the long run. And, more interestingly, information can only be perfectly conserved when it is carried by deterministic constraints - i.e. quantities which are exactly equal between two parts of the system.
>
> [...]
>
> Why am I excited about the Telephone Theorem? First and foremost: **finding deterministic constraints does not involve computing any high-dimensional integrals**. It just involves equation-solving/optimization - not exactly easy, in general, but much more tractable than integrals! *(highlight his*)

We are personally more skeptical about just how much the Telephone Theorem shows: the theorem itself seems much more narrow than this quote suggests (see the appendix for a more detailed discussion of this point).

The generalized KPD theorem tackles a different aspect of efficient representations of abstractions. Let’s say we have some way of finding the natural abstractions, e.g. by looking for deterministic constraints as in the Telephone theorem. Then far-away low-level parts of the system should be independent conditional on this abstraction. But even if the abstraction itself is simple, the distribution of these low-level parts given the abstraction could still be quite complicated a priori. The gKPD theorem could be a way to show that, instead, the distribution of low-level parts is an exponential family distribution, which is easier to handle. While the gKPD

theorem is suggestive of such a result, there is currently no formal theorem. In May 2022, John wrote a post giving an overview of some heuristic arguments for abstractions inducing exponential family distributions. In his 2022 Plan update, he mentioned a proof, but based on private communication it seems that proof didn't work after all and it's currently less clear how helpful the gKPD results are for natural abstractions.

The redundant information and minimal latent results can be understood as making natural abstractions less reliant on a local graph structure. The Telephone theorem requires some notion of "far away", defined by a choice of Markov blankets. Which abstraction you get depends on these Markov blankets. In contrast, the resampling definition of redundant information defines natural abstractions based only on a joint distribution over some variables. If these variables happen to form a causal graph, then a Telephone-like result holds *for the redundant information abstraction*: far away parts are independent given the abstraction for *any* choice of Markov blankets (see our earlier math section). John also told us about a new version of the Telephone theorem that gets rid of any requirement of local graph structure. That result is not yet published and we won't discuss it as much, though see the appendix for a sketch.

Finally, the theoretical results provide some evidence for Claim 2a (natural abstractions are functions of redundant information). Specifically, information at a distance and minimal latents both are intuitively plausible guesses for properties that good abstractions might have. The fact that they both end up being contained by redundant information (another intuitive guess) is promising.

Selection theorems

In parallel to the natural abstractions agenda, John is also working on the selection theorems agenda. Briefly, selection theorems are theorems of the form "a system under selection pressure X will develop property Y". The selection pressure could be natural selection or a machine learning training setup, and the property could be something like "the system has a world model" or "the system behaves like an expected utility maximizer". We won't discuss selection theorems in general here, but will highlight a connection to natural abstractions. Namely, one selection theorem we can hope for is that many cognitive systems use natural abstractions. This is a *theoretical* approach to testing the Universality Hypothesis, as opposed to empirical tests discussed in the previous subsection. In this aspect, the selection theorems agenda and natural abstractions agenda can thus support each other: proving such a selection theorem would give clarity about natural abstractions, and conversely having a good theory of what natural abstractions even are should make it easier to state and prove such a selection theorem.

Discussion, limitations, and critiques

The previous sections were our attempt to explain the natural abstractions agenda mostly without introducing our opinions. Now we instead discuss our own views on the agenda. We start by outlining some key pieces that we think are currently missing in the theory of natural abstractions—John might agree with these but they aren't discussed as much as we think they should be. Second, we discuss the connections between natural abstractions and alignment that we described in the previous section. We conclude with some meta-level critiques about research methodology.

Note that our discussion of current limitations is based on published work. We know John is thinking about a few of these points already (and he might have thoughts on most or all of the rest), but we still list them.

Gaps in the theory

We think there has been significant conceptual progress on natural abstractions, but that key pieces of the formalism are missing. We aren't convinced that "the core theory of natural abstractions is now 80% nailed down"—we will discuss some questions that we would consider part of the "core theory" but that remain open as far as we know.

