

Inductive Biases for Deep Learning of Higher-Level Cognition

Anirudh Goyal

Mila, University of Montreal

ANIRUDHGOYAL9119@GMAIL.COM

Yoshua Bengio

Mila, University of Montreal

YOSHUA.BENGIO@MILA.QUEBEC

Abstract

A fascinating hypothesis is that human and animal intelligence could be explained by a few principles (rather than an encyclopedic list of heuristics). If that hypothesis was correct, we could more easily both understand our own intelligence and build intelligent machines. Just like in physics, the principles themselves would not be sufficient to predict the behavior of complex systems like brains, and substantial computation might be needed to simulate human-like intelligence. This hypothesis would suggest that studying the kind of inductive biases that humans and animals exploit could help both clarify these principles and provide inspiration for AI research and neuroscience theories. Deep learning already exploits several key inductive biases, and this work considers a larger list, focusing on those which concern mostly higher-level and sequential conscious processing. The objective of clarifying these particular principles is that they could potentially help us build AI systems benefiting from humans' abilities in terms of flexible out-of-distribution and systematic generalization, which is currently an area where a large gap exists between state-of-the-art machine learning and human intelligence.

1. Has Deep Learning Converged?

Is 100% accuracy on the test set enough? Many machine learning systems have achieved excellent accuracy across a variety of tasks (Deng et al., 2009; Mnih et al., 2013; Schrittwieser et al., 2019), yet the question of whether their reasoning or judgement is correct has come under question, and answers seem to be wildly inconsistent, depending on the task, architecture, training data, and interestingly, the extent to which test conditions match the training distribution. Have the main principles required for deep learning to achieve human-level performance been discovered, with the main remaining obstacle being to scale up? Or do we need to follow a completely different research direction not built on the principles discovered with deep learning, in order to achieve the kind of cognitive competence displayed by humans? Our goal here is to better understand the gap between current deep learning and human cognitive abilities so as to help answer these questions and suggest research directions for deep learning with the aim of bridging the gap towards human-level AI. Our main hypothesis is that deep learning succeeded in part because of a set of inductive biases, but that additional ones should be added in order to go from good in-distribution generalization in highly supervised learning tasks (or where strong and dense rewards are available), such as object recognition in images, to strong out-of-distribution generalization and transfer learning to new tasks with low sample complexity. To make that concrete, we consider

some of the inductive biases humans may exploit for higher-level and highly sequential cognition operating at the level of conscious processing, and review some early work exploring these “high-level cognitive inductive priors” in deep learning. We argue that the progression from MLPs to convnets to transformers has in many ways been an (incomplete) progression towards the original goals of deep learning, i.e., to enable the discovery of a hierarchy of representations, with the most abstract ones, often associated with language, at the top. In other words our arguments suggest that deep learning brought remarkable progress but needs to be extended in qualitative and not just quantitative ways (larger datasets and more computing resources). We argue that having larger and more diverse datasets (Brown et al., 2020) is important but insufficient without good architectural inductive biases. At the same time, we make the case that evolutionary forces, the interactions between multiple agents, the non-stationary and competition systems put pressure on the learner to achieve the kind of flexibility, robustness and ability to adapt quickly which humans seem to have when they are faced with new environments (Bansal et al., 2017; Liu et al., 2019; Baker et al., 2019; Leibo et al., 2019). In addition to thinking about the learning advantage, this paper focuses on knowledge representation in neural networks, with the idea that by decomposing knowledge in small pieces which can be recomposed dynamically as needed (to reason, imagine or explain at an *explicit* level), one may achieve the kind of systematic generalization which humans enjoy and is obvious in natural language (Marcus, 1998, 2019; Lake and Baroni, 2017; Bahdanau et al., 2018).

1.1 Data, Statistical Models and Causality

Our current state-of-the-art machine learning systems sometimes achieve good performance on a specific and narrow task, using very large quantities of labeled data, either by supervised learning or reinforcement learning (RL) with strong and frequent rewards. Instead, humans are able to understand their environment in a more unified way (rather than with a separate set of parameters for each task) which allows them to quickly generalize (from few examples) on a new task, thanks to their ability to reuse previously acquired knowledge. Instead current systems are generally not robust to changes in distribution (Peters et al., 2017b), adversarial examples (Goodfellow et al., 2014; Kurakin et al., 2016) etc.

One possibility studied in the machine learning literature is that we should train our models with multiple datasets, each providing a different view of the underlying model of the world shared by humans (Baxter, 2000). Whereas multi-task learning usually just pools the different datasets (Caruana, 1997; Collobert and Weston, 2008; Ruder, 2017), we believe that there is something more to consider: we want our learner to perform well on a completely new task or distribution, either immediately (with zero-shot out-of-distribution generalization), or with a few examples (i.e. with efficient transfer learning) (Ravi and Larochelle, 2016; Wang et al., 2016; Finn et al., 2017).

This raises the question of changes in distribution or task. Whereas the traditional train-test scenario and learning theory assumes that test examples come from the same distribution as the training data, dropping that assumption means that we cannot say anything about generalization to a modified distribution. Hence new assumptions are required about how the different tasks or the different distributions encountered by a learning agent are related to each other.

We use the term *structural-mechanistic* (Schölkopf, 2015) to characterize models which follow an underlying mechanistic understanding of reality. They are closely related to the structural causal models used to capture causal structure (Pearl, 2009). The key property of such models is that they will make correct predictions over a variety of data distributions which are drawn from the same underlying causal system, rather than being specific to a particular distribution. To give a concrete example, the equation $E = MC^2$ relates mass and energy in a way which we expect to hold regardless of other properties in the world. On the other hand, an equation like $GDP_t = 1.05GDP_{t-1}$ may be correct under a particular data distribution (for example a country with some growth pattern) but will fail to hold when some aspects of the world are changed, even in ways which did not happen or could not happen, i.e., in a counterfactual.

However, humans do not represent all of their knowledge in such a neat verbalizable way as Newton’s equations. Most humans understand physics first of all at an *intuitive* level and in solving practical problems we typically combine such implicit knowledge with explicit verbalizable knowledge. We can name high-level variables like position and velocity but may find it difficult to explain intuitive inference mechanisms which relate them to each other, in everyday life (by opposition to a physicist running a simulation of Newton’s equations). An important question for us is how knowledge can be represented in these two forms, the implicit – intuitive and difficult to verbalize – and the explicit – which allows humans to share part of their thinking process through natural language.

This suggests that one possible direction that deep learning need to incorporate is more notions about agency and causality, even when the application only involves single inputs like an image and not actually learning a policy. For this purpose we need to examine how to go beyond the statistical learning framework which has dominated deep learning and machine learning in recent decades. Instead of thinking of data as a set of examples drawn independently from the same distribution, we should probably reflect on the origin of the data through a real-world non-stationary process. This is after all the perspective which learning agents, such as babies or robots, need to have in order to succeed in the changing environment they face.

2. About Inductive Biases

The no-free-lunch theorem for machine learning (Wolpert et al., 1995; Baxter, 2000) basically says that some set of preferences (or inductive bias) over the space of all functions is necessary to obtain generalization, that there is no completely general-purpose learning algorithm, that any learning algorithm implicitly or explicitly will generalize better on some distributions and worse on others. Typically, given a particular dataset and loss function, there are many possible solutions (e.g. parameter assignments) to the learning problem that exhibit equally “good” performance on the training points. Given a finite training set, the only way to generalize to new input configurations is then to rely on some assumption about the solution we are looking for. The question for AI research aiming at human-level performance then is to identify inductive biases that are most relevant to the human perspective on the world around us. Inductive biases, broadly speaking, encourage the learning algorithm to prioritise solutions with certain properties. Table 1 lists some of the inductive biases already used in various neural networks, and the corresponding properties.

Inductive Bias	Corresponding property
Distributed representations	Patterns of features
Convolution	group equivariance (usually over space)
Deep architectures	Complicated functions = composition of simpler ones
Graph Neural Networks	equivariance over entities and relations
Recurrent Nets	equivariance over time
Soft attention	equivariance over permutations

Table 1: Examples of current inductive biases in deep learning

From Inductive Biases to Algorithms. There are many ways to encode such biases—e.g. explicit regularisation objectives (Bishop et al., 1995; Bishop, 1995; Srivastava et al., 2014; Kukačka et al., 2017; Zhang et al., 2017), architectural constraints (Yu and Koltun, 2015; Long et al., 2015; Dumoulin and Visin, 2016; He et al., 2016; Huang et al., 2017), parameter sharing (Hochreiter and Schmidhuber, 1997; Pham et al., 2018), implicit effects of the choice of the optimization method, or choices of prior distributions in a Bayesian model. For example, one can build translation invariance of a neural network output by replacing matrix multiplication by convolutions (LeCun et al., 1995) and pooling (Krizhevsky et al., 2012), or by averaging the network predictions over transformations of the input (feature averaging) (Zhang et al., 2017), or by training on a dataset augmented with these transformations (data augmentation) (Krizhevsky et al., 2012). Whereas some inductive biases can easily be encoded into the learning algorithm (e.g. with convolutions), the preference over functions is sometimes implicit and not intended by the designer of the learning system, and it is sometimes not obvious how to turn an inductive bias into a machine learning method, this often being the core contribution of machine learning papers.

Inductive Biases as Data. We can think of inductive biases or priors and built-in structure as “training data in disguise”, and one can compensate lack of sufficiently powerful priors by more data. Interestingly, different inductive biases may be equivalent to more or less data (could be exponentially more data): we suspect that inductive biases based on a form of compositionality (like distributed representations (Pascanu et al., 2013), depth (Montufar et al., 2014) and attention (Bahdanau et al., 2014; Vaswani et al., 2017)) can potentially also provide a larger advantage (to the extent that they apply well to the function to be learned). On very large datasets, the advantage of inductive biases may be smaller, which suggests that transfer settings (where only few examples are available for the new distribution) are interesting to evaluate the advantage of inductive biases and of their implementation.

Inductive Biases and Ease of Learning. Interestingly, we observe a link between the advantage that an inductive bias brings in terms of generalization and the advantage it brings in terms of ease of optimization (e.g. attention, or residual connections). This makes sense if one thinks of the training process as a form of generalization inside the training set. With a strong inductive bias, the first half of the training set already strongly constrains what the learner should predict on the second half of the training set, which means that when we encounter these examples, their error is already low and training thus converges faster.

Inductive Biases about How the World Works. Note that in certain reinforcement learning (RL) problems, such as learning to play chess or Go, the rules of the domain’s dynamics are so simple that you can easily build an almost perfect simulator, and hence generate an infinite data set without looking at the real world. In that case the most constraining factor is computational resources and not data. A simulator is a generative world model, and in most games, that ‘world’ is fairly simple. In many real-world settings (e.g., which involve images or natural language), we do not have a good simulator of the environment, and the hard task is to learn one (or enough aspects of it). Although our world looks very complicated, reflecting aspects of it which humans understand suggests that there is a way to decompose our knowledge about the world in small pieces which can be independently discovered and then combined in new ways, and that events happen according to causality and according to stable laws like the laws of physics, in spite of the changes in distribution we encounter. A generative causal understanding of the world facilitates generalization to new unseen domains. Generative models allow us to learn from a single example because we can embed that example in a structured construction of background knowledge. Humans have a remarkable ability to simulate counterfactual worlds that will never be but can nonetheless exist in our minds. We can imagine the consequences of our actions by simulating the world at an abstract level as it unfolds under some putative past or future action, or we can derive what caused current events by simulating possible worlds that may have led to it. Clearly, this ability depends on our intuitive understanding of the dynamics of the world around us, but it is interesting to note that these internal simulations are very different from physics simulations or those of current model-based RL systems: they don’t involve the full state of the world but only a few chosen variables and attributes, which matter to that simulation. This sparsity of the dependencies and of the causal reasoning steps which humans capture in their verbalizable thoughts is another inductive prior which we wish to add to the deep learning toolbox.

