PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Perspective

Cite this article: Goyal A, Bengio Y. 2022 Inductive biases for deep learning of higher-level cognition. *Proc. R. Soc. A* **478**: 20210068.
https://doi.org/10.1098/rspa.2021.0068

Received: 8 December 2021
Accepted: 13 September 2022

Subject Areas:
artificial intelligence

Keywords:
deep learning, causality, reasoning, systematic and out-of-distribution generalization, system 2

Authors for correspondence:
Anirudh Goyal
e-mail: anirudhgoyal9119@gmail.com
Yoshua Bengio
e-mail: yoshua.bengio@mila.quebec

'An invited Perspective to mark the election of Yoshua Bengio to the fellowship of the Royal Society in 2020.'

THE ROYAL SOCIETY PUBLISHING

Inductive biases for deep learning of higher-level cognition

Anirudh Goyal and Yoshua Bengio

Mila, University of Montreal, Montreal, Quebec Canada

AG, 0000-0002-4080-1940

A fascinating hypothesis is that human and animal intelligence could be explained by a few principles (rather than an encyclopaedic 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 behaviour 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 [1–3], 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

© 2022 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

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 (preferences, priors or assumptions), but that additional ones should be included 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 (OOD) generalization and transfer learning to new tasks with low sample complexity (few examples needed to generalize well). To make that concrete, we consider some of the inductive biases humans may exploit in conscious thought using 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 use the term *high-level* to talk about variables that are manipulated at the conscious level of processing and are thus generally verbalizable. However, humans can consciously focus attention on low-level or intermediate-level features, e.g. by describing an odd-coloured pixel, not just very abstract concepts like objects or social situations. We argue that the deep learning 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. Note, however, that although language may give us a view on system 2, these abilities are likely to pre-exist language as there is evidence of surprisingly strong forms of on-the-fly reasoning in some non-human animals, like corvids [4]. Our arguments suggest that while deep learning brought remarkable progress, it needs to be extended in qualitative and not just quantitative ways: larger and more diverse datasets and more computing resources [5] are important but insufficient without additional inductive biases [6–14]. We make the case that evolutionary forces, the interactions between multiple agents, the non-stationary and competition systems put pressure on the learning machinery to achieve the kind of flexibility, robustness and ability to adapt quickly which humans seem to have when they are faced with new environments [15–18] but needs to be improved with deep learning. The sought-after inductive biases should thus especially help AI to progress on these fronts. In addition to thinking about the learning and sample complexity advantage of these inductive biases, this paper links them with knowledge representation in neural networks, with the idea that by decomposing knowledge into its stable parts (like causal mechanisms) and volatile parts (random variables), and factorizing knowledge in small and somewhat independent pieces that can be recomposed dynamically as needed (to reason, imagine or explain at an *explicit* and verbalizable level), one may achieve the kind of systematic generalization which humans enjoy and is common in natural language [19–23].

(a) 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 labelled 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 [24–28], adversarial examples [29,30], spurious correlations [31–33], 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 [34]. Whereas multi-task learning usually just pools the different datasets [35–37], 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 OOD generalization), or with a few examples (i.e. with efficient transfer learning) [5,38–50].

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, just 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* [51] to characterize models which follow an underlying mechanistic understanding of reality. They are closely related to the structural causal models (SCMs) used to capture causal structure [52]. 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 '$\mathrm{GDP}_t = 1.05\,\mathrm{GDP}_{t-1} + \text{noise}$' 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 [53–56]. We can name high-level variables like position and velocity but may find it difficult to explain the intuitively known mechanisms which relate them to each other, in everyday life (by opposition to a physicist running a simulation of Newton's equations).

Implicit and explicit knowledge

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.

Humans frequently explain their perception (at the explicit level) and reason in terms of causal structure, and causal structure is really about how a joint distribution between causal random variables can change under interventions, i.e. actions. This suggests that one possible direction that deep learning needs to incorporate includes more notions about agency, reasoning 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 that 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. We claim that this perspective would help learning agents, such as babies or robots to succeed in the changing environments. This paper mostly discusses inductive biases inspired by higher-level cognition and aimed at facing these generalization challenges, pointing to existing work to implement some of them. However, for the most part, how to efficiently implement and combine these inductive biases in a single system remains an open question.

Table 1. Examples of current inductive biases in deep learning. Some have to do with the architecture while the last one influences the training framework and objective.

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
self-supervised pre-training	$P(X)$ is informative about $P(Y\|X)$

2. About inductive biases

The no-free-lunch theorem for machine learning [34,57] 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 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 assumptions or preferences about the solution we are looking for. An important 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 prioritize solutions with certain properties. Table 1 lists some of the inductive biases already used in various neural networks, and the corresponding properties. Although they are often expressed in terms of a neural architecture, they can also be about how the networks are trained, e.g. unsupervised pre-training, self-supervised learning and semi-supervised training, which all have to do with the input distribution $P(X)$ being informative about future tasks $P(Y|X)$. Other relevant elements which are not directly about inductive biases (and not discussed further in this paper) include for example the ability of a learning agent to actively seek knowledge (e.g. in active learning or RL) or to obtain information from other agents (e.g. social learning, multi-agent learning).