Results don't discuss encoding/representation of abstractions

All existing results in the natural abstractions agenda are formulated in information-theoretic terms, but information theory doesn't discuss how information is represented. As an extreme example, consider a one-way permutation f, i.e. an invertible function that's easy to compute but cryptographically hard to invert. The mutual information between X and f(X) is maximal (i.e. the entropy H(X)) for any random variable X. But in practice, knowing f(X) isn't helpful for figuring out X because the necessary computations are completely intractable.

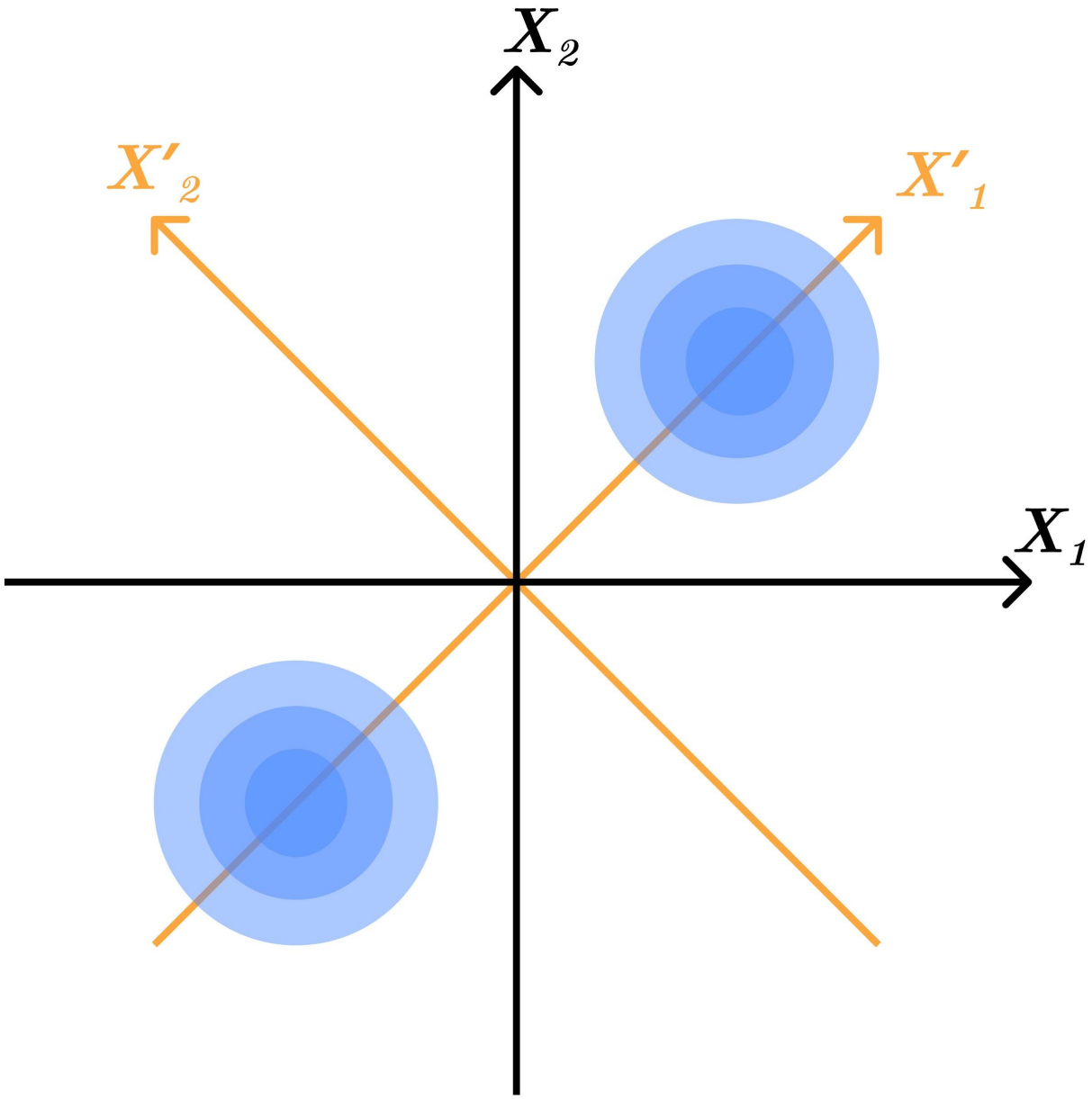
When talking about different cognitive systems "learning the same abstractions" in the Universality Hypothesis, the intuitive claim is that the abstractions will be *recognizably* the same—that it will be relatively easy to translate between them. Indeed, the common claim that the Universality Hypothesis being true would make alignment much easier relies on such an interpretation. But information theory alone doesn't seem suitable to even formally state a claim of this form. Notably, Chris Olah's formulation of the Universality Hypothesis *does* talk about universality of circuits, not just information. We think that a complete theory of natural abstractions will likewise need to consider how abstractions are represented. It may turn out that results from information theory mostly transfer (for example, there is existing work on a version of information theory that takes computational limits into account). However, it also seems very plausible that this will involve significant additional work and important changes.