Agency, Sequential Decision Making and Non-Stationary Data Streams. The classical framework for machine learning is based on the assumption of identically and independently distributed data (i.i.d.), i.e test data has the same distribution as the training data. This is a very important assumption, because if we did not have that assumption, then we would not be able to say anything about generalization to new examples from the same distribution. Unfortunately, this assumption is too strong, and reality is not like this, especially for agents taking decisions one at a time in an environment from which they also get observations. The distribution of observations seen by an agent may change for many reasons: the agent acts (intervenes) in the environment, other agents intervene in the environment, or simply our agent is learning and exploring, visiting different parts of the state-space as it does so, discovering new parts of it along the way, thus experiencing non-stationarities along the way. Although sequential decision-making is ubiquitous in real life, there are scenarios where thinking about these non-stationarities may seem unnecessary (like object recognition in static images). However, if we want to build learning systems which are robust to changes in distribution, it may be necessary to train them in settings where the distribution changes! And then of course there are applications of machine learning where the data is sequential and non-stationary (like historical records of anything) or even more so, where the learner is also an agent or is an agent interacting with other agents

(like in robotics, autonomous driving or dialogue systems). That means we may need to go away from large curated datasets typical of supervised learning frameworks and instead construct non-stationary controllable environments as the training grounds and benchmarks for our learners. This complicates the task of evaluating and comparing learning algorithms but is necessary and we believe, feasible, e.g. see Yu et al. (2017); Packer et al. (2018); Chevalier-Boisvert et al. (2018); Dulac-Arnold et al. (2020); Ahmed et al. (2020).

Transfer Learning and Continual Learning. Instead of a fixed data distribution and searching for an inductive bias which works well with this distribution, we are thus interested in transfer learning (Pratt et al., 1991; Pratt, 1993) and continual learning (Ring, 1998) scenarios, with a potentially infinite stream of tasks, and where the learner must extract information from past experiences and tasks to improve its learning speed (i.e., sample complexity, which is different from asymptotic performance which is currently the standard) on future and yet unseen tasks. Suppose the learner faces a sequence of tasks, A, B, C and then we want the learner to perform well on a new task D. Short of any assumptions it is nearly impossible to expect the learner to perform well on D. However if there is some shared structure, between the transfer task (i.e task D) and source tasks (i.e tasks A, B and C), then it is possible to generalize or transfer knowledge from the source task to the target task. Hence, if we want to talk meaningfully about knowledge transfer, it is important to talk about the assumptions on the kind of data distribution that the learner is going to face, i.e., (a) what they may have in common, what is stable and stationary across the environments experienced and (b) how they differ or how changes occur from one to the next in case we consider a sequential decision-making scenario. This division should be reminiscent of the work on *meta-learning* (Bengio et al., 1990; Schmidhuber, 1987; Finn et al., 2017; Ravi and Larochelle, 2016), which we can understand as dividing learning into slow learning (of stable and stationary aspects of the world) and fast learning (of task-specific aspects of the world). This involves two time scales of learning, with an outer loop for meta-learning of meta-parameters and an inner loop for regular learning of regular parameters. In fact we could have more than two time scales (Clune, 2019): think about the outer loop of evolution, the slightly faster loop of cultural learning (Bengio, 2014) which is somewhat stable across generations, the faster learning of individual humans, the even faster learning of specific tasks and new environments within a lifetime, and the even faster inner loops of motor control and planning which adapt policies to the specifics of an immediate objective like reaching for a fruit. Ideally, we want to build an understanding of the world which shifts as much of the learning to the slower and more stable parts so that the inner learning loops can succeed faster, requiring less data for adaptation.

Systematic Generalization and Out-of-Distribution Generalization. In this paper, we focus on the objective of out-of-distribution (OOD) generalization, i.e., generalizing outside of the specific distribution(s) from which training observations were drawn. A more general way to conceive of OOD generalization is with the concept of sample complexity in the face of new tasks or changed distributions. One extreme is zero-shot OOD generalization while the more general case, often studied in meta-learning setups, involves k -shot generalization (from k examples of the new distribution).

Whereas the notions of OOD generalization and OOD sample complexity tell us what we want to achieve (and hint at how we might measure it) they say nothing about how to achieve

it. This is where the notion of *systematic generalization* becomes interesting. Systematic generalization is a phenomenon which was first studied in linguistics (Lake and Baroni, 2017; Bahdanau et al., 2018) because it is a core property of language, that meaning for a novel composition of existing concepts (e.g. words) can be derived systematically from the meaning of the composed concepts. This very clearly exists in language, but humans benefit from it in other settings, e.g., understanding a new object by combining properties of different parts which compose it. Systematic generalization even makes it possible to generalize to new combinations that have zero probability under the training distribution: it is not just that they did not occur in the training data, but that even if we had seen an infinite amount of training data from our training distribution, we would not have any sample showing this particular combination. For example, when you read a science fiction scenario for the first time, that scenario could be impossible in your life, or even in the aggregate experiences of billions of humans living today, but you can still imagine such a counterfactual and make sense of it (e.g., predict the end of the scenario from the beginning). Empirical studies of systematic generalization were performed by Bahdanau et al. (2018, 2019), where particular forms of combinations of linguistic concepts were present in the training distribution but not in the test distribution, and current methods take a hit in performance, whereas humans would be able to answer such questions easily.

Humans use inductive biases providing forms of compositionality, making it possible to generalize from a finite set of combinations to a larger set of combinations of concepts. Deep learning already benefits from a form of compositional advantage with distributed representations (Hinton, 1984; Bengio and Bengio, 2000; Bengio et al., 2001), which are at the heart of why neural networks work so well. There are theoretical arguments about why distributed representations can bring a potentially exponential advantage (Pascanu et al., 2013), if this matches properties of the underlying data distribution. Another advantageous form of compositionality in deep nets arises from the depth itself, i.e., the composition of functions, again with provable exponential advantages under the appropriate assumptions (Montufar et al., 2014). However, a form of compositionality which we propose here should be better incorporated in deep learning is the form called systematicity (Lake and Baroni, 2018) defined by linguists, and more recently systematic generalization in machine learning papers (Bahdanau et al., 2018).

Current deep learning methods tend to overfit the training *distribution*. This would not be visible by looking at a test set from the same distribution as the training set, so we need to change our ways of evaluating the success of learning because we would like our learning agents to generalize in a systematic way, out-of-distribution. This only makes sense if the new environment has enough shared components or structure with previously seen environments, which corresponds to certain assumptions on distributional changes, bringing back the need for appropriate inductive biases, about distributions (e.g., shared components) as well as about how they change (e.g., via agents’ interventions).

3. Inductive biases based on higher-level cognition as a path towards systems that generalize OOD

This section starts by reviewing findings from cognitive neuroscience which may inspire us in hypothesizing a set of inductive biases for higher-level cognition.

3.1 Conscious vs Unconscious Processing in Brains

Imagine that you are driving a car from your office, back home. You do not need to pay lot of attention to the road and you can talk to you. Now imagine encountering a road block due to construction then you have to pay slightly more attention, you have to be on lookout for new information, if someone tries talking to you, then you may have to tell the person, “please let me drive”. It is interesting to consider that when humans are confronted with a new situation, very commonly they require their *conscious* attention. For example in the driving example, when there is a road block you need to pay attention in order to think what to do next, and you probably don’t want to be disturbed, because your conscious attention can only focus on one thing at a time.

There is something in the way humans process information which seems to be different – both functionally and in terms of neural signature in the brain – when we deal with conscious processing and novel situations (changes in the distribution) which require our conscious attention, compared to our habitual routines. In those situations, we generally have to *think*, *focus* and *attend* to specific elements of our perception, actions or memory and sometimes inhibit our reactions based on context (e.g., facing new traffic rules or a road block). Why would humans have evolved to deal with such an ability with changes in distribution? Maybe simply because life experience is highly non-stationary.

Cognitive scientists distinguish (Botvinick et al., 2001) *habitual* versus *controlled* processing, where the former corresponding to default behaviors, whereas the latter require attention and *mental effort*. Daniel Kahneman introduced the framework of fast and slow thinking (Kahneman, 2011), and the *system 1* and *system 2* styles of processing in our brain. Some tasks can be achieved using only system 1 abilities whereas others also require system 2 and conscious processing. There are also notions of explicit (verbalizable) knowledge and explicit processing (which roughly correspond to system 2) and implicit (intuitive) knowledge and corresponding neural computations. The default (or unconscious) processing of system 1 can take place very rapidly (as fast as about 100ms) and mobilize many areas of the brain in parallel. On the other hand, controlled (or conscious) processing involves a sequence of thoughts, usually verbalizable, typically requiring seconds to achieve. Whereas we can act in fast and precise habitual ways without having to think consciously, the reverse is not true: controlled processing (i.e., system 2 cognition) generally requires unconscious processing to perform much of its work. It is as if the conscious part of the computation was just the top-level program and the tip of the iceberg. Yet, it seems to be a very powerful one, which makes it possible for us to solve new problems creatively by recombining old pieces of knowledge, to reason, to imagine explanations and future outcomes, to plan and to apply or discover causal dependencies. It is also at that level that we interface with other humans through natural language. And when a word refers to a complex concept for which we do not have a clear verbalizable and precise explanation (like how we manage to drive our bike), we can still talk about it, reason with it, etc. Even imagination and planning (which are hallmarks of system 2 abilities) require system 1 computations to sample candidate solutions to a problem (from a possibly astronomical number, which we never have to explicitly examine).

Our brain seems to thus harbour two very different types of knowledge: the kind we can explicitly reason about and communicate verbally (system 2 knowledge) and the kind that is intuitive and implicit (system 1 knowledge). When we learn something new, it typically starts being represented explicitly, and then as we practice it more, it may migrate to a different, implicit form. When you learn the grammar of a new language, you may be given some set of rules, which you try to apply on the fly, but that requires a lot of effort and is done painfully slowly. As you practice this skill, it can gradually migrate to a habitual form, you make less mistakes (for the common cases), you can read / translate / write more fluently, and you may even eventually forget the original rules. When a new rule is introduced, you may have to move back some of that processing to system 2 computation to avoid inconsistencies. It looks as if one of the key roles of conscious processing is to integrate different sources of knowledge (from perception and memory) in a coherent way. This is at the heart of the Global Workspace Theory (or GWT) from Baars (1993) and its extension, the Global Neuronal Workspace model (Dehaene et al., 2017). The GWT theory posits that conscious processing revolves around a communication bottleneck between selected parts of the brain which are called upon when addressing a current task. There is threshold of relevance beyond which information which was previously handled unconsciously gains access to this bottleneck, instantiated in a working memory. When that happens, that information is broadcast to the whole brain, allowing the different relevant parts of it to synchronize and choose more configurations and interpretations of their piece of the action which are more globally coherent with the configurations chosen in other parts of the brain.

It looks like current deep learning systems are fairly good at perception and system 1 tasks. They can rapidly produce an answer (if you have parallel computing like that of GPUs) through a complex calculation which is difficult (or impossible) to dissect into the application of a few simple verbalizable operations. They require a lot of practice to learn and can become razor sharp good at the kinds of data they are trained on. On the other hand, humans enjoy system 2 abilities which permit fast learning (I can tell you a new rule in one sentence and you do not have to practice it in order to be able to apply it, albeit awkwardly and slowly) and systematic generalization, both of which should be important characteristics of the next generation of deep learning systems. For this purpose, we are proposing to take inspiration from cognition and build machines which integrate two very different kinds of representations and computations corresponding to the system 1 / implicit / unconscious vs system 2 / explicit / conscious divide. This paper is about inductive biases not yet integrated in state-of-the-art deep learning systems but which could help us achieve these system 2 abilities. In the next subsection, we summarize some of these system 2 inductive biases.