(a) From inductive biases to algorithms

There are many ways to encode such biases—e.g. explicit regularization objectives [58–62], architectural constraints [7,63–66], parameter sharing [67,68], implicit effects of the choice of the optimization method [69–71], self-supervised learning or self-supervised pre-training [72–77], invariance or equivariance to known transformations [78–83] or choices of prior distributions in a Bayesian model [84–87]. For example, one can build translation invariance of a neural network output by replacing matrix multiplication by convolutions [88] and pooling [89], or by averaging the network predictions over transformations of the input (feature averaging) [62], or by training on a dataset augmented with these transformations (data augmentation) [89]. 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 conversion often being the core contribution of machine learning papers (see table 2).

Table 2. Proposed additional inductive biases for deep learning: much progress has been made in learning representation of high-level variables (entities or objects). Much more progress is needed on other inductive biases such as the ones listed above. It would also be useful to think about integrating these inductive biases into a unified architecture.

inductive bias	corresponding property	relevant references
high-level variables play a causal role	learning representations of latent entities/attributes	[90–108]
changes in distribution are due to causal interventions	changes in distribution are localized in the appropriate semantic space	[103,109–113]
knowledge is generic, defined over abstract variables, and can be applied on different instances	factorizing knowledge in terms of abstract variables and functions that encapsulate how these variables interact with each other	[99,114,115]
sparsity of the factor graph	learned functions operate on a sparse set of variables (like arguments in typed-programming languages)	[99,114]
relevant causal chains tend to be very short (in time)	causal chains used to perform learning or inference (to obtain explanations or plans for achieving some goal) are broken down into short causal chains of events that may be far in time from each other	[116–121]
context-dependent processing involving goals, top-down influence and bottom-up competition	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	[122–124]

(b) 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 a lack of sufficiently powerful priors by more data [12]. Interestingly, different inductive biases may be equivalent to more or less data (even possibly exponentially more data): we suspect that inductive biases based on a form of compositionality (like distributed representations [125], depth [126] and attention [6,127]) 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 a few examples are available for the new distribution) are interesting to evaluate the advantage of inductive biases and of their implementation.

(c) 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 (iid), i.e. test data have 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 that 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 are 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 [102,128–131].

(d) Transfer learning and continual learning

Instead of a fixed data distribution and searching for an inductive bias that works well with this distribution, we are thus interested in transfer learning [132,133] and continual learning [134] 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. (i) what they may have in common, what is stable and stationary across the environments experienced and (ii) 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* [38,40,135,136], 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 [137]: think about the outer loop of evolution, the slightly faster loop of cultural learning [138], 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 that 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.

(e) Systematic generalization and OOD generalization

In this paper, we focus on the objective of 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 set-ups, 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 [19,23,139,140]. Systematic generalization is a phenomenon that was first studied in linguistics [21,22] because it is a core

property of language: the 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 it 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.* [22,141], 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 [142–144], 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 [125], 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 [126]. However, a form of compositionality which we propose here that should be better incorporated in deep learning is the form called systematicity [145] defined by linguists, and more recently systematic generalization in machine learning papers [22,146,147].

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, OOD. 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 better OOD

(a) 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 OOD abilities or in the context of causal environments changing due to interventions, e.g. as in [102]. Because they have to be very formal, AI models can also suggest hypotheses for how brains might implement an equivalent strategy with biological 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.

(b) Conscious versus 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 the passenger. Now imagine encountering a road block due to construction: you have to pay more attention, you have to be on the 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 [148,149]. For example in the driving example, when there is a road block you need to pay attention in order to think through what to do next, and you probably do not 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 that seems to be different—both functionally and in terms of the neural signature in the brain—when we deal with conscious processing and novel situations (changes in the distribution) which require our conscious attention, compared with our habitual routines. In those novel situations, we generally have to *think*, *focus* and *attend* to specific elements of our perception, actions or memories 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.

(i) System 1 and system 2

Cognitive scientists distinguish [150–155] *habitual* versus *controlled* processing, where the former correspond to default behaviours, whereas the latter require attention and *mental effort*. Daniel Kahneman introduced the framework of fast and slow thinking [156], 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 system 1 neural computations. The default (or unconscious) processing of system 1 can take place very rapidly (as fast as about 100 ms) 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 name it and reason about how it relates with other pieces of knowledge, 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 practise 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 practise 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.

(ii) The global workspace theory

The above division of labour is at the heart of the cognitive neuroscience global workspace theory (or GWT) from Baars [157,158] and its extension, the global neuronal workspace model [159–164]. The GWT suggests an architecture allowing specialist components to interact. The key claim of the GWT is the existence of a shared representation—sometimes called a blackboard, sometimes a workspace—that can be modified by any selected specialist and whose content is broadcast to all specialists. That selection is based on a form of attention and can correspond to dynamically selecting (based on the input) a module or a few modules in a modular neural net that are most appropriate for a particular context and task. 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 [159,165]. The GWT theory posits that conscious processing revolves around a communication bottleneck between selected parts of the brain that are called upon when addressing a current task. There is a threshold of relevance beyond which information that 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, forcing each module to learn to exchange with other modules in a way that allows swapping one module for another as a source or destination of communicated content, i.e. with a shared 'language'. These shared representations can be interpreted by many other modules. This gives rise to semantic representations that are not tied to a particular modality but can be triggered by any of the sensory channels. This makes it possible to flexibly obtain new combinations of pieces of knowledge, enabling a compositional advantage aligned with the needs of systematic generalization OOD.