Definitions depend on choice of variables X_i

All current attempts to define natural abstractions—whether via resampling, minimal latents, or information at a distance—rely on some factorization of the system into subsystems or variables

X_i. For resampling, these variables are important because we resample one variable (or some small number of variables) at a time. For minimal latents, we want to make the variables independent conditional on the abstraction. And for information at a distance, we need variables to form a Markov chain (the need for variables still exists in John's unpublished new Telephone theorem).

This wouldn't be too much of a problem if the choice of variables didn't matter much, or at least if all "reasonable" choices gave the same result. However, there are simple transformations that can completely change the redundant information in a system. A trivial example is adding copies of variables or combining several variables into one. But there are also more interesting cases, such as a simple rotation of the variable axes:



Even simple transformations of variables can completely change the redundant information: In

the original coordinates X_1, X_2 (black axes), there is 1 bit of redundant information for distinguishing the two modes of the blue distribution. Resampling one variable keeps us near the same mode. But rotating the variables by 45° (orange) removes the redundant information: resampling X'_1 can switch between modes.

For concreteness, imagine that X_1 and X_2 are the positions of two particles, so we know that either they are both are at positive positions or they are both at negative positions. From this description, this system contains redundant information. But we could equivalently specify the state of this system by giving the center of mass and the distance vector between the two particles (that's exactly the orange X' coordinate system). Now, there's no redundant information anymore! Both of these descriptions are often used in physics and it's unclear which choice of variables is the "right" one.

Perhaps in many practical settings, there is one intuitively "right" choice of variables. But this seems extremely speculative, and even if it's true, we currently don't have a good theory for extracting these "right" variables from a system.

Theorems focus on infinite limits, but abstractions happen in finite regimes

The literal mathematical results in the natural abstractions agenda often discuss some form of infinite limit. For example, the Telephone theorem only makes statements about the infinite distance limit: at any finite point, constraints may be non-deterministic.

This wouldn't be as big an issue if the abstractions we care about in practice were practically the same ones we get from the infinite limit. But we think that in practice, most interesting abstractions "live" firmly in the finite regime. Consider the example of the rotational velocity of a gear. This piece of information is relevant if you are standing a few feet away and looking at the gear. But if we increase the distance sufficiently, e.g. by considering the information relevant in a different galaxy, then the rotational velocity of this gear becomes an irrelevant low-level detail, washed out by noise. The same principle applies to distances in time rather than space. As an extreme case, if we consider information still relevant at the heat death of the universe, the gear's rotational velocity certainly doesn't qualify.

One might hope that many ideas derived from the infinite distance limit are still relevant in these finite regimes. But we think that the finite regime suggests research questions that differ from the ones that current theorems address. For example, are there clearly separable "scales" or levels of abstraction? Or can you just transition smoothly between levels?