3.2 High-level Representations Describe Verbalizable Concepts as Semantic Variables

We know that a way to define conscious processing for humans is that it is the part of our mental activity which can be verbally reported (possibly not always perfectly, though). What follows is maybe the most influential inductive bias we want to consider in this paper: that *high-level variables (manipulated consciously) are generally verbalizable*. To put it in simple terms, we can imagine the high-level semantic variables captured at this top level

of a representation to be associated with single words. In practice, the notion of word is not always the same across different languages, and the same semantic concept may be represented by a single word or by a phrase. There may also be more subtlety in the mental representations (such as accounting for uncertainty or continuous-valued properties) which is not always or not easily well reflected in their verbal rendering.

Because we know a lot about natural language, this inductive bias may give rise to many more inductive biases derived from it and from properties of natural language, as we will see in the rest of this section. Keep in mind that these inductive biases do not need to cover all the aspects of our internal model of the world (they couldn't), only those aspects of our knowledge which we are able to communicate with language. The rest would have to be represented in pure system 1 (non system 2) machinery, such as in an encoder-decoder that could relate low-level actions and low-level perception to high-level variables. If there is some set of properties that apply well to some aspects of the world, then it would be advantageous for a learner to have a subsystem that takes advantage of these properties (the inductive priors described here) and a subsystem which models the other aspects. These inductive priors then allow faster learning and potentially other advantages like systematic generalization, at least concerning these aspects of the world which are consistent with these assumptions (system 2 knowledge, in our case).

The assumption we are making is thus that there is a simple lossy mapping from high-level semantic representations to natural language expressions of them in natural language. This is an inductive bias which could be exploited in grounded language learning scenarios (Winograd, 1972; Hermann et al., 2017; Chevalier-Boisvert et al., 2018) where we couple language data with observations and actions by an agent. We can constrain the learner's architecture so that there would be a simple transformation from the high-level representation to sentences. To encourage modularity (discussed below), we can make that transformation use attention mechanisms so that few of the quantities in the representation mostly suffice to produce a single word or phrase (some representation of context is always going to be also necessary, though, since we know that language is context-dependent and that there are grammatical constraints in the sequence of words generated). Grounded language learning would thus put pressure on the top-level representation so that it captures the kinds of concepts expressed with language. One can view this as a form of weak supervision, where we don't force the top-level representations to be human-specified labels, only that there is a simple relationship between these representations and utterances which humans would often associate with the corresponding meaning.

3.3 Semantic Variables Play a Causal Role and Knowledge about them is Modular

In natural environments, biological systems achieve sophisticated and scalable group behavior among vast numbers of independent organisms using simple control strategies, e.g. bird flocks, fish schools, and multi-cellular organisms. Such biological phenomena have inspired the design of several distributed multi-agent systems, for example, swarm robotic systems, sensor networks, and modular robots. Despite this, most machine learning models employ the opposite inductive bias, i.e., with all elements (e.g., artificial neurons) interacting all the time. The GWT (Baars, 1997; Dehaene, 2020) also posits that the brain is composed

in a modular way, with a set of expert modules which need to communicate but only do so sparingly and via a bottleneck through which only a few elements can squeeze at any time. If we believe that theory, these elements are the concepts present to our mind at any moment, and a few of them are called upon and joined in working memory in order to reconcile the interpretations made by different modular experts across the brain. The decomposition of knowledge into recomposable pieces also makes sense as a requirement for obtaining systematic generalization: conscious attention would then select which expert and which concepts (which we can think of as variables with different attributes and values) interact with which pieces of knowledge (which could be verbalizable rules or non-verbalizable intuitive knowledge about these variables) stored in the modular experts. On the other hand, the modules which are not brought to bear in this conscious processing may continue working in the background in a form of default or habitual computation (which would be the form of most of perception). For example, consider the task of predicting from the pixel-level information the motion of balls sometimes colliding against each other as well as the walls. It is interesting to note that all the balls follow their default dynamics, and only when balls collide do we need to intersect information from several bouncing balls in order to make an inference about their future states. Saying that the brain modularizes knowledge is not sufficient, since there could be a huge number of ways of factorizing knowledge in a modular way. We need to think about the desired properties of modular decompositions of the acquired knowledge, and we propose here to take inspiration from the causal perspective on understanding how the world works, to help us define both the right set of variables and their relationship.

We are calling these variables manipulated consciously semantic variables because they can be verbalized. *We hypothesize that semantic variables are also causal variables.* Words in natural language often refer to agents (subjects, which cause things to happen), objects (which are controlled by agents), actions (often through verbs) and modalities or properties of agents, objects and actions (for example we can talk about future actions, as intentions, or we can talk about time and space where events happen, or properties of objects). It is thus plausible to assume that causal reasoning of the kind we can verbalize involves as variables of interest those high-level semantic variables which may sit at the top of a processing hierarchy in the brain, at a place where signals from all modalities join, such as pre-frontal cortex (Cohen et al., 2000).

The connection between causal representations and modularity is profound: an assumption which is commonly associated with structural causal models is that it should break down the knowledge about the causal influences into *independent mechanisms* (Peters et al., 2017b). As explained in Section 4.1, each such mechanism relates direct causes to one or more effect variables and knowledge of one such mechanism should not tell us anything about another mechanism (otherwise we should restructure our representations and decomposition of knowledge to satisfy this information-theoretic independence property). This is not about independence of the corresponding random variables but about the algorithmic mutual information between the descriptions of these mechanisms. What it means practically and importantly for out-of-distribution adaptation is that if a mechanism changes (e.g. because of an intervention), the representation of that mechanism (e.g. the parameters used to capture a corresponding conditional distribution) needs to be adapted but that of the others do not need to be changed to account for that change.

These mechanisms may be organized in the form of an acyclic causal schema which scientists attempt to identify. The sparsity of the change in the joint distribution between the semantic variables (discussed more in Section 3.4) is different but related to a property of such high-level structural causal model: the sparsity of the graph capturing the joint distribution itself (discussed in Section 3.6). In addition, the causal structure, the causal mechanisms and the definition of the high-level causal variables tend to be stable across changes in distribution, as discussed in Section 3.5.

3.4 Local Changes in Distribution in Semantic Space

Consider a learning agent (like a learning robot or a learning child). What are the sources of non-stationarity for the distribution of observations seen by such an agent, assuming the environment is in some (generally unobserved) state at any particular moment? Two main sources are (1) the non-stationarity due to the environmental dynamics (including the learner’s actions and policy) not having converged to an equilibrium distribution (or equivalently the mixing time of the environment’s stochastic dynamics is much longer than the lifetime of the learning agent) and (2) causal interventions by agents (either the learner of interest or some other agents). The first type of change includes for example the case of a person moving to a different country, or a videogame player playing a new game or a never-seen level of an existing game. That first type also includes the non-stationarity due to changes in the agent’s policy arising from learning. The second case includes the effect of actions such as locking some doors in a labyrinth (which may have a drastic effect on the optimal policy). The two types can intersect, as the actions of agents (including those of the learner, like moving from one place to another) contribute to the first type of non-stationarity.

Let us consider how humans describe these changes with language. For many of these changes, they are able to explain the change with a few words (a single sentence, often). This is a very strong clue for our proposal to include as an inductive bias the assumption that *most changes in distribution are localized in the appropriate semantic space*: only one or a few variables or mechanisms need to be modified to account for the change. Note how humans will even create new words when they are not able to explain a change with a few words, with the new words corresponding to new latent variables, which when introduced, make the changes explainable “easily” (assuming one understands the definition of these variables and of the mechanisms relating them to other variables).

For type-2 changes (due to interventions), we automatically get locality, since by virtue of being localized in time and space, actions can only affect very few variables, with most other effects being consequences of the initial intervention. This sparsity of change is a strong assumption which can put pressure on the learning process to discover high-level representations which have that property. Here, we are assuming that the learner has to jointly discover these high-level representations (i.e. how they relate to low-level observations and low-level actions) as well as how the high-level variables relate to each other via causal mechanisms.

3.5 Stable Properties of the World

Above, we have talked about changes in distribution due to non-stationarities, but there are aspects of the world that are stationary, which means that learning about them eventually converges. In an ideal scenario, our learner has an infinite lifetime and the chance to learn everything about the world (a world where there are no other agents) and build a perfect model of it, at which point nothing is new and all of the above sources of non-stationarity are gone. In practice, only a small part of the world will be understood by the learning agent, and interactions between agents (especially if they are learning) will perpetually keep the world out of equilibrium. If we divide the knowledge about the world captured by the agent into the stationary aspects (which should converge) and the non-stationary aspects (which would generally keep changing), we would like to have as much knowledge as possible in the stationary category. The stationary part of the model might require many observations for it to converge, which is fine because learning these parts can be amortized over the whole lifetime (or even multiple lifetimes in the case of multiple cooperating cultural agents, e.g., in human societies). On the other hand, the learner should be able to quickly learn the non-stationary parts (or those the learner has not yet realized can be incorporated in the stationary parts), ideally because very few of these parts need to change, if knowledge is well structured. Hence we see the need for at least two speeds of learning, similar to the division found in meta-learning of learnable coefficients into meta-parameters on one hand (for the stable, slowly learned aspects) and parameters on the other hand (for the non-stationary, fast to learn aspects), as already discussed above in Section 2. The proposed inductive bias (which now involves the whole system, not just system 2), is that *there should be several speeds of learning, with more stable aspects learned more slowly and more non-stationary or novel ones learned faster, and pressure to discover stable aspects among the quickly changing ones*. This pressure would mean that more aspects of the agent’s represented knowledge of the world become stable and thus less needs to be adapted when there are changes in distribution.

For example, consider scientific laws, which are most powerful when they are universal. At another level, consider the mapping between the perceptual input, low level actions, and the high-level semantic variables. The encoder that implements this mapping should ideally be highly stable, or else downstream computations would need to track those changes (and indeed the low-level visual cortex seems to compute features that are very stable across life, contrary to high-level concepts like new visual categories). Causal interventions are taking place at a higher level than the encoder, changing the value of an unobserved high-level variables or changing one of the mechanisms. If a new concept is needed, it can be added without having to disturb the rest of it, especially if it can be represented as a composition of existing high-level features and concepts. We know from observing humans and their brain that new concepts which are not obtained from a combination of old concepts (like a new skill or a completely new object category not obtained by composing existing features) take more time to learn, while new high-level concepts which can be readily defined from other high-level concepts can be learned very quickly (as fast as with a single example / definition).

Another example arising from the analysis of causal systems is that causal interventions (which are in the non-stationary, quickly inferred or quickly learned category) may tem-

porarily modify the causal graph structure (which variable is a direct cause of which) by breaking causal links (when we set a variable we break the causal link from its direct causes) but that most of the causal graph is a stable property of the environment. Hence, we need neural architectures which make it easy to quickly adapt the relationship between existing concepts, or to define new concepts from existing ones.

3.6 Sparse Factor Graph in the Space of Semantic Variables

Our next inductive bias for high-level variables can be stated simply: *the joint distribution between high-level concepts can be represented by a sparse factor graph*. Any joint distribution can be expressed as a factor graph, but we claim that the ones which can be conveniently described with natural language have the property that they should be sparse. A factor graph is a particular factorization of the joint distribution. A factor graph is bipartite, with variable nodes on one hand and factor nodes on the other. Factor nodes represent dependencies between the variables to which they are connected. To illustrate the sparsity of verbalizable knowledge, consider knowledge graphs and other relational systems, in which relations between variables often involve only two arguments (i.e., two variables). In practice, we may want factors with more than two arguments, but probably not a lot more. We understand each factor to capture a causal mechanism between its argument variables, and thus we should add an additional semantic element to these factor graphs: each argument of a factor should either play the role of cause or of effect, making the bipartite graph directed.