(c) Attention as dynamic information flow

The GWT suggests a fleeting memory capacity in which only one consistent content can be dominant at any given moment, which suggests a sharper form of attention than the soft attention currently dominant in deep learning. 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 generating the next French word, we normally focus especially on the 'right' few words in the source English sentence that may be relevant to do the translation. This is the motivation that stimulated our work on content-based soft self-attention [127] 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.

(i) 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. With stochastic hard attention [166] one samples from a distribution over elements to choose the attended content, whereas with soft attention [127] one mixes these contents with different positive convex weights. 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 [5,167] and is the standard tool for memory-augmented neural networks [168–171]. Attention and memory can also help address the problem of credit assignment through long-term dependencies [116,120] by creating dynamic skip connections through time (i.e. a memory access) which unlock the problems of vanishing gradients and learning long-term dependencies [172,173]. 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 [6,171]. 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. With soft attention, one obtains a convex combination of the values in the rows of V, whereas stochastic hard attention would sample one of the value vectors with probability equal to that weight. If the soft attention is focused on one element for a particular row (i.e. the softmax is saturated), we get deterministic hard attention: only one of the objects is selected and its value copied 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. Note that hard attention is more biologically plausible (we only see one interpretation of the Necker cube [174] at once, and have one thought at a time) but soft attention enables end-to-end training and has been the most commonly used in deep learning architectures up to now, e.g. with transformers [6]. However, there is recent evidence [175] that if the communication bottleneck is discretized, better OOD generalization is observed, maybe because the simpler lingua franca would make it easier to swap one module for another in the attention-controlled communication between modules.

(ii) 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).

(iii) Attention implements variable binding

When the inputs and outputs of each of the modules 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 that 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 kth 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 head. 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, albeit in a soft differentiable way (that mixes multiple possiblities) in the case of soft attention.

(d) Blend of 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 and receive inputs from other modules, but when they do need to communicate information with another *arbitrary* module, the information has to go through a routing bottleneck (the global workspace) controlled by an attention mechanism. Because so few elements can be put in coherence at each step of the GWT selection, the 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). 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) [176,177]. 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.

(i) Partial state

From an RL perspective, it is interesting to note that if the GWT holds an important part of the state (including imagined future states, when planning), it does not describe all the aspects of the environment, only a handful of them, as already explored in the RL literature [178]. This is different from standard RL approaches where the input (or the sequence of past inputs) is mapped to a fixed-size (estimated and latent) state vector. The GWT suggests instead that, besides long-term memory content (which mostly does not change), the rapidly changing state should be seen as a very small set of entities (e.g. objects or particular attributes of objects, and their relation), with an information content similar to that of a single sentence. This suggests neural net architectures in which very few modules and specific (variable, value) pairs are selected at every inference step, based on those that were recently selected, the current sensory input and the current contents of memory (which can also compete for write-access to the workspace). 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.* [109] (see §3d above) where just a few relevant modules need to adapt to a change in distribution.

(ii) System 2 to system 1 consolidation

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 OOD in surprisingly 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 unfamiliar traffic regulations, which requires full conscious attention (§2b). 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 [179,180]. 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 practise it in order to be able to apply it, albeit awkwardly and slowly at first) and systematic

generalization, both of which should be important characteristics of the next generation of deep learning systems.

(iii) 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 [158,159,181]: 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 that 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 [182,183]: (i) modules that 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); and (ii) modules that 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 Baars' GTW. 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. 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. This enables a coherence-seeking mechanism: 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 machinery seeking to obtain coherent configurations of the variables interacting according to some piece of knowledge (a factor of the factor graph, a causal dependency).

System 2 inductive biases

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 versus system 2/explicit/conscious divide.

This paper is about inductive biases not yet sufficiently integrated in state-of-the-art deep learning systems but which could help us achieve these system 2 abilities. In the next section, we summarize some of these system 2 inductive biases.

(e) Semantic representations describe verbalizable concepts

Conscious content is revealed by reporting it, often with language [184]. This suggests that high-level variables manipulated consciously are closely related with their verbal forms (like words and phrases). This yields 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 (although we can also use words to identify some lower-level variables). 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, concept representation and continuous-valued properties) which is not always or not easily well reflected in their verbal rendering. Much of what our brains know actually cannot be easily translated in natural language and forms the content of system 1 knowledge. This means that system 2 (verbalizable) knowledge is incomplete: words are mostly pointers to knowledge which belongs to system 1 and thus is in great part not consciously accessible. The system 2 inductive biases do not need to cover all the aspects of our internal model of the world (they could not), 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 semantic variables that can be operated on at the system-2 level. 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).

High-level representations describe verbalizable concepts

There is a simple lossy mapping from semantic representations going through the GWT bottleneck to natural language expressions of them in natural language. This is an inductive bias which could be exploited in grounded language learning scenarios [130,185–187] where we couple language data with observations and actions by an agent.