Missing theoretical support for several key claims

While there are several theorems in the natural abstractions agenda (as we discussed above), we think it's important to remember that they don't directly support many of the key claims we identified earlier. In particular:

- There are no selection theorems implying any form of the Universality Hypothesis (Claim 1) yet.
- None of the results show any discreteness of natural abstractions (Claim 1b). In fact, the math currently only defines the entire abstraction—all the redundant information at once. Discrete "sub-abstractions" aren't discussed at all.
- The theorems don't show that redundant information abstractions are low-dimensional (Claim 2c).

To be clear, it may be better to look for empirical evidence of these claims instead of proving theorems about them! John has said himself several times that the Natural Abstraction Hypothesis ultimately needs to be tested empirically. (For example, redundant information abstractions are clearly not low-dimensional in *all possible mathematical structures*—this is a claim about *our universe*.)

On the other hand, John has also said:

> *"For most physical systems, the information relevant "far away" can be represented by a summary much lower-dimensional than the system itself."*
>
> Assuming the proofs in this post basically hold up, and the loopholes aren't critical, I think this claim is now basically proven. There's still some operationalization to be done (e.g. the "dimension" of the summary hasn't actually been addressed yet) [...]

While we strongly agree that our universe has good "low-dimensional" summaries at large distances, we disagree with this characterization of the state of the theory: given that the claim is about the low dimensionality of summaries, and this is exactly the part that the theorems don't yet address, we wouldn't call this claim "basically proven".

Overall, we think there is substantial evidence about many of the key claims from intuition and just by looking at examples. Reducing the remaining uncertainty may often best be done by empirical research. What we want to advocate against here is using the theorems as significant evidence for most of the key claims—we think whether you believe them or not should mostly be informed by other sources. To be clear, John might in fact agree with this (with the quote above being an exception), but we think it's an easy misconception for readers to develop given the informal discussion of theorems and the close connections to conceptual work. We discuss this in more detail in a case study in the appendix, using the Telephone theorem as an example.

Missing formalizations

In this post and the mathematical pdf appendix, we have presented formal statements and proofs of the Telephone theorem and the generalized KPD theorem. (In both cases, John had given a reasonably detailed proof sketch already.) However, the claims surrounding redundant information and minimal latents only have rudimentary proof sketches and in some cases only high-level intuitive arguments. We are still short of a full formalization and proofs.

Relevance to alignment

Having discussed some of the open problems not yet addressed by existing work on natural abstractions, let's zoom out and ask: how helpful is progress on natural abstractions for alignment?

In summary, we agree that the connections between abstractions and alignment outlined above are plausible, though with varying amounts of disagreement. We especially agree that the extent to which the Universality Hypothesis is true is a crucial factor for the difficulty of alignment, and to some extent for prioritization between agendas. We also strongly agree that interpretability methods need to tell us about how internal representations are connected to real-world things in order to be useful. We are more skeptical about the possibility of "accelerating all alignment research" with a formalization of abstractions, and we disagree with John about current interpretability methods. In several cases, we're also not convinced that the current direction of the natural abstractions agenda is the best approach. The rest of this section discusses these points in more detail.

Figuring out whether the Universality Hypothesis is true: This was the original stated motivation for developing the theory of natural abstractions. We agree that figuring out to what extent ML systems learn human-understandable concepts is very valuable. What we're less convinced of is that the current theoretical approach is a good way to tackle this question. One worrying sign is that almost two years after the project announcement (and over three years after work on natural abstractions began), there still haven't been major empirical tests, even though that was the original motivation for developing all of the theory. John seemed optimistic about running experiments soon in April 2021, September 2021, and December 2021. The 2022 update mentions that progress on crossing the theory-practice gap has been a bit slower than expected, though not enough that John is too worried for now. Of course sometimes experiments do require upfront theory work. But in this case, we think that e.g. empirical interpretability work is already making progress on the Universality Hypothesis, whereas we're unsure whether the natural abstractions agenda is much closer to major empirical tests than it was two years ago.[4]

Abstractions as a bottleneck for agent foundations: The high-level story for why abstractions seem important for formalizing e.g. values seems very plausible to us. It's less clear to us whether they are *necessary* (or at least a good first step). You could make a structurally similar argument about probability theory:

> "Probability theory talks about random variables, which are functions on some joint sample space. But to talk about what the type of this sample space even is, we first need measure theory."