It is easy to see that linguistically expressed knowledge satisfies this property by noting that statements about the world can be expressed with a sentence and each sentence typically has only a few words, and thus relates very few concepts. When we write “If I drop the ball, it will fall on the ground”, the sentence clearly involves very few variables, and yet it can make very strong prediction about the position of the ball. A factor in a factor graph involving a subset S of variables is simply stating a probabilistic constraint among these variables. It allows one to predict the value of one variable given the others (if we ignore other constraints or factors), or more generally it allows to describe a preference for some set of values for a subset of S , given the rest of S . The fact that natural language allows us to make such strong predictions conditioned on so few variables should be seen as surprising: it only works because the variables are semantic ones. If we consider the space of pixel values in images, it is very difficult to find such strongly predictive rules to say predict the value of one pixel given the value of three other pixels. What this means is that pixel space does not satisfy the sparsity prior associated with the proposed inductive bias.

The proposed inductive bias is closely related to the bottleneck of the GWT of conscious processing (see more details in Section 5.4), and one can also make links with the original von Neumann architecture of computers. In both the GWT and the von Neumann architecture, we have a communication bottleneck with in the former the working memory and in the latter the CPU registers where operations are performed. The communication bottleneck only allows a few variables to be brought to the nexus (working memory in brains, registers in the CPU). In addition, the operations on these variables are extremely sparse, in the sense that they take very few variables at a time as arguments (no more than the handful in working memory, in the case of brains, and generally no more than two or three in typical assembly languages). This sparsity constraint is consistent with a decomposition

of computation in small chunks, each involving only a few elements. In the case of the sparse factor graph assumption we only consider that sparsity constraint for declarative knowledge ("how the world works", its dynamics and statistical or causal structure), but a similar constraint may make sense for the computations taking place at the top level of the representation hierarchy (those we can describe verbally), which may capture inference mechanisms ("how to interpret some observations", "what to do next").

This assumption about the joint distribution between the high-level variables at the top of our deep learning hierarchy is different from the assumption commonly found in many papers on disentangling factors of variation (Higgins et al., 2016; Burgess et al., 2018; Chen et al., 2018; Kim and Mnih, 2018; Locatello et al., 2019), where the high-level variables are assumed to be marginally independent of each other, i.e., their joint distribution factorizes into independent marginals. We think this deviates from the original goals of deep learning to learn abstract high-level representations which capture the underlying explanations for the data. Note that one can easily transform one representation (with a factorized joint) into another (with a non-factorized joint) by some transformation (think about the independent noise variables in a structural causal model, Section 4). However, we would then lose the properties introduced up to now (that each variable is causal and corresponds to a word or phrase, that the factor graph is sparse, and that changes in distribution can be originated to one or very few variables or factors).

Instead of thinking about the high-level variables as completely independent, we propose to see them as having a very structured joint distribution, with a sparse factor graph and other characteristics (such as dependencies which can be instantiated on particular variables from generic schemas or rules, described above). We argue that if these high-level variables has to capture semantic variables expressible with natural language, then the joint distribution of these high-level semantic variables must have sparse dependencies rather than being independent. For example, high-level concepts such as "table" and "chair" are not statistically independent, instead they come in very powerful and strong but sparse relationships. Instead of imposing a very strong prior of complete independence at the highest level of representation, we can have this slightly weaker but very structured prior, that the joint is represented by a sparse factor graph. Interestingly, recent studies confirm that the top-level variables in generative adversarial networks (GANs), which are independent by construction, generally do not have a semantic interpretation (as a word or short phrase), whereas many units in slightly lower layers do have a semantic interpretation (Bau et al., 2018).

Why not represent the causal structure with a directed graphical model? In these models, which are the basis of standard representations of causal structure (e.g., in structural causal models, described below), knowledge to be learned is stored in the conditional distribution of each variable (given its direct causal parents). However, it is not clear that this is consistent with the requirements of independent mechanisms. For example, typical verbally expressed rules have the property that many rules could apply to the same variable. Insisting that the units of knowledge are conditionals would then necessarily lump the corresponding factors in the same conditional. This issue becomes even more severe if we think of the rules as generic pieces of knowledge which can be reused to be applied to many different tuples of instances, as elaborated in the next subsection.

3.7 Variables, Instances and Reusable Knowledge Pieces

A standard graphical model is static, with a separate set of parameters for each conditional (in a directed acyclic graph) or factor (in a factor graph). There are extensions which allow parameter sharing, e.g. through time with dynamic Bayes nets, or in undirected graphical models such as Markov Networks (Kok and Domingos, 2005) which allow one to “instantiate” general “patterns” into multiple factors of the factor graph. Markov Networks can for example implement forms of recursively applied probabilistic rules. But they do not take advantage of distributed representations and other inductive biases of deep learning.

The idea we are presenting here, is that instead of separately defining specific factors in the factor graph (maybe each with a piece of neural network), each having its separate set of parameters, we would define “generic factors”, “factor-rules”, “schemas” or “factor templates” or “factor factories” or “relations”. A schema, or generic factor, is a probabilistic analogue of a logical rule with quantifiers, i.e., with variables or arguments which can be bound. A static rule is a thing like ‘if John is hungry then he looks for food’. Instead, a more general rule is a thing like, ‘for all X , if X is a human and X is hungry, then X looks for food’ (with some probability). X can be bound to specific instances (or to other variables which may involve more constraints on the acceptable set). In classical AI, we have unification mechanisms to match together variables, instances or expressions involving variables and instances, and thus keep track of how variables can ultimately be ‘bound’ to instances (or to variables with more constraints on their attributes), when exploring whether some schema can be applied to some objects (instances or more generic objects) with properties (constituting a database of entities).

The proposed inductive bias is also inspired by the presence of such a structure in the semantics of natural language and the way we tend to organize knowledge according to relations, e.g., in knowledge graphs (Sowa, 1987). Natural language allows to state rules involving variables and is not limited to making statements about specific instances. The assumption is that *the independent mechanisms (with separate parameters) which specify dependencies between variables are generic, i.e., they can be instantiated in many possible ways to specific sets of arguments with the appropriate types or constraints.*

What this means in practice is that we never hold in memory the full instantiated graph of with all possible instances and all possible mechanisms relating them. Instead, we generate the needed pieces of the graph and even perform reasoning (i.e. deduction) at an abstract level where nodes in the graph (random variables) stand not for instances but for sets of instances belonging to some category or satisfying some constraints. Whereas one can unfold a recurrent neural network or a Bayesian network to obtain the fully instantiated graph, in the case we are talking about, similarly to Markov network, it is generally not feasible to do that. It means that inference procedures always look at a small piece of the (partially) unfolded graph at a time or even reasons about how to combine these generic schemas without having to fully instantiate them with concrete instances or concrete objects in the world. One way to think about this inspired by how we do programming is that we have functions with generic and possibly typed variables as arguments and we have instances on which a program is going to be applied. At any time (as you would have in Prolog), an inference engine must match the rules with the current instances (so the types and other constraints between arguments are respected) as well as other elements (such as

what we are trying to achieve with this computation) in order to combine the appropriate computations. It would make sense to think of such a computation controller, based on attention mechanisms, to select which pieces of knowledge and which pieces of the short-term (and occasionally long-term memory) need to be combined in order to push new values in working memory.

An interesting outcome of such a representation is that one can apply the same knowledge (i.e knowledge specified by a schema which links multiple abstract entities together) to different instances (i.e different object files). For example, you can apply the same laws of physics to two different balls that are visually different (and maybe have different colors and masses). This is also related to notions of arguments and indirection in programming. The power of such relational reasoning resides in its capacity to generate inferences and generalizations that are constrained by the roles that elements play, and the roles they can play may depend on the properties of these elements, but these schemas specify how entities can be related to each other in systematic (and possibly novel) ways. In the limit, relational reasoning yields universal inductive generalization from a finite and often very small set of observed cases to a potentially infinite set of novel instances, so long as those instances can be described by attributes (specifying types) allowing to bound them to appropriate schemas.

3.8 Relevant causal chains (for learning or inference) tend to be very short

Our brains seem to segment streams of sensory input into meaningful representations of episodes and *events*. Event segmentation allows functional representations that support temporal reasoning, an ability that arguably relies on mechanisms to encode and retrieve information to and from memory (Zacks et al., 2007; Radvansky and Zacks, 2017). Indeed, faced with a task, our brains appear to easily and *selectively* pluck context-relevant past information from memory, enabling both powerful multi-scale associations as well as flexible computations to relate temporally distant events. As we argue here, the ability of the brain to efficiently segment sensory input into events, and the ability to *selectively* recall information from the distant past based on the current context helps to efficiently propagate information (such as credit assignment or causal dependencies) over long time spans. Both at the cognitive and at the physiological levels, there is evidence of information “routing” mechanisms that enable this efficient propagation of information, although they are far from being understood.

Our next inductive bias is almost a consequence of the biases on causal variables and the bias on the sparsity of the factor graph for the joint distribution between high-level variables. It states that *causal chains used to perform learning (to propagate and assign credit) or inference (to obtain explanations or plans for achieving some goal) are broken down into short causal chains of events which may be far in time but linked by the top-level factor graph over semantic variables.*

3.9 Context-dependent processing involving goals, top-down influence, and bottom-up competition

Successful perception in humans clearly relies on both top-down and bottom-up signals. Top-down information encode relevant context, priors and preconceptions about the current

scene: for example, what we might expect to see when we enter a familiar place. Bottom-up signals consist of what is literally observed through sensation. The best way to combine top-down and bottom-up signals remains an open question, but it is clear that these signals need to be combined in a way which is dynamic and depends on context - in particular top-down signals are especially important when stimuli are noisy or hard to interpret by themselves (for example walking into a dark room). Additionally, which top-down signals are relevant also changes depending on the context. It is possible that combining specific top-down and bottom-up signals that can be weighted dynamically (for example using attention) could improve robustness to distractions and noisy data.

In addition to the general requirement of dynamically combining top-down and bottom-up signals, it makes sense to do so at every level of processing hierarchy to make the best use of both sources of information at every stage of that computation, as is observed in the visual cortex (with very rich top-down signals influencing the activity at every level). This can be summarized by stating an inductive bias in favour of architectures in which *top-down contextual information is dynamically combined with bottom-up sensory signals at every level of the hierarchy of computations relating low-level and high-level representations*.

3.10 Inspiration from Computer Programming

A central question for this research program is where we can find better inductive biases. Up to now we have mostly taken inspiration from human cognition, but an indirect mark of human cognitive biases can be found in the field of programming languages, already hinted at in subsections above.

Programming languages provide high-level tools and constructs for controlling computers. Computer programs have a number of interesting properties. They are concrete and literal, as they always reduce to machine instructions. At the same time, they provide efficient abstractions for writing complicated programs with a small amount of reusable code (even within the same program, a piece of code can be instantiated and executed many times). They have also seen a significant amount of human effort and design, which has led to a number of different frameworks and concepts.

Some of the goals of a programming language are shared with the goals of higher-level cognition. High-level conscious processing manipulates a small number of high-level variables at a time (those which are present to our mind at that moment), and abstract tasks are decomposed into more specific sub-tasks. With programming languages, a programmer wishes to express a simple high-level specification of a family of computations and to translate it into low-level machine code. For these reasons, many of the structures present in programming languages could also serve as inspiration for system 2 deep learning.

Classes and Objects: Exploration has started of a recurrent neural net in which modules with distinct states dynamically learn when and how to share parameters (Goyal et al., 2020). This was directly inspired by the idea from object-oriented programming in which a single class, which has a set of methods, can be used to create many objects which have their own distinct states (called properties in programming languages).

Recursion: Functions which can call themselves can be surprisingly dynamic and adaptable. For example trees and search routines are often most simply written with recursive

functions. Recursion requires calling the same function many times (in principle an unlimited number of times) and ideally in a consistent way. Despite the importance of this capability, many machine learning models don't have an easy route towards learning to handle recursive functions. For example, feed-forward networks (including MLPs, convnets, and transformers) typically have no parameter sharing between layers, making exact recursion require very precise correspondence between parameter values.