This suggests that natural language understanding systems should be trained in a way that couples natural language with what it refers to. This is the idea of *grounded language learning*. IT would 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 do not force the top-level GWT 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. Our discussion about causality should also suggest that passive observation 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 [188,189]. Studying this kind of set-up was the motivation for our work on the Baby AI environment [130].

(f) Semantic variables play a causal role and knowledge about them is modular

Biological phenomena such as bird flocks 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 [158,164] 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 selected bits of information can squeeze at any time. If we believe that theory, these selected 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, a hallmark of classical AI based on rules [190] also makes sense as a requirement for obtaining systematic generalization [22]: 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 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.

Semantic variables are often also causal variables

We hypothesize that semantic variables are often 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 or actions). However, note that we can also name many low-level (like pixels) and intermediate features (like L-shaped edges). It is thus plausible to assume that causal reasoning of the kind we can verbalize involves as variables of interest those semantic variables which we can name, and that they can be at any level of the processing hierarchy in the brain, including at the highest levels of abstraction, where signals from all modalities join, such as pre-frontal cortex [191], and where concepts can be manipulated in a way that is not specific to a single modality.

The connection between causal representations and modularity is profound: an assumption which is commonly associated with SCMs is that it should break down the knowledge about the causal influences into *independent mechanisms* [24]. As explained in §3a, each such mechanism relates direct causes to their direct effect 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 statistical 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 OOD 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) may need to be adapted but that of the others do not need to be tuned 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 §3g) is different but related to a property of such high-level SCM: the sparsity of the graph capturing the joint distribution itself (discussed in §3i). 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 §3h.

(g) 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 (i) 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 longer than the lifetime of the learning agent) and (ii) 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 learning to play 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.

Changes in distribution are localized in the appropriate semantic space

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 existing 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 system 2 changes (due to interventions), we automatically get locality of the source of changes (which start at one or a few nodes of the causal graph), since by virtue of being localized in time and space, actions can only directly affect very few variables, with other effects (on downstream variables) 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.

(h) 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 would eventually converge. 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 §2.

Stable versus unstable properties of the world

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. An encoder that would implement 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 variable 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 brains 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 temporarily modify the causal graph structure (which specifies 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.

(i) Sparse factor graph in the space of semantic variables

Sparsity as to how variables and factors interact with each other

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 [192–194], 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. A factor may capture a causal mechanism between its argument variables, and thus we should introduce an additional semantic element to these factors: each argument of a causal 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 sparsity 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 a 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, e.g. to 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. Our interpretation of this restriction on write access to the GWT 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, e.g. consider loopy belief propagation [195,196]. By constraining the size of the working memory, evolution may have thus enforced the sparsity of the factor graph. The GWT also makes a claim that the workspace is associated with the conscious contents of cognition, which can be reported verbally. 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 (verbalizing 'how the world works', its dynamics and statistical or causal structure).

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 [197–201], 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 SCM, §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 have 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 [202].

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. Another reason for a formulation that is not constrained to an acyclic graph is that humans also reason about relations between variables at equilibrium (such as voltage and current), which can mutually be causes of each other (i.e. arrows can go both ways).

(j) 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 (DAG)) or factor (in a factor graph). There are extensions which allow parameter sharing, e.g. through time with dynamic Bayes nets [203], or in undirected graphical models such as Markov Networks [194] 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 inductive bias 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, 'schemas' or 'factor templates'. A schema, or generic factor is a reusable probabilistic relation, i.e. with argument variables, which can be bound to instances. A static instantiated 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 [204]. Natural language allows us to state rules involving variables and is not limited to making statements about specific instances.

Knowledge is generic and can be instantiated on different instances

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 do not need to hold in memory the full instantiated graph with all possible instances and all possible mechanisms relating them (or worse, all the generic factor instantiations that are compatible with the data, in a Bayesian posterior). Instead, inference involves generating the needed pieces of the graph and even performing 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 a 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 and they can reason 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, as an internal policy with attention and memory access as actions, 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 [157–159,205].

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' in cognitive psychology [206–208]). For example, you can

apply the same laws of physics to two different balls that are visually different (and maybe have different colours 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.

There are two forms of knowledge representation we have discussed: declarative knowledge or hypotheses, i.e. that can be verbalized (e.g. of facts, hypotheses, explicit causal dependencies, etc.), and inference machinery used to reason with these pieces of knowledge. Standard graphical models only represent the declarative knowledge and typically require expensive but generic iterative computations (such as Monte-Carlo Markov chains) to perform approximate inference [209,210]. However, brains need fast inference [211], 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. In general, searching for a good configuration of the values of top-level variables which is consistent with the given context 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). 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 inference 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 consolidation into cortex [212,213] so that they can be presumably committed to cortical long-term memory and fast inference.

(k) Relevant causal chains (for learning or inference) can be approximated as very short chains

In a *clock-based segmentation*, the boundaries between discrete time steps are spaced equally [214–216]. In an *event-based segmentation*, the boundaries depend on the state of the environment, resulting in dynamic-duration of intervals [217]. Our brains seem to segment streams of sensory inputs into meaningful representations of variable-length episodes and *events* [218–222].

The 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 travelling 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 that detects novel events is triggered for a large number of such events over the course of the trip. By contrast, few novel events occur on a familiar route. If our perception of time is event-based, meaning that higher centres 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.