Measure theory is indeed helpful for formalizing probability theory, but you can do a lot of very useful probability theory without it. To be clear, we don't think this is a tight enough analogy to show that the argument in favor of abstractions must be flawed, it just makes us cautious.

Overall, we agree that abstractions seem important for several concepts in alignment and that this is a good argument to study them.

Accelerating alignment research: The promise behind this motivation is that having a theory of natural abstractions will make it much easier to find robust formalizations of abstractions such as “agency”, “optimizer”, or “modularity”. This seems “big if true”: a way to find “good concepts” more quickly and reliably would be valuable for alignment research but also much more broadly applicable. A very successful version of this could amount to a paradigm for evaluating definitions in a similar way to proofs as a paradigm for evaluating certain types of claims and arguments. To us, such an outcome seems unlikely, though it may still be worth pursuing—highly ambitious projects can be very good in expectation. One specific obstacle is that many of these concepts seem more like *mathematical abstractions* than physical abstractions like “tree”. While it’s possible that ideas developed for physical abstractions will work anyway, we think that people focused on this motivation should focus much more on also understanding mathematical abstraction, until the two either converge or become clearly distinct.

Interpretability: As mentioned, we strongly agree that interpretability methods should tell us about how internal representations are connected to real-world things; we mainly disagree with John’s view of the current state of interpretability. Figuring out the real-world meaning of internal network activations is one of the core themes of safety-motivated interpretability work. And reverse-engineering a network into “pseudocode” is not just some separate problem, it’s deeply intertwined. We typically understand the *inputs* of a network, so if we can figure out how the network transforms these inputs, that can let us test hypotheses for what the meaning of internal activations is. See e.g. Zoom In: An Introduction to Circuits for many early examples of circuits being used to validate hypotheses about the meaning of neurons. It’s certainly possible that thinking about natural abstractions will at some point contribute to interpretability in concrete ways. But we don’t see the crucial missing parts in current interpretability research that John seems to be pointing at.

Concluding thoughts on relevance to alignment: While we’ve made critical remarks on several of the details, we also want to reiterate that overall, we think (natural) abstractions are an important direction for alignment and it’s good that someone is working on them! In particular, the fact that there are at least four distinct stories for how abstractions could help with alignment is promising.

Methodological critiques

We’ve discussed what we see as important missing pieces and given our opinions on the relevance of natural abstractions to alignment. We now move away from the object-level discussion to a few critiques of the research methodology in the natural abstractions agenda. We won’t justify these in too much detail because we think they can and should be discussed more generally than just in the context of this agenda. Nevertheless, we think it’s valuable to explicitly note these disagreements here.

Low level of precision and formalization

John's writing emphasizes intuition and examples over precise claims or formal proofs. This definitely has advantages, and we think it's a great choice for *first introducing* ideas to new audiences. What we would like to see more of is more precise statements and more formalism after ideas have been introduced for the first time. This is an almost universally accepted best practice in most scientific fields, and rightfully so in our view. Outlining a few reasons:

- Making precise arguments is a way to verify claims and spot mistakes. Errors in mathematical claims do happen (e.g. John told us that the first theorem in the redundant information post has an incorrect proof sketch and might be wrong). Formal proofs certainly don't protect against these entirely, but they help. (To be clear, we think that intuitive arguments *also* help figure out the truth of mathematical claims!)
- Stating claims (and proofs) precisely makes it much easier for others to point out mistakes. If a claim is stated in a way that has many slightly different formal interpretations, then giving a strong critique requires disproving each one of these versions. In contrast, a formal claim can be disproven by a single counterexample—at that point, the next step is to figure out whether the claim can be patched or not, but at least there are easy atomic steps to make progress, instead of putting the entire burden on the person trying to disprove the claim. The same principle applies to proofs vs informal proof sketches.
- Stating claims precisely makes it clearer which parts are supported by theorems and which parts are speculative interpretations or conceptual claims on top of what's been proven. With some work on the reader's part, it's also possible to figure this out based on only informal descriptions, but a cursory reading can easily lead to wrong impressions. We think this is the case for e.g. the Telephone theorem and discuss this more in the appendix.