Functions with named and typed arguments: There is an intriguing connection between key-value attention mechanisms used in deep learning (Bahdanau et al., 2014) and variables with typed arguments in programming languages or the use of names in programming so that the same instance is used as argument on different calls to possibly different functions: a query can be used to select a key which has an associated value appropriate for the use linked with the query. The keys and queries in deep learning attention mechanisms can be seen as a soft variant of the hard name-based selection which is used in programming languages. With typed languages, the type of the provided argument matches (in the desired dimensions) the expected type, we know that the function can be applied to the provided argument. Queries and keys can also be seen as representing types (the expected type of an argument and the actual type of an object instance, respectively), and this double meaning as both a type and a name is consistent with the name being a distributed (and thus rich and unique) representation of the type, with type matching not necessarily requiring a perfect match but sometimes only a match on some dimensions. If the individual modules in such a system 2 neural network can represent functions with typed arguments, then they can bind them to a currently available input which matches the role or type expected for this function, thus enabling a crucial form of systematic generalization. With this perspective, there are several other ideas from programming languages which are not captured well by current deep learning. For example, programming languages allow for nested variables, whose value can contain references to other variables, with arbitrarily many levels of nesting.

4. Declarative Knowledge of Causal Dependencies

Whereas a statistical model captures a single joint distribution, a causal model captures a large family of joint distributions, each corresponding to a different intervention (or set of interventions), which modifies the (unperturbed) distribution (e.g., by removing parents of a node and setting a value for that node). Whereas the joint distribution $P(A, B)$ can be factored either as $P(A)P(B|A)$ or $P(B)P(A|B)$ (where in general both graph structures can fit the data equally well), only one of the graphs corresponds to the correct causal structure and can thus consistently predict the effect of interventions. The asymmetry is best illustrated by an example: if A is altitude and B is average temperature, we can see that intervening on A will change B but not vice-versa.

Preliminaries Given a set of random variables X_i , a bayesian network is commonly used to describe the dependency structure of both probabilistic and causal models via a Directed Acyclic Graph (DAG). In this graph structure, a variable (represented by a particular node) is independent of all the other variables, given all the direct neighbors of a variable. The edge

direction identify a specific factorization of the joint distribution of the graph’s variables:

$$p(X_1, \dots, X_n) = \prod_{i=1}^m p(X_i \mid \mathbf{PA}_i). \quad (1)$$

Structural causal models (SCMs). A Structural Causal Model (SCM) (Peters et al., 2017b) over a finite number M of random variables X_i given a set of *observables* X_1, \dots, X_M (modelled as random variables) associated with the vertices of a DAG G , is a set of structural assignments

$$X_i := f_i(X_{pa(i,C)}, N_i), \quad \forall i \in \{1, \dots, M\} \quad (2)$$

where f_i is a deterministic function, the set of noises N_1, \dots, N_m are assumed to be *jointly independent*, and $pa(i, C)$ is the set of parents (direct causes) of variable i under configuration C of the SCM directed acyclic graph, i.e., $C \in \{0, 1\}^{M \times M}$, with $c_{ij} = 1$ if node i has node j as a parent (equivalently, $X_j \in X_{pa(i,C)}$; i.e. X_j is a direct cause of X_i). Causal structure learning is the recovery of the ground-truth C from observational and interventional studies.

Interventions. Without experiments, or interventions i.e., in a purely-observational setting, it is known that causal graphs can be distinguished only up to a Markov equivalence class, i.e., the set of graphs compatible with the observed dependencies. In order to identify the true causal graph, the learner needs to perform interventions or experiments i.e., interventional data is needed (Eberhardt et al., 2012).

4.1 Independent Causal Mechanisms.

A powerful assumption about how the world works which arises from research in causality (Peters et al., 2017b) and briefly introduced earlier is that the causal structure of the world can be described via the composition of independent causal mechanisms.

Independent Causal Mechanisms (ICM) Principle. A complex generative model, temporal or not, can be thought of as composed of independent mechanisms that do not inform or influence each other. In the probabilistic case, this means a particular mechanism should not inform (in information sense) or influence the other mechanisms.

This principle subsumes several notions important to causality, including separate intervenability of causal variables, modularity and autonomy of subsystems, and invariance (Pearl, 2009; Peters et al., 2017a).

This principle applied to the factorization in (1), tells us that the different factors should be independent in the sense that (a) performing an intervention on one of the mechanisms $p(X_i \mid \mathbf{PA}_i)$ does not change any of the other mechanisms $p(X_j \mid \mathbf{PA}_j)$ ($i \neq j$), (b) knowing some other mechanisms $p(X_i \mid \mathbf{PA}_i)$ ($i \neq j$) does not give us information about any another mechanism $p(X_j \mid \mathbf{PA}_j)$.

Causal Factor Graphs. We propose that the formalism of directed graphical models, even extended to that of structural causal models (Eq. 2), may not be consistent with the idea of independent causal mechanisms, and that parametrizing the causal structure with a particular form of factor graph (with directed edges to represent causal direction) could be more appropriate. See Section 3.6 for an introduction to factor graphs. Our argument is the

following. If we force the conditionals $P(X_i|X_{pa(i,C)})$ (equivalently the parametrization of f_i in Eq. 2) to be the independent units of parametrization, then we cannot represent other decomposition’s that better factor out the independent knowledge pieces. For example, consider two independent causal ‘rules’ (not necessarily logical and discrete), rule R_1 which tells us how X_3 is affected by X_1 and rule R_2 which tells us how X_3 is affected by X_2 . They can of course be encapsulated in a single conditional $P(X_3|X_1, X_2)$ but we then lose the ability to localize an intervention which would affect only one of these two rules. If only the first rule is modified by an intervention, we still have to adapt the whole conditional to cope with that intervention. To make things even worse, imagine that R_1 is a generic rule, which can be applied across many different random variables, i.e., its parameters are shared across many parts of the structural causal model. It is difficult to capture that sharing if R_1 is hidden inside the black box of $P(X_3|X_1, X_2)$ and the other conditionals in which it appears. A proper factorization of knowledge into its independent pieces thus needs more flexibility in the parametrization of the causal dependencies, and this is one which factor graphs can offer. However, factor graphs are missing the information about causal direction which is why we suggest to extend them by associated with each argument of each factor’s potential function a binary indicator to specify whether the argument acts as a cause or as an effect. Seeing the factor graph as bipartite (with nodes for variables and nodes for factors), the bipartite graph now has directed edges, as well as a sharing structure (so that the same underlying ‘rule’ can be instantiated on different factor nodes of the graph).

4.2 Exploit changes in distribution due to causal interventions

Nature doesn’t shuffle examples. Real data arrives to us in a form which is not iid, and so in practice what many practitioners of data science or researchers do when they collect data is to *shuffle* it to make it iid. “Nature doesn’t shuffle data, and we should not” Bottou (2019). When we shuffle the data, we destroy useful information about those changes in distribution that are inherent in the data we collect and contain information about causal structure. Instead of destroying that information about non-stationarities, we should use it, in order to learn how the world changes.

4.3 Challenges for Deep Learning

Inducing causal relationships from observations is a classic problem in machine learning. Most work in causality starts from the premise that the causal variables themselves have known semantics or are observed. However, for AI agents such as robots trying to make sense of their environment, the only observables are low-level variables like pixels in images or low-level motor actions. To generalize well, an agent must induce high-level variables, particularly those which are causal (i.e., can play the role of cause or effect). A central goal for AI and causality research is thus the joint discovery of abstract high-level representations (how they relate to low-level observations and low-level actions) at the same time as of causal structure at the high level. This objective also dovetails with the goals of the representation learning and structure learning communities.

4.4 Relation between meta-learning, causality, OOD generalization and fast transfer learning

To illustrate the link between meta-learning, causality, OOD generalization and fast transfer learning, consider the example from Bengio et al. (2019). We consider two discrete random variables A and B , each taking N possible values. We assume that A and B are correlated, without any hidden confounder. The goal is to determine whether the underlying causal graph is $A \rightarrow B$ (A causes B), or $B \rightarrow A$. Note that this underlying causal graph cannot be identified from observational data from a single (training) distribution p only, since both graphs are Markov equivalent for p . In order to disambiguate between these two hypotheses, they use samples from some transfer distribution \tilde{p} in addition to our original samples from the training distribution p .

Without loss of generality, they fix the true causal graph to be $A \rightarrow B$, which is unknown to the learner. Moreover, to make the case stronger, they consider a setting called *covariate shift*, where they assume that the change (again, whose nature is unknown to the learner) between the training and transfer distributions occurs after an intervention on the cause A . In other words, the marginal of A changes, while the conditional $p(B | A)$ does not, i.e. $p(B | A) = \tilde{p}(B | A)$. Changes on the cause will be most informative, since they will have direct effects on B . Bengio et al. (2019) find experimentally that this is sufficient to identify the causal graph, while Priol et al. (2020) justify this with theoretical arguments in the case where the intervention is on the cause.

In order to demonstrate the advantage of choosing the causal model $A \rightarrow B$ over the anti-causal $B \rightarrow A$, Bengio et al. (2019) compare how fast the two models can adapt to samples from the transfer distribution \tilde{p} . They quantify the speed of adaptation as the log-likelihood after multiple steps of fine-tuning via (stochastic) gradient ascent, starting with both models trained on a large amount of data from the training distribution. They show via simulations that the model corresponding to the underlying causal model adapts faster. Moreover, the difference between the quality of the predictions made by the causal and anti-causal models as they see more post-intervention examples is more significant when adapting on a small amount of data, of the order of 10 to 30 samples from the transfer distribution. Indeed, asymptotically, both models recover from the intervention perfectly and are not distinguishable. This is interesting because it shows that generalization from few examples (after a change in distribution) actually contains more signal about the causal structure than generalization from a lot of examples (whereas in machine learning we tend to think that more data is always better). Bengio et al. (2019) make use of this property (the difference in performance between the two models) as a noisy signal to infer the direction of causality, which here is equivalent to choosing how to modularize the joint distribution. The connection to meta-learning is that in the inner loop of meta-learning we adapt to changes in the distribution, whereas in the outer loop we slowly converge towards a good model of the causal structure. Here the meta-parameters capture the belief about the causal graph structure and the default (unperturbed) conditional dependencies, while the parameters are those which capture the effect of the intervention.

(Ke et al., 2019) further expanded this idea to deal with more than two variables. To model causal relations and out-of-distribution generalization we view real-world distributions as a product of causal mechanisms. Any change in such a distribution (e.g., when moving

from one setting/domain to a related one) will always be due to changes in at least one of those mechanisms (Goyal et al., 2019; Bengio et al., 2019; Ke et al., 2019). An intelligent agent should be able to recognize and make sense of these sparse changes and quickly adapt their pre-existing knowledge to this new domain. A current hypothesis is that a causal graphical model defined on the appropriate causal variables would be more efficiently learned than one defined on the wrong representation. Preliminary work based on meta-learning (Ke et al., 2019; Dasgupta et al., 2019; Bengio et al., 2019) suggests that, parameterizing the correct variables and causal structures, the parameters of the graphical model capturing the (joint) observational distribution can be adapted faster to changes in distribution due to interventions. This comes as a consequence of the fact that fewer parameters need to be adapted as only few variables were affected by the intervention (Priol et al., 2020). In this sense, learning causal representations may bring immediate benefits to machine learning models.

4.5 Actions and affordances as part of the causal model

Understanding causes and effects is a crucial component of the human cognitive experience. Humans are agents and their actions change the world (sometimes only in little ways), and those actions can inform them about the causal structure in the world. Understanding that causal structure is important in order to plan further actions in order to achieve desired consequences, or to attribute credit to one’s or others’ actions, i.e., to understand and cope with changes in distribution occurring in the world. However, in realistic settings such as those experienced by a child or a robot, the agent typically does not have full knowledge of what abstract action was performed and needs to perform inference over that too: in addition to discovering the causal variables (from the low-level observations) and how they are related to each other, the agent needs to learn how high-level intentions and actions are linked with low-level observations and with interventions on the high-level causal variables.