Event segmentation allows functional representations that support temporal reasoning, an ability that arguably relies on neural circuits to encode and retrieve information to and from memory [223–225]. 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 inputs 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 sufficiently understood [226–228].

Relevant causal chains tend to be sparse

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. Causal chains used to perform learning (to imagine counterfactuals and 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.

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 [158]. 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 [116,120] is that this prior assumption is the strongest ingredient to bound the degree of vanishing of gradients.

(l) 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 [229–234]. Top-down information encodes 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 the 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).

Dynamic integration of bottom-up and top-down information

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.

4. Declarative knowledge of causal structure

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 or default 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.

(a) 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 DAG. In this graph structure, a variable (represented by a particular node) is independent of all the other variables, given all the direct neighbours of a variable. The edge direction identify a specific factorization of the joint distribution of the graph's variables:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{m} p(X_i \mid \mathbf{PA}_i). \tag{4.1}$$

(b) Structural causal models

A SCM [24] over a finite number M of random variables X_i given a set of *observables* $X_1, \ldots, X_M$ (modelled as random variables) associated with the vertices of a DAG G, is a set of structural assignments

$$X_i := f_i(X_{pa(i,C)}, N_i), \quad \forall i \in \{1, \ldots, M\}, \tag{4.2}$$

where f_i is a deterministic function, the set of noises $N_1, \ldots, N_m$ 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 data, possibly yielding a posterior distribution over causal structures compatible with the data, and a neural network can be trained to generate graphs from that posterior [235].

(c) 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 are generally needed [236].

(d) Independent causal mechanisms

A powerful assumption about how the world works which arises from research in causality [24] 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 the information theory 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 [52,237].

This principle applied to the factorization in equation (4.1), tells us that the different factors should be independent in the sense that (i) performing an intervention on one of the mechanisms $p(X_i|\mathbf{PA}_i)$ does not change any of the other mechanisms $p(X_j|\mathbf{PA}_j)$ $(i \neq j)$, (ii) knowing some other mechanisms $p(X_i|\mathbf{PA}_i)$ $(i \neq j)$ does not give us information about any another mechanism $p(X_j|\mathbf{PA}_j)$.

(i) Causal factor graphs

We propose that the formalism of directed graphical models, even extended to that of structural causal models (equation (4.2)), may not be consistent with the idea of ICM, and that parametrizing the causal structure with a particular form of factor graph (with directed edges to represent causal direction when there is one) could be more appropriate. Our argument is the following. If we force the conditionals $P(X_i|X_{pa(i,C)})$ (equivalently the parametrization of f_i in equation (4.2)) to be the independent units of parametrization, then we cannot represent other decompositions that may 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 SCMs. 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. Ordinary factor graphs are, however, missing the information about causal direction. They can, however, be extended by optionally associating with each argument of each factor's potential function an indicator to specify whether the argument acts as a cause or as an effect.

(e) Exploit changes in distribution due to causal interventions

(i) Nature does not 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' [238]. 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.

(f) 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.* [109]. 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 $A \to B$ (A *causes* B), or $B \to A$. 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, i.e. consistent with observational data of any size. In order to disambiguate between these two hypotheses, Bengio *et al.* [109] 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 $A \to B$, 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 $p(B \mid A)$ does not, i.e. $p(B \mid A) = \tilde{p}(B \mid A)$. Changes on the cause will be most informative, since they will have direct effects on B. Bengio *et al.* [109] find experimentally that this is sufficient to identify the causal graph, while [239] 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 $A \to B$ over the anti-causal $B \to A$, Bengio *et al.* [109] 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 on the example wise log-likelihood, 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 structure 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 10–30 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.* [109] 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 (which describes what is shared across environments and interventions). Here the meta-parameters capture the belief about the causal graph structure and the default (unperturbed) conditional dependencies, while the inner loop parameters are those which capture the change in the graph due to the intervention.

Ke *et al.* [113] further expanded this idea to deal with more than two variables. To model causal relations and OOD generalization one can view real-world distributions as arising from the composition of causal mechanisms. Any change in distribution (e.g. when moving from one setting/domain to a related one) is attributed to changes in as few as possible (but at least one) of those mechanisms [103,109,113]. A do-intervention or hard intervention would set the value of a variable to some value irrespective of the causal parents of that variable, thus disconnecting that node from its parents in the causal graph. By inferring this graph surgery, 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 [109,113,240] 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 to account for the intervention [239]. In this sense, learning causal representations may bring immediate benefits to machine learning models in terms of reduced sample complexity.

(g) 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 occuring 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. The agent would thus have a causal model of latent causal variables (how they influence each other and relate to each other), an intervention model relating low-level actions with interventions (or intentions to change specific high-level variables), as well as an observation model (relating high-level causal variables and sensory observations). In addition to these models, it would have inference machinery associated with them.

A human-centric version of this viewpoint is the psychological theory of affordances [241–243] that can be linked to predictive state representations in RL: 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 us 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 RL on controllability of aspects of the environment [244,245]. 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 most others (i.e. except as an effect of the causal variable which is being controlled). A learner thus needs to discover an intervention model (what actions give rise to what interventions), but this can also help the learner figure out a good representation space for causal variables.