These points apply most straightforwardly to mathematical claims and arguments, but high levels of precision are still desirable and achievable even for purely conceptual claims that are not yet at the stage where they can be entirely formalized. For example, we think our breakdown of the key claims on natural abstractions into nine subclaims clarifies several points that John's usual breakdown of the Natural Abstraction Hypothesis doesn't mention.

Few experiments

As we briefly discussed earlier, we think it's worrying that there haven't been major experiments on the Natural Abstraction Hypothesis, given that John thinks of it as mostly an empirical claim. We would be excited to see more discussion on experiments that can be done right now to test (parts of) the natural abstractions agenda! We elaborate on a preliminary idea in the appendix (though it has a number of issues).

Little engagement with existing work

As our overview of related work hopefully shows, many people have thought about concepts similar to natural abstractions before. The Universality Hypothesis in the context of interpretability research is an especially notable case.

An obvious reason to connect with these other subfields is to make use of their ideas and evidence. But explicitly discussing the relation to existing work also makes it easier for others with background knowledge in these fields to parse new content. Jacob Steinhardt wrote a good explanation of this point: stating clearly how new research is connected to existing work, and in particular which parts are meant to be new and which parts are meant to be different framings on the same idea, helps others decide what to read at all. Of course it also makes it easier for readers to incorporate the new content into their existing mental models.

Should this all be delegated?

One response to all of these points might be that it's better to divide labor: some researchers should work on generating conceptual ideas and sketches of formal results, and then others should formalize these claims, do empirical tests, and improve exposition (including connections to existing work). This is certainly something John has written about and we agree this can be great if done right (the invention of the transistor is a famous example of collaboration between people with different strengths). But for this approach to work, there need to be people actually working on each of those aspects. The Telephone theorem and generalized KPD theorem have been out for about 1.5 years and yet we are the first to provide a full formalization. In the redundant information post, John says:

> I'll handle those subtleties mainly by ignoring them and hoping a mathematician comes along to clean it up later.

So far, no mathematician has come along to clean it up. To sum up: delegating to others is a perfectly valid approach in research, but it can be hard to do and doesn't always happen automatically. In our view, researchers generally shouldn't simply rely on others to independently formalize, distill, or empirically test their ideas, at least as long as the ecosystem doesn't guarantee that this actually happens comprehensively.

Conclusion

In this work, we clarified the Natural Abstraction Hypothesis by dividing it into two main claims: the Universality Hypothesis, which states that many cognitive systems converge to learning roughly the same ("natural") abstractions, and the Redundant Information Hypothesis, which describes an approach to mathematically formalize natural abstractions. Both claims can be further broken down into more precise subclaims. This includes subclaims that tend to be mentioned less frequently, such as that the space of natural abstractions is roughly discrete. The Universality Hypothesis and Redundant Information Hypothesis both have many connections to existing academic work, as we've briefly outlined.

The theoretical results developed in the natural abstractions agenda form three clusters: the Telephone theorem, the generalized KPD theorem, and several claims surrounding redundant information (defined via resampling or minimal latents). Detailed proof sketches for the Telephone theorem and the generalized KPD theorem already existed and we turned these into formal proofs (while also formalizing the theorem statements). Claims about redundant information remain at a lower level of formalization.

We also outlined four different ways in which the natural abstractions agenda could help for AI alignment:

1. The truth/falsehood of the Universality Hypothesis affects which other research agendas are likely to be promising.
2. Defining abstractions appears as a subproblem of defining many concepts in agent foundations (such as “agency” or “values”).
3. A definition of what makes an abstraction “good” or natural could accelerate research by serving as a tool for evaluation.
4. An understanding of natural abstractions could help advance interpretability.