A human-centric version of this viewpoint is the psychological theory of affordances (Gibson, 1977) that can be linked to predictive state representations in reinforcement learning: what can we do with an object? What are the consequences of these actions? Learning affordances as representations of how agents can cause changes in their environment by controlling objects and influencing other agents is more powerful than learning a data distribution. It would not only allow to predict the consequences of actions we may not have observed at all, but it also allows us to envision which potentialities would result from a different mix of interacting objects and agents. This line of thinking is directly related to the work in machine learning and reinforcement learning on controllability of aspects of the environment (Bengio et al., 2017; Thomas et al., 2017). A clue about a good way to define causal variables is precisely that there exist actions which can control one causal variable while not directly influencing the others (i.e., except via the causal variable which is being controlled). A learner thus needs to discover what actions give rise to interventions and how, but this can also help the learner figure out a good representation space for causal variables.

Right now, for most of the representation learning methods in deep reinforcement learning, the perceptual subsystem first collects sensory information to build an internal descriptive representation of environment, and then this information is used along with representations of memories of past experience to decide upon a course of action. But continuous

interaction with the world often does not allow the agent to stop to think or to collect information and build a complete knowledge of one’s surroundings. To adapt in an environment, an agent should be ready to adapt on short notice, executing actions which are only partially prepared. Hence one strategy could be for agents to process sensory information in an action-dependent manner, to build representations of the potential high-level actions which the environment currently affords. In other words, the perception of a given natural setting may involve not only representations which capture information about the environment, but also representations which specify the parameters of possible actions that can be taken in this context.

In essence, it is possible that we as humans address the questions of specification (“how to do it”) before deciding “what to do”. To summarize, sensory information arriving from the world is continuously used to identify several currently available potential actions, while other kinds of information are collected to select from among these one that will be executed at a given moment. The complete action plans do not need to be prepared for all of the possible actions that one might take at a given moment. First, only the actions which are currently afforded by the environment would be evaluated in this manner. Second, many possible actions may be eliminated by selective attention mechanisms which limit the sensory information that is transformed into representations of action.

5. Biological Inspiration and Characterization of High-Level Cognition

Whereas the last section delved deeper into causality, this one reviews biological and cognitive inspiration for inductive biases relevant for higher-level cognition.

5.1 Synergy between AI research and cognitive neuroscience

Our aim is to take inspiration from (and further develop) research into the cognitive science of conscious processing, to deliver greatly enhanced AI, with abilities observed in humans thanks to high-level reasoning leading among other things to greater abilities to face unusual or novel situations by reasoning, reusing existing knowledge and being able to communicate about that high-level semantic knowledge. At the same time, new AI models could drive new insights into the neural mechanisms underlying conscious processing, instantiating a virtuous circle. Machine learning procedures have the advantage that they can be tested for their effective learning abilities, and in our case in terms of out-of-distribution abilities or in the context of causal environments changing due to interventions, e.g., as in Ahmed et al. (2020). Because they have to be very formal, they can also suggest hypotheses for how brains might implement an equivalent strategy with neural machinery. Testing these hypotheses could in turn provide more understanding about how brains solve the same problems and help to refine the deep learning systems.

5.2 Attention

Attention is about sequentially selecting what computation to perform on what quantities. Let us consider a machine translation task from English to French. To obtain a good translation, we normally focus especially on the “right” few words in the English sentence that may be relevant to do the translation. This is the motivation that stimulated our work

on content-based soft attention (Bahdanau et al., 2014) but may also be at the heart of conscious processing in humans as well as in future deep learning systems with both system 1 and system 2 abilities.

Content Based Soft attention: soft attention forms a soft selection of one element (or multiple elements) from a previous level of computations, i.e we are taking a convex combination of the values of the elements at the previous level. These convex weights are coming from a softmax that is conditioned on how each of the elements’ key vector matches some query vector. In a way, attention is parallel, because computing these attention weights considers all the possible elements in some set, yielding a score for each of them, to decide which of them are going to receive the most attention. Content-based attention also introduces a non-local inductive bias into neural network processing, allowing it to infer long-range dependencies that might be difficult to discern if computations are biased by local proximity. Attention is at the heart of the current state-of-the art NLP systems (Devlin et al., 2018; Brown et al., 2020) and is the standard tool for memory-augmented neural networks (Graves et al., 2014; Sukhbaatar et al., 2015; Gulcehre et al., 2016; Santoro et al., 2018). Attention and memory can also help address the problem of credit assignment through long-term dependencies (Ke et al., 2018; Kerg et al., 2020) by creating dynamic skip connections through time (i.e., a memory access) which unlock the problem of vanishing gradients. Attention also transforms neural networks from machines that are processing vectors (e.g., each layer of a deep net), to machines that are processing sets, more particularly sets of key/value pairs, as with Transformers (Vaswani et al., 2017; Santoro et al., 2018). Soft attention uses the product of a *query* (or *read key*) represented as a matrix Q of dimensionality $N_r \times d$, with d the dimension of each key, with a set of N_o objects each associated with a *key* (or *write-key*) as a row in matrix K^T ($N_o \times d$), and after normalization with a softmax yields outputs in the convex hull of the *values* (or *write-values*) V_i (row i of matrix V). The result is

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V,$$

where the softmax is applied to each row of its argument matrix, yielding a set of convex weights. As a result, one obtains a convex combination of the values in the rows of V . If the attention is focused on one element for a particular row (i.e., the softmax is saturated), this simply selects one of the objects and copies its value to row j of the result. Note that the d dimensions in the key can be split into *heads* which then have their attention matrix and write values computed separately.

Attention as dynamic connections: We can think of attention as a way to create a dynamic connection between different blocks of computation, whereas in the traditional neural net setting, connections are fixed. On the receiving end (downstream module) of an attention-selected input, it is difficult to tell from the selected value vector from where it comes (among the selected upstream modules which competed for attention). To resolve this, it would make sense that the information being propagated along with the selected value includes a notion of key or type or name, i.e., of where the information comes from, hence creating a form of indirection (a reference to where the information came from, which can be passed to downstream computations).

Attention implements variable binding When the inputs and outputs of each of the factor are a set of objects or entities (each associated with a key and value vector), we have a generic object-processing machinery which can operate on “variables” in a sense analogous to variables in a programming language: as interchangeable arguments of functions. Because each object has a key embedding (which one can understand both as a name and as a type), the same computation can be applied to any variable which fits an expected “distributed type” (specified by a query vector). Each attention head then corresponds to a typed argument of the function computed by the factor. When the key of an object matches the query of head k , it can be used as the k -th input vector argument for the desired computation. Whereas in regular neural networks (without attention) neurons operate on fixed input variables (the neurons which are feeding them from the previous layer), the key-value attention mechanisms make it possible to select on the fly which variable instance (i.e. which entity or object) is going to be used as input for each of the arguments of some computation, with a different set of query embeddings for each argument. The computations performed on the selected inputs can be seen as *functions with typed arguments*, and attention is used to *bind* their formal argument to the selected input.

5.3 Modularity in the brain and in neural networks

An important goal for machine learning, perhaps its most important goal, is to learn to understand and interact with the world in a way that is inherently flexible. There are a lot of regularities in the real world and on each time interval not all elements of the observable environment change. A lot of the elements remain static or changing in uninteresting ways. Thus it may be wasteful to model each and every aspect of the world’s dynamics at each step of an imagined trajectory. The standard approach in machine learning represent the world at a point in time with a single hidden state which attempts to describe the full state of the world. However, not all aspects may matter to a particular plan or goal, and nearly independent systems may only occasionally interact. For example, imagine two different rooms in a house separated by a wall. For a model which represents transition dynamics with a single hidden state, it can be difficult to keep these two processes separate, and thus we would expect to have long-term interference which is not justified by the true dynamics of the world.

Implications for AI: The existence of modularity in the brain and of the theories of conscious processing such the GWT described below suggests an interesting connection: dynamically selecting modules for solving a task seems appropriate for obtaining systematic generalization. Each module captures a piece of knowledge and conscious processing selects on-the-fly just the ones which are currently relevant and need to be coherently combined. This would enable a compositional advantage aligned with the needs of systematic generalization out-of-distribution. In addition, only the selected modules would be under pressure to adapt when the result of the combination needs to be tuned, leading to selective adaptation similar to that explored by Bengio et al. (2019) (see Section 4.4 above) where just a few relevant modules need to adapt to a change in distribution.

5.4 Global workspace theory

The global workspace theory (GWT) was introduced by Baars (1993, 1997) and refined by Dehaene and Changeux (2011); Dehaene et al. (2017); Dehaene (2020) and others. Many of its ingredients motivate the inductive biases discussed in this paper.

Coherence through a shared workspace. In cognitive science, the Global Workspace Theory (GWT) (Baars, 1993) suggests an architecture allowing specialist components to interact. As developed in Section 5.4, the key claim of GWT is the existence of a shared representation—sometimes called a blackboard, sometimes a workspace—that can be modified by any selected specialist and that is broadcast to all specialists. The GWT suggests a fleeting memory capacity in which only one consistent content can be dominant at any given moment. Our interpretation of this restriction on write access by a very small number of specialists selected on-the-fly by an attention mechanism is that it stems from an assumption on the form of the joint distribution between high-level variables whose values are broadcast. If the joint distribution factor graph is sparse, then only a few variables (those involved in one factor or a few connected factors) need to be synchronized at each step of an inference process. By constraining the size of the working memory, we enforce the sparsity of the factor graph, thus implementing the inductive bias discussed earlier (Section 3.6). GWT also makes a claim that the workspace is associated with the conscious contents of cognition, which can be reported verbally. The basic idea of deep learning frameworks inspired by the GWT is to explore a similar communication and coordination scheme for a neural net comprising of distinct modules.

Serial and Parallel Computations: From a computational perspective, one hypothesis about the dynamics of communication between different modules is that different modules generally act in parallel, but when they do need to communicate information with another *arbitrary* module, the information has to go through a routing bottleneck. Among the results of computations performed by different parts of the brain, some elements get selected and broadcast to the rest of the brain, influencing the rest of the brain. The contents which have thus been selected are essentially the only ones which can be committed to memory, starting with short-term memory. Working memory refers to the ability of the brain to operate on a few recently accessed elements (i.e., those in short-term memory). These elements can be remembered and have a heavy influence on the next thought, action or perception, as well as on what learning focuses on, possibly playing a role similar to desired outputs, goals or targets in supervised learning for system 1 computations.

One way to think about this is when an agent uses their imagination, different modules can actually participate, but for the imagined scenario to make sense, the different pieces of knowledge from different modules need to be coordinated and form a coherent configuration. This would be the role of the GWT bottleneck, as part of an inference process which selects and uses existing pieces of knowledge to form coherent interpretations of selected aspects of the world, at each moment. Because so few elements can be put in coherence at each step, this inference process generally requires several such steps, leading to the highly sequential nature of system 2 computation (compared with the highly parallel nature of system 1 computation).

5.5 Verbal Reporting and Grounded Language Learning

Conscious content is revealed by reporting it, often with language. This suggests that high-level variables manipulated consciously are closely related with their verbal forms (like words and phrases), as discussed in Section 3.2, and this is why we call them semantic variables. However, much of what our brains knows cannot be easily translated in natural language and forms the content of system 1 knowledge, as discussed in Sections 1.1 and 3.1. This means that system 2 (verbalizable) knowledge is incomplete: words are mostly pointers to semantic content which belongs to system 1 and thus in great part not consciously accessible, e.g., capturing what instances of the "door" category look like or the kind of motor control schemes that allow one to open a door. This suggests that training an AI system using only text corpora, even huge ones, will not be sufficient to build systems which understand the meaning of sentences in a way comparable to even a young child. It will be necessary for natural language understanding systems to be trained in a way that couples natural language with what it refers to. This is the idea of *grounded language learning*, but the discussion in this paper should also suggest that passively observing for example videos and verbal explanations of them may be insufficient: in order to capture the causal structure understood by humans, it may be necessary for learning agents to be embedded in an environment in which they can act and thus discover its causal structure. Studying this kind of setup was the motivation for our work on the Baby AI environment (Chevalier-Boisvert et al., 2018).