5. Conclusion

To be able to handle dynamic, changing conditions, we want to move from deep statistical models which are 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, OOD 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. We have motivated these inductive biases in terms of expected (and observed in recent work) gains in terms of OOD generalization and fast adaptation in transfer settings rather than the standard test set from the same distribution as the training set. The general insight here is that the proposed inductive biases should help organize knowledge into the stable reusable parts that are likely to be useful in new settings and tasks (such as causal mechanisms), separating them from the more volatile pieces of information (the values of variables) that can be changed by agents (through causal intervention) or those that are affected by these changes and may vary across environments or tasks. Some of the more salient inductive biases we propose deserve especially more attention in deep learning research

include (i) the fairly direct connection between high-level variables and natural language or more generally how humans communicate knowledge among them, i.e. we can verbalize our thoughts to a large extent and this can provide rich insights about underlying inductive biases such as these: (ii) the modular decomposition of knowledge into independent reusable pieces that can be composed on the fly to address new contexts, (iii) the causal interpretation of actions by agents and of changes in distribution, with agents generally intending to affect a single or very few (generally latent) variables and (iv) the sparsity of dependencies between high-level variables (and thus the small number of variables that are linked by causal mechanisms imagined by humans to explain their environment). Finally, we would also like to mention that inductive biases are not the only way to bridge the gap to high-level human cognition: we may gain by improving our optimization algorithms, by scaling up neural networks [246] and by moving to other frameworks that better capture uncertainty about the world (e.g. by learning a Bayesian posterior over neural network models as compared with learning a point estimate). It would also be intriguing to think of ways to combine all these different components together.

Ethics. Research conducted in this study is purely technical. The authors do not foresee negative social impact of this work beyond that which could arise from general improvements in machine learning.

Data accessibility. This article has no additional data.

Authors' contributions. A.G.: conceptualization, methodology, writing—original draft; Y.B.: conceptualization, supervision, writing—original draft, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. The authors are grateful to NSERC, CIFAR, Google, Samsung, Nuance, IBM, Canada Research Chairs, Canada Graduate Scholarship Program, Nvidia for funding and Compute Canada for computing resources.

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 and Bernhard Schölkopf for many brainstorming discussions. The authors would also like to thank Stefan Bauer, Aniket Didolkar, Nasim Rahaman, Kanika Madan, Philippe Beaudoin, Charles Blundell, Dianbo Liu and Moksh Jain for useful feedback. The authors would also like to acknowledge comments by the reviewers which helped in improving the manuscript.

References

1. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009 Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE.
2. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M. 2013 Playing atari with deep reinforcement learning. (http://arxiv.org/abs/1312.5602)
3. Schrittwieser J *et al.* 2019 Mastering atari, go, chess and shogi by planning with a learned model. (http://arxiv.org/abs/1911.08265)
4. Taylor AH, Hunt GR, Medina FS, Gray RD. 2009 Do New Caledonian crows solve physical problems through causal reasoning?. *Proc. R. Soc. B* **276**, 247–254. (doi:10.1098/rspb.2008.1107)
5. Brown TB *et al.* 2020 Language models are few-shot learners. (http://arxiv.org/abs/2005.14165)
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017 Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008.
7. He K, Zhang X, Ren S, Sun J. 2016 Deep residual learning for image recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778.
8. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. 2017 Neural message passing for quantum chemistry. In *Proc. of the 34th Int. Conf. on Machine Learning-Volume 70*, pp. 1263–1272. JMLR. org.
9. Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, Dean J. 2017 Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. (http://arxiv.org/abs/1701.06538)