We explained how the theoretical results we discussed earlier fit into this picture: they started as an attempt to make empirical tests of the Natural Abstraction Hypothesis feasible (1.), but also try to formalize natural abstractions (2.-4.).

Finally, we have given some of our own views on the natural abstractions agenda. In particular:

- We’ve described several areas where we see a need for more theoretical work, such as moving beyond information theory to representations, considering the finite regime instead of just infinite limits, and dealing with the fact that current definitions depend massively on the choice of variables.
- We agree natural abstractions have multiple different plausible connections to alignment, which is very promising. On the other hand, we discuss a few reservations and ways in which other research agendas such as empirical interpretability can address the same questions.
- We discuss how we would approach the natural abstractions agenda in methodologically different ways: aim for more precision in claims and formalization of proofs, more experiments, and connect ideas to existing work.

We expect there will be some disagreement about these views but hope they will lead to fruitful discussions. Beyond that, we hope that the earlier sections of this post can serve as an easier way for people to get up to speed on the natural abstractions agenda than existing writing, while still being comprehensive.

Acknowledgments

We would like to start by extending a big thank you to John Wentworth! His feedback on early drafts and discussions with him have made this project significantly easier. We also really appreciated his openness about his uncertainties and about how his views have changed over time.

Thanks as well to Jan Kirchner for writing a summary of related work in Neuroscience!

Thanks also to Ryan Greenblatt, Alexander Oldenziel, Lisa Thiergart, and Dan Hendrycks for feedback and helpful conversations.

Our TL;DR was much longer in an earlier version of this post—thanks to Raemon for suggesting we shorten it.

Leon Lang worked on this project as part of the SERI ML Alignment Theory Scholars Program - Winter 2022 Cohort.

1. ^

   John mentioned a caveat on this to us:
   Note that I sometimes hedge about whether "the natural abstractions" are F(X) itself, or whether they're a latent variable of which F(X) is an estimate. The latter is probably the right answer, but we'd expect in typical systems that the estimate is very precise, so the distinction doesn't matter much. (Prototypical example: average particle energy in one chunk of a gas as an estimate of the temperature of the gas.)
   [Further explanation after some discussion with us:]
   Latent variables, in general, are not necessarily fully determined by the physical state of the universe; that much just naturally drops out of the math. Latents are just these mathematical constructs. They can be predictively useful and powerful, while still mathematically having uncertainty separate from the state of the world.

   Another way to frame it: consider the Kolmogorov complexity/Solomonoff induction view. From a God's-eye view, we could observe the entire low-level state of the universe, then find the shortest program which outputs that state. And it's entirely possible that that shortest program contains some variables whose values we are unable to perfectly estimate, even knowing the entire low-level state of the universe. (In the Kolmogorov context, this means that there are multiple different programs with approximately-the-same length which all output the observed universe-state, and all have very similar structure, but assign different values to corresponding variables.) What our uncertainty is over is the values of the latent variables - i.e. the internal variables used by the programs which approximately-maximally compress the low-level universe state. Insofar as the programs are near-optimal compressions, that uncertainty should be small, but it's not necessarily zero. And of course those internal variables can be predictively useful and powerful for modeling the world, even if their values are not fully determinable from the full world-state.

We're not sure whether we fully understand his views here, and in any case think this distinction shouldn't matter too much for the rest of our post, so we won't discuss it further.

2. ^

   The (slight) difference is that Gibbs sampling is typically defined as resampling X_1, then X_2, and so on, wrapping around to X_1 after each variable has been resampled once. In contrast, John proposes randomly choosing which variable to resample at each step.

3. ^

   Note that it's currently not quite clear in which sense anything converges here, see appendix for some notes on further formalization of X_∞.

4. ^

   It's certainly possible that the connection between theoretical progress so far and future empirical tests is just not meant to be fully legible based on John's public writing.