5.6 Slow processing and out-of-distribution problem solving

As an agent, a human being is facing frequent changes because of their actions or the actions of other agents in the environment. Most of the time, humans follow their habitual policy, but tend to use system 2 cognition when having to deal with unfamiliar settings. It allows humans to generalize out-of-distribution in suprisingly powerful ways, and understanding this style of processing would help us build these abilities in AI as well. This is illustrated with our early example of driving in an area with different traffic regulations, which requires full conscious attention (Section 3.1). This observation suggests that system 2 cognition is crucial in order to achieve the kind of flexibility and robustness to changes in distribution required in the natural world.

5.7 Between-Modules Interlingua and Communication Topology

If the brain is composed of different modules, it is interesting to think about what code or lingua franca is used to communicate between them, such that it can lead to interchangeable pieces of knowledge being dynamically selected and combined to solve a new problem. The GWT bottleneck may thus also play a role in forcing the emergence of such a lingua franca (Baars, 1997): the same information received by module A (e.g. "there is a fire") can come from any other module (say B, which detected a fire by smell, or C which detected a fire by sight). Hence B and C need to use a compatible representation which is broadcast via the GWT bottleneck for A's use. Again, we see the crucial importance of attention mechanisms to force the emergence of shared representations and indirect references exchanged between the modules via the conscious bottleneck.

However, the GWT bottleneck is by far not the only way for modules to communicate with each other. Regarding the topology of the communication channels between modules, it is known that modules in the brain satisfy some spatial topology such that the computation is not all-to-all between all the modules. It is plausible that the brain uses both fixed local or spatially nearby connections as well as the global broadcasting system with top-down influence. We also know that there are hierarchical communication routes in the visual cortex (on the path from pixels to object recognition), and we know how successful that has been in computer vision with convnets. Combining these different kinds of inter-module communication modalities in deep network thus seems well advised as well: (1) Modules which are near each other in the brain layout can probably communicate directly without the need to clog the global broadcast channel (and this would not be reportable consciously). (2) Modules which are arbitrarily far from each other in the spatial layout of the brain can exchange information via the global workspace, following the theatre analogy of Baar’s. The other advantage of this communication route is of course the exchangeability of the sources of information being broadcast, which we hypothesize leads to better systematic generalization.

However, the role of working memory in the GWT is not just as a communication buffer. It also serves as a blackboard (or analogously the “registers” in CPUs) where operations can be done locally to improve coherence. The objective of such a coherence-seeking mechanism is that the different modules (especially the active ones) should adopt a configuration of their internal variables (and especially the more abstract entities they communicate to other modules) which is consistent with what other active modules “believe”. It is possible, that a large part of the functional role of conscious processing is for that purpose, which is consistent with the view of the working memory as a central element of the inference mechanism seeking to obtain coherent configurations of the variables interacting according to some piece of knowledge (a factor of the factor graph, a causal dependency).

5.8 Inference versus declarative knowledge

There are two forms of knowledge representation we have discussed: declarative knowledge (in particular of causal dependencies and environment dynamics, with an explicit causal graph structure), and inference mechanisms (in particular with modules which communicate with each other to solve problems, answer questions, imagine solutions). Standard graphical models only represent the declarative knowledge and typically require expensive iterative computations (such as Monte-Carlo Markov chains) to perform approximate inference. However, brains need fast inference mechanisms, and most of the advances made with deep learning concern such learned fast inference computations. Doing inference using only the declarative knowledge (the graphical model) is very flexible (any question of the form “predict some variables given other variables or imagined interventions” can be answered) but also very slow. We also know that after system 2 has been called upon to deal with novel situations repeatedly, the brain tends to bake these patterns of response in habitual system 1 circuits which can do the job faster and more accurately but have lost some flexibility. When a new rule is introduced, the system 2 is flexible enough to handle it and slow inference needs to be called upon again. Neuroscientists have also accumulated evidence that the hippocampus is involved in replaying sequences (from memory or imagination) for consol-

idation into cortex (Alvarez and Squire, 1994) so that they can be presumably committed to cortical long-term memory and fast inference mechanisms.

In general, searching for a good configuration of the values of top-level variables which is consistent with the conditions being given is computationally intractable. However, different approximations can be made which trade-off computational cost for quality of the solutions found. This difference could also be an important ingredient of the difference between system 1 (fast and parallel approximate and inflexible inference) and system 2 (slower and sequential but more flexible inference).

5.9 Reasoning through Sequences of Relevant Events

Temporal information-processing tasks generally require that the continuous stream of time be segmented into distinct intervals or a sequence of selected events. In a *clock-based segmentation*, the boundaries are spaced equally in time, resulting in fixed-duration intervals. In an *event-based segmentation*, the boundaries depend on the state of the environment, resulting in variable-duration intervals. Models of temporal information processing in cognitive science and artificial intelligence generally rely on clock-based segmentation. However, event-based segmentation can greatly simplify temporal information-processing tasks. An agent interacting with the world is faced with visual data from the environment and does not have the capacity to consciously process all that information in real time. Some form of orienting mechanism could alert the agent when a relevant stimulus appears, allowing the agent to process and respond to the stimulus. The orienting mechanism does not need to know how to respond to the stimulus, but only that the stimulus is relevant and must be dealt with. The orienting mechanism achieves a form of event-based segmentation. Its detection of a relevant event in the temporal stream triggers information processing of the event. The psychological reality of event-based segmentation can be illustrated through a familiar phenomenon. Consider the experience of traveling from one location to another, such as from home to office. If the route is unfamiliar, as when one first starts a new job, the trip is confusing and lengthy, but as one gains more experience following the route, one has the sense that the trip becomes shorter. One explanation for this phenomenon is as follows. On an unfamiliar route, the orienting mechanism constantly detects novel events, and a large number of such events will accumulate over the course of the trip. In contrast, few novel events occur on a familiar route. If our perception of time is event based, meaning that higher centers of cognition count the number of events occurring in a temporal window, not the number of milliseconds, then one will have the sense that a familiar trip is shorter than an unfamiliar trip. Experimental evidence from cognitive science clearly supports the notion that the event stream influences the perception of duration (Zacks et al., 2007), although the relationship between novelty and the perception of duration is complicated due to interacting variables such as sequence complexity and attention. At least at a conscious level, humans are not able to reason about many such events at a time, due to the limitations on short-term memory and the bottleneck of conscious processing Baars (1997). Hence it is plausible that humans would exploit an assumption on temporal dependencies in the data: that the most relevant ones only involve short dependency chains, or a small-depth graph of direct dependencies. Depth here refers to the longest path in the relevant graph of dependencies between events. What we showed earlier (Ke et al., 2018; Kerg et al., 2020)

is that this prior assumption is the strongest ingredient to bound the degree of vanishing of gradients.

6. Recent and Ongoing Work

The inductive biases discussed in this paper have been behind our recent or current research summarized below.

6.1 Recurrent Independent Mechanisms (Goyal et al., 2019)

RIMs are inspired by the decomposition of knowledge in small exchangeable pieces implied jointly by the sparse factor graph assumption (Section 3.6) and the inductive bias to favour modularity of causal knowledge (Section 3.3).

6.2 Learning to combine top-down and bottom-up information. (Mittal et al., 2020)

We explore deep recurrent neural net architectures in which bottom-up and top-down signals are dynamically combined using attention. The RIM-based modularity of the architecture and use of attention to control exchange of information further restricts the sharing and communication of information. Together, attention and modularity direct bottom-up vs top-down information flow, which leads to reliable performance improvements in perceptual and language tasks, and in particular improves robustness to distractions and noisy data. This work is based on the same inductive biases as RIMs but adds the inductive bias about the need for combining top-down and bottom-up information flows discussed in Section 3.9.

6.3 Object Files and Schemata (Goyal et al., 2020)

In addition to the inductive biases already exploited in RIMs, this architecture includes the inductive bias on generic knowledge (rules, schemata) which can be instantiated on different objects, with multiple versions of the generic knowledge applicable to different objects (Section 3.7).

6.4 Sparse Attentive Backtracking (Ke et al., 2018)

Humans make limited and focused uses of memory during cognitive processing (at least in terms of conscious access). In parallel, inference in architectures like SAB (Ke et al., 2018) involves examining the whole memory of past states and selecting a finite subset of them. Ideally, the attention mechanism should only select a few of these possibly relevant memories. One way to do so is to only attend the top- κ relevant memories, according to some appropriately learned attention score.

This work exploits the inductive bias on the shortness of relevant causal chains (Section 3.8), even though it does not explicitly handles causality and uses a form of back-propagation for credit assignment. Future work should investigate how to apply these ideas in a reinforcement learning setting, in a manner which could also be related to the recent work on hindsight credit assignment (Harutyunyan et al., 2019).

6.5 A meta-transfer objective for learning to disentangle causal mechanisms (Bengio et al., 2019)

In this paper we examine the inductive bias about changes in distribution derived from research in causality, introduced in Section 3.4. Causal models specify how the joint distribution of interest could change under interventions (Pearl, 1988), or more generally actions (Sutton and Barto, 2018). A consequence of this causal view is that changes of distribution are captured by the change in a single random variable (which can itself influence many others directly or indirectly). This suggests as an inductive bias that changes in distribution have such a locality in the appropriate representation space (Goyal et al., 2019; Bengio et al., 2019), and our work, both theoretically and through simulations, suggests that such an assumption indeed helps to adapt faster after an intervention Bengio et al. (2019). The parameter which controls which causal graph is believed to be the correct one is meta-learned in the outer-loop of training, whereas the inner loop modifies the parameters of the conditionals and marginals so as to adapt them to the modified distribution (after an intervention).

6.6 Learning neural causal models from unknown interventions (Ke et al., 2019)

This paper extends the previous one by extending the application of the same inductive bias from a generative model with two causal variables to one with an arbitrary number of causal variables. To avoid the exponential growth in the number of possible causal graphs, the paper proposes to factorize the belief distribution over graphs (which is gradually adapted), with one free parameter per directed edge. In addition, the paper finds that making an inference about which variable was the victim of the intervention helps convergence of training, and it proposes a way to combine prior knowledge about some edges (which are supposed known) with learning of the missing edges (whether they should be present or not). It also exploits a prior (as a regularizer) about the absence of cycles in the graph and uses an efficient parametrization of the all the possible conditionals (with the same neural network used for all possible choices of direct causal parents, the only difference being that the corresponding inputs may be masked or not).

7. Projects Looking Forward

The ideas presented in this paper are still in early stages of maturation, with only a few papers beginning the necessary work of ironing out the devil in the detail. Many open questions and paths remain and we highlight a few here.

- One of the big remaining challenges in line with the ideas discussed here remains to jointly learn a large-scale encoder (mapping low-level pixels to high-level causal variables) and a large-scale causal model of these high-level variables. An ideal scenario for this would be that of model-based reinforcement learning, where the causal model would learn the stochastic system dynamics. We have done this (Bengio et al., 2019) only at a small scale (with two causal variables) and using an encoder guaranteed to not have singular values 1 for its Jacobian, which avoids a potential collapse of the

encoder. In order to avoid collapse, one possibility is to use a contrastive loss at the high level (e.g. as in Deep Infomax (Hjelm et al., 2018) for example).