10. Fedus W, Zoph B, Shazeer N. 2021 Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. (https://arXiv:2101.03961)
11. Hinton G. 2021 How to represent part-whole hierarchies in a neural network. (http://arxiv.org/abs/2102.12627)
12. Welling M. 2019 Do we still need models or just more data and compute? University of Amsterdam, April 20.
13. Dosovitskiy A *et al.* 2020 An image is worth 16×16 words: transformers for image recognition at scale. (http://arxiv.org/abs/2010.11929)
14. Battaglia PW *et al.* 2018 Relational inductive biases, deep learning, and graph networks. (http://arxiv.org/abs/1806.01261)
15. Bansal T, Pachocki J, Sidor S, Sutskever I, Mordatch I. 2017 Emergent complexity via multi-agent competition. (http://arxiv.org/abs/1710.03748)
16. Liu S, Lever G, Merel J, Tunyasuvunakool S, Heess N, Graepel T. 2019 Emergent coordination through competition. (http://arxiv.org/abs/1902.07151)
17. Baker B, Kanitscheider I, Markov T, Wu Y, Powell G, McGrew B, Mordatch I. 2019 Emergent tool use from multi-agent autocurricula. (https://arXiv.1909.07528)
18. Leibo JZ, Hughes E, Lanctot M, Graepel T. 2019 Autocurricula and the emergence of innovation from social interaction: a manifesto for multi-agent intelligence research. (http://arxiv.org/abs/1903.00742)
19. Marcus GF. 1998 Rethinking eliminative connectionism. *Cognit. Psychol.* **37**, 243–282. (doi:10.1006/cogp.1998.0694)
20. Marcus GF. 2019 *The algebraic mind: integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.
21. Lake BM, Baroni M. 2017 Generalization without systematicity: on the compositional skills of sequence-to-sequence recurrent networks. (http://arxiv.org/abs/1711.00350)
22. Bahdanau D, Murty S, Noukhovitch M, Nguyen TH, de Vries H, Courville A. 2018 Systematic generalization: what is required and can it be learned?. (http://arxiv.org/abs/1811.12889)
23. McClelland JL, Rumelhart DE, PDP Research Group. 1987 *Parallel distributed processing, Volume 2: Explorations in the microstructure of cognition: psychological and biological models*, vol. 2. Cambridge, MA: MIT Press.
24. Peters J, Janzing D, Schölkopf B. 2017 *Elements of causal inference: foundations and learning algorithms*. Cambridge, MA: MIT press.
25. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA. 2020 Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673. (doi:10.1038/s42256-020-00257-z)
26. Hendrycks D *et al.* 2021 The many faces of robustness: a critical analysis of out-of-distribution generalization. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pp. 8340–8349.
27. Koh PW *et al.* 2021 Wilds: a benchmark of in-the-wild distribution shifts. In *Int. Conf. on Machine Learning*, pp. 5637–5664. PMLR.
28. Schneider S, Rusak E, Eck L, Bringmann O, Brendel W, Bethge M. 2020 Improving robustness against common corruptions by covariate shift adaptation. *Adv. Neural Inf. Process. Syst.* **33**, 11 539–11 551.
29. Goodfellow IJ, Shlens J, Szegedy C. 2014 Explaining and harnessing adversarial examples. (http://arxiv.org/abs/1412.6572)
30. Kurakin A, Goodfellow I, Bengio S. 2016 Adversarial examples in the physical world. (http://arxiv.org/abs/1607.02533)
31. Krueger D, Caballero E, Jacobsen JH, Zhang A, Binas J, Zhang D, Le Priol R, Courville A. 2021 Out-of-distribution generalization via risk extrapolation (REx). In *Int. Conf. on Machine Learning*, pp. 5815–5826. PMLR.
32. Beery S, Van Horn G, Perona P. 2018 Recognition in terra incognita. In *Proc. of the European conf. on computer vision (ECCV)*, pp. 456–473.
33. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. 2019 Invariant risk minimization. (http://arxiv.org/abs/1907.02893)
34. Baxter J. 2000 A model of inductive bias learning. *J. Artif. Intell. Res.* **12**, 149–198. (doi:10.1613/jair.731)
35. Caruana R. 1997 Multitask learning. *Mach. Learn.* **28**, 41–75. (doi:10.1023/A:1007379606734)
36. Collobert R, Weston J. 2008 A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proc. of the 25th Int. Conf. on Machine Learning*, pp. 160–167.

37. Ruder S. 2017 An overview of multi-task learning in deep neural networks. (http://arxiv.org/abs/1706.05098)
38. Ravi S, Larochelle H. 2016 Optimization as a model for few-shot learning.
39. Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, Blundell C, Kumaran D, Botvinick M. 2016 Learning to reinforcement learn. (http://arxiv.org/abs/1611.05763)
40. Finn C, Abbeel P, Levine S. 2017 Model-agnostic meta-learning for fast adaptation of deep networks. (http://arxiv.org/abs/1703.03400)
41. Cabi S *et al.* 2019 Scaling data-driven robotics with reward sketching and batch reinforcement learning. (http://arxiv.org/abs/1909.12200)
42. Jang E, Irpan A, Khansari M, Kappler D, Ebert F, Lynch C, Levine S, Finn C. 2022 Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conf. on Robot Learning*, pp. 991–1002. PMLR.
43. Reed S *et al.* 2022 A generalist agent. (http://arxiv.org/abs/2205.06175)
44. Ahn M *et al.* 2022 Do as I can, not as I say: grounding language in robotic affordances. (http://arxiv.org/abs/2204.01691)
45. Alayrac JB *et al.* 2022 Flamingo: a visual language model for few-shot learning. (http://arxiv.org/abs/2204.14198)
46. Borgeaud S *et al.* 2022 Improving language models by retrieving from trillions of tokens. In *Int. Conf. on Machine Learning*, pp. 2206–2240. PMLR.
47. Chowdhery A *et al.* 2022 Palm: scaling language modeling with pathways. (http://arxiv.org/abs/2204.02311)
48. Sanh V *et al.* 2021 Multitask prompted training enables zero-shot task generalization. (http://arxiv.org/abs/2110.08207)
49. Lu J, Clark C, Zellers R, Mottaghi R, Kembhavi A. 2022 Unified-IO: a unified model for vision, language, and multi-modal tasks. (http://arxiv.org/abs/2206.08916)
50. Raffel C *et al.* 2020 Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67.
51. Schölkopf B. 2015 Artificial intelligence: learning to see and act. *Nature* **518**, 486–487. (doi:10.1038/518486a)
52. Pearl J. 2009 *Causality: models, reasoning, and inference*, 2nd edn. New York, NY: Cambridge University Press.
53. McCloskey M. 1983 Intuitive physics. *Sci. Am.* **248**, 122–131. (doi:10.1038/scientificamerican0483-122)
54. Baillargeon R, Spelke ES, Wasserman S. 1985 Object permanence in five-month-old infants. *Cognition* **20**, 191–208. (doi:10.1016/0010-0277(85)90008-3)
55. Spelke ES, Breinlinger K, Macomber J, Jacobson K. 1992 Origins of knowledge. *Psychol. Rev.* **99**, 605. (doi:10.1037/0033-295X.99.4.605)
56. Battaglia PW, Hamrick JB, Tenenbaum JB. 2013 Simulation as an engine of physical scene understanding. *Proc. Natl Acad. Sci. USA* **110**, 18 327–18 332. (doi:10.1073/pnas.1306572110)
57. Wolpert DH, Macready WG. 1995 No free lunch theorems for search. Technical report Technical Report SFI-TR-95-02-010, Santa Fe Institute.
58. Bishop CM. 1995 *Neural networks for pattern recognition*. Oxford, NJ: Oxford University Press.
59. Bishop CM. 1995 Training with noise is equivalent to Tikhonov regularization. *Neural Comput.* **7**, 108–116. (doi:10.1162/neco.1995.7.1.108)
60. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014 Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958.
61. Kukačka J, Golkov V, Cremers D. 2017 Regularization for deep learning: a taxonomy. (http://arxiv.org/abs/1710.10686)
62. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. 2017 mixup: beyond empirical risk minimization. (http://arxiv.org/abs/1710.09412)
63. Yu F, Koltun V. 2015 Multi-scale context aggregation by dilated convolutions. (http://arxiv.org/abs/1511.07122)
64. Long J, Shelhamer E, Darrell T. 2015 Fully convolutional networks for semantic segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3431–3440.
65. Dumoulin V, Visin F. 2016 A guide to convolution arithmetic for deep learning. (http://arxiv.org/abs/1603.07285)
66. Huang G, Liu Z, Van D, Weinberger KQ. 2017 Densely connected convolutional networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4700–4708.