- Another major challenge is to unify in a single architecture both the declarative knowledge representation (like a structural causal model) and the inference mechanism (maybe implemented with attention and modularity, like with RIMs and their variants). There is a lot of data from human cognition about the consolidation of rule-based behavior into fast habitual skills which could serve as inspiration, e.g., maybe using replay from the hippocampus to train cortical modules to be consistent with the declarative knowledge. Existing work on variational auto-encoders can serve as inspiration as well (with the encoder being the inference machine and the decoder being the causal model, in that case). See Section 5.8 for a relevant discussion.
- Most current deep learning models use fixed parameter sharing and fixed regular memory access patterns which are well tailored to modern computing hardware (such as GPUs and TPUs) relying on SIMD parallelism. However, the form of attention-driven computation described in this paper may require dynamic, irregular and sparse memory access and parameter sharing which does not fit well with GPUs and makes it difficult to parallelize computation across examples of a minibatch. Tackling this may require innovation in neural architectures, low-level programming and hardware design. In terms of model-induced parallelism, the SCOFF approach (Goyal et al., 2020) shows some promise in this direction, where most of the computation is decentralized in each of the experts, and the conscious processing is just the tip of the iceberg in terms of computational cost.
- The way humans plan is very different from the approach currently used in model-based RL (or hybrids such as AlphaZero based on MCTS and value functions). Humans seem to exploit the inductive bias about the sparsity of the causal factor graph and the fact that reasoning sequences in the abstract space can be very short. It means that when humans plan, they do not build trajectories of the full state but instead partial state trajectories where only some aspects (variables) of the state are considered. In addition, they do not unfold future trajectories for every discrete time step but directly learn how to relate temporally distant events, similar to how credit assignment is considered by Ke et al. (2018). It would be interesting to explore these inductive biases in novel planning methods, which might be much more efficient than standard ones. When we plan, we can consider the possibility of novel situations, and if a model misses important aspects of the causal structure, it may not generalize well to these novel changes, and the plans may radically overestimate or underestimate some novel possibilities.
- We really want to have sparsity in the computation over modules and over data points. This is again difficult to handle with mini-batches, which are necessary for taking full advantage of GPUs. Efficiently optimizing these computations is challenging but could greatly help move the research forward.
- Scaling to large number of modules: the brain is probably composed of a very large number of independent modules, whereas most of the current work in modular deep

learning deals with much smaller numbers of modules, like 20. It would be interesting to consider new algorithms, architectures that can help in extending to a very large number of modules.

- Macro and Micro Modules: the kinds of modules that are usually considered in GWT are pretty high-level, e.g., face recognition, gait recognition, object recognition, visual routines, auditory speech perception, auditory object recognition, tactile object recognition. These are *macro-modules* rather than modules that carve up the visual input into single objects i.e *micro-modules*. Most of the work we have done focuses on micro-modules. How should a modular hierarchy be structured which accounts for both these large-scale and fine-scale ways of modularizing knowledge and computation?

8. Looking Backward: Relation to Good Old-Fashioned Symbolic AI

How are the approaches proposed here for system 2 computation different from and related to classical symbolic AI (GOFAI) approaches? Let us first review some of the issues with these classical approaches which have motivated solutions building on top of deep learning instead.

1. We want efficient large scale learning, e.g., brought by variants of stochastic gradient descent and end-to-end learning behind the state-of-the-art in modern deep learning. It is challenging to learn to perform pure symbol manipulation on a large scale because of the discreteness of the operations.
2. We want semantic grounding of the higher-level concept in terms of the lower-level observations and lower-level actions (which is done by system 1 computation in the brain). This is important because some of the understanding of the world (maybe a large part) is not represented at the conscious system 2 level, and is completely lacking when representing knowledge purely in symbolic terms.
3. We want distributed representations of higher-level concepts (to benefit from the remarkable advantages this brings in terms of generalization): whereas pure symbolic representations put every symbol at the same distance of every other symbol, distributed representations represent symbols through a vector of attributes, with related symbols having overlapping representations.
4. We want efficient search and inference. A computational bottleneck of GOFAI is search, which in probabilistic terms is equivalent to the problem of inference and is generally intractable and needs to be approximated. Variational auto-encoders have shown how this computational cost can be amortized by training the inference machinery. This the only way we currently know how to do this in a general enough way, and it is consistent with cognitive neuroscience of transfer from system 2 to system 1 of habitual skills.
5. We want to handle uncertainty, which most machine learning approaches are meant to handle.

We already have these capabilities with the current deep learning toolbox. What is missing is to integrate to that toolbox what will be required to achieve the kind of systematic generalization and decomposition of knowledge into small exchangeable pieces which is typically associated with GOFAL. We believe that it will not be sufficient to slap GOFAL methods on top of representations produced by neural nets, for the above reasons, but especially points 1, 3 and 4.

9. Conclusions

To be able to handle dynamic, changing conditions, we want to move from deep statistical models able to perform system 1 tasks to deep structural models also able to perform system 2 tasks by taking advantage of the computational workhorse of system 1 abilities. Today's deep networks may benefit from additional structure and inductive biases to do substantially better on system 2 tasks, natural language understanding, out-of-distribution systematic generalization and efficient transfer learning. We have tried to clarify what some of these inductive biases may be, but much work needs to be done to improve that understanding and find appropriate ways to incorporate these priors in neural architectures and training frameworks.

10. Acknowledgements

The authors are grateful to Alex Lamb, Rosemary Nan Ke and Olexa Bilaniuk for leading many projects some of which are also discussed here. The authors are grateful to Mike Mozer for many brainstorming discussions. AG is also grateful to Bernhard Scholkopf, Sergey Levine, Matthew Botvinick and Charles Blundell for arranging research visits and collaborations which helped shape the intuitions regarding problems discussed here. The authors would also like to thank Stefan Bauer, Aniket Didolkar, Nasim Rahaman, Kanika Madan, Philippe Beaudoin for useful feedback. This document contains a review of AG's research as part of the requirement of his predoctoral exam, an overview of the main contributions of the author's few recent papers (co-authored with several other co-authors) as well as a vision of proposed future research.

References

- Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wüthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.
- Pablo Alvarez and Larry R Squire. Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the national academy of sciences*, 91(15):7041–7045, 1994.
- Bernard J Baars. *A cognitive theory of consciousness*. Cambridge University Press, 1993.
- Bernard J Baars. In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4):292–309, 1997.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints*, art. arXiv:1409.0473, Sep 2014.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: what is required and can it be learned? *arXiv preprint arXiv:1811.12889*, 2018.
- Dzmitry Bahdanau, Harm de Vries, Timothy J O’Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. Closure: Assessing systematic generalization of clevr models. *arXiv preprint arXiv:1912.05783*, 2019.
- Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autotutorials, 2019.
- Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*, 2017.
- David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv:1811.10597, ICLR’2019*, 2018.
- Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.
- Emmanuel Bengio, Valentin Thomas, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. *CoRR*, abs/1703.07718, 2017. URL <http://arxiv.org/abs/1703.07718>.
- Samy Bengio and Yoshua Bengio. Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks*, 11(3):550–557, 2000.

- Yoshua Bengio. Evolving culture versus local minima. In *Growing Adaptive Machines*, pages 109–138. Springer, 2014.
- Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. *Learning a synaptic learning rule*. Citeseer, 1990.
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, pages 932–938, 2001.
- Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv:1901.10912*, 2019.
- Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.
- Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.
- Leon Bottou. Learning representations using causal invariance. *ICLR Keynote Talk*, 2019.
- Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β -vae. *arXiv preprint arXiv:1804.03599*, 2018.
- Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.
- Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: First steps towards grounded language learning with a human in the loop. *arXiv preprint arXiv:1810.08272*, 2018.
- Jeff Clune. Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985*, 2019.
- Jonathan D Cohen, Matthew Botvinick, and Cameron S Carter. Anterior cingulate and prefrontal cortex: who’s in control? *Nature neuroscience*, 3(5):421–423, 2000.
- Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

- Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.
- S. Dehaene, H. Lau, and S. Kouider. What is consciousness, and could machines have it? *Science*, 358(6362):486–492, 2017.
- Stanislas Dehaene. *How We Learn: Why Brains Learn Better Than Any Machine... for Now*. Penguin, 2020.
- Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>.
- Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. An empirical investigation of the challenges of real-world reinforcement learning. *arXiv preprint arXiv:2003.11881*, 2020.
- Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.
- Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *arXiv preprint arXiv:1207.1389*, 2012.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.
- James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2), 1977.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms, 2019. URL [arXiv:1909.10893](https://arxiv.org/abs/1909.10893).
- Anirudh Goyal, Alex Lamb, Phanideep Gampa, Philippe Beaudoin, Sergey Levine, Charles Blundell, Yoshua Bengio, and Michael Mozer. Object files and schemata: Factorizing declarative and procedural knowledge in dynamical systems. *arXiv preprint arXiv:2006.16225*, 2020.
- Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

- Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. Dynamic neural turing machine with soft and hard addressing schemes. *arXiv preprint arXiv:1607.00036*, 2016.
- Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, et al. Hindsight credit assignment. In *Advances in neural information processing systems*, pages 12488–12497, 2019.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, et al. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*, 2017.
- Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.
- Geoffrey E Hinton. Distributed representations. 1984.
- R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.
- Nan Rosemary Ke, Anirudh Goyal ALIAS PARTH GOYAL, Olexa Bilaniuk, Jonathan Binas, Michael C Mozer, Chris Pal, and Yoshua Bengio. Sparse attentive backtracking: Temporal credit assignment through reminding. In *Advances in neural information processing systems*, pages 7640–7651, 2018.
- Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.
- Giancarlo Kerg, Bhargav Kanuparthi, Anirudh Goyal, Kyle Goyette, Yoshua Bengio, and Guillaume Lajoie. Untangling tradeoffs between recurrence and self-attention in neural networks. *arXiv preprint arXiv:2006.09471*, 2020.

- Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.
- Stanley Kok and Pedro Domingos. Learning the structure of markov logic networks. In *Proceedings of the 22nd international conference on Machine learning*, pages 441–448, 2005.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- Jan Kukačka, Vladimir Golkov, and Daniel Cremers. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*, 2017.
- Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
- Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2879–2888, 2018.
- Brenden M Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*, 2017.
- Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.
- Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*, 2019.
- Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess, and Thore Graepel. Emergent coordination through competition. *arXiv preprint arXiv:1902.07151*, 2019.
- Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124, 2019.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- Gary F Marcus. Rethinking eliminative connectionism. *Cognitive psychology*, 37(3):243–282, 1998.
- Gary F Marcus. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press, 2019.

- Sarthak Mittal, Alex Lamb, Anirudh Goyal, Vikram Voleti, Murray Shanahan, Guillaume Lajoie, Michael Mozer, and Yoshua Bengio. Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules. *arXiv preprint arXiv:2006.16981*, 2020.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.
- Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.
- Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of inference regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098, ICLR’2014*, 2013.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1988.
- Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, 2nd edition, 2009.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017a. ISBN 978-0-262-03731-0.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017b.
- Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *arXiv preprint arXiv:1802.03268*, 2018.
- Lorien Y Pratt. Discriminability-based transfer between neural networks. In *Advances in neural information processing systems*, pages 204–211, 1993.
- Lorien Y Pratt, Jack Mostow, Candace A Kamm, and Ace A Kamm. Direct transfer of learned information among neural networks. In *Aaai*, volume 91, pages 584–589, 1991.
- Rémi Le Priol, Reza Babanezhad Harikandeh, Yoshua Bengio, and Simon Lacoste-Julien. An analysis of the adaptation speed of causal models. *arXiv preprint arXiv:2005.09136*, 2020.
- Gabriel A Radvansky and Jeffrey M Zacks. Event boundaries in memory and cognition. *Current opinion in behavioral sciences*, 17:133–140, 2017.

- Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.
- Mark B Ring. Child: A first step towards continual learning. In *Learning to learn*, pages 261–292. Springer, 1998.
- Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.
- Adam Santoro, Ryan Faulkner, David Raposo, Jack W. Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy P. Lillicrap. Relational recurrent neural networks. *CoRR*, abs/1806.01822, 2018. URL <http://arxiv.org/abs/1806.01822>.
- Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.
- B. Schölkopf. Artificial intelligence: Learning to see and act. *Nature*, 518(7540):486–487, 2015.
- Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.
- John F Sowa. Semantic networks. 1987.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc., 2015. URL <http://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf>.
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. *arXiv preprint arXiv:1708.01289*, 2017.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharmashan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

- Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.
- David H Wolpert, William G Macready, et al. No free lunch theorems for search. Technical report, Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.
- Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
- Wenhao Yu, Jie Tan, C Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. *arXiv preprint arXiv:1702.02453*, 2017.
- Jeffrey M Zacks, Nicole K Speer, Khen M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007.
- Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.