67. Hochreiter S, Schmidhuber J. 1997 Long short-term memory. *Neural Comput.* **9**, 1735–1780. (doi:10.1162/neco.1997.9.8.1735)
68. Pham H, Guan MY, Zoph B, Le QV, Dean J. 2018 Efficient neural architecture search via parameter sharing. (http://arxiv.org/abs/1802.03268)
69. Jastrzębski S, Kenton Z, Arpit D, Ballas N, Fischer A, Bengio Y, Storkey A. 2017 Three factors influencing minima in SGD. (http://arxiv.org/abs1711.04623)
70. Smith SL, Le QV. 2017 A bayesian perspective on generalization and stochastic gradient descent. (http://arxiv.org/abs/1710.06451)
71. Chaudhari P, Soatto S. 2018 Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE.
72. Hinton GE, Osindero S, Teh YW. 2006 A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554. (doi:10.1162/neco.2006.18.7.1527)
73. Erhan D, Courville A, Bengio Y, Vincent P. 2010 Why does unsupervised pre-training help deep learning?. In *Proc. of the Thirteenth Int. Conf. on Artificial Intelligence and Statistics*, pp. 201–208. JMLR Workshop and Conference Proceedings.
74. Devlin J, Chang MW, Lee K, Toutanova K. 2018 Bert: pre-training of deep bidirectional transformers for language understanding. (http://arxiv.org/abs/1810.04805)
75. Chen T, Kornblith S, Norouzi M, Hinton G. 2020 A simple framework for contrastive learning of visual representations. In *Int. Conf. on Machine Learning*, pp. 1597–1607. PMLR.
76. Chen X, Fan H, Girshick R, He K. 2020 Improved baselines with momentum contrastive learning. (http://arxiv.org/abs/2003.04297)
77. Grill JB *et al.* 2020 Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **33**, 21 271–21 284.
78. Bruna J, Zaremba W, Szlam A, LeCun Y. 2013 Spectral networks and locally connected networks on graphs. (http://arxiv.org/abs/1312.6203)
79. Defferrard M, Bresson X, Vandergheynst P. 2016 Convolutional neural networks on graphs with fast localized spectral filtering. *Adv. Neural Inf. Process. Syst.* **29**.
80. Ravanbakhsh S, Schneider J, Poczos B. 2017 Equivariance through parameter-sharing. In *Int. Conf. on Machine Learning*, pp. 2892–2901. PMLR.
81. Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K, Riley P. 2018 Tensor field networks: rotation-and translation-equivariant neural networks for 3D point clouds. (http://arxiv.org/abs/1802.08219)
82. Finzi M, Stanton S, Izmailov P, Wilson AG. 2020 Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *Int. Conf. on Machine Learning*, pp. 3165–3176. PMLR.
83. Satorras VG, Hoogeboom E, Welling M. 2021 E (n) equivariant graph neural networks. In *Int. Conf. on Machine Learning*, pp. 9323–9332. PMLR.
84. Jeffreys H. 1946 An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A.* **186**, 453–461. (doi:10.1098/rspa.1946.0056)
85. Berger JO, Bernardo JM. 1992 On the development of the reference prior method. *Bayesian Stat.* **4**, 35–60.
86. Gelman A. 1996 Bayesian model-building by pure thought: some principles and examples. *Stat. Sin.* **6**, 215–232.
87. Fortuin V. 2022 Priors in bayesian deep learning: a review. *Int. Stat. Rev.* (https://doi.org/10.1111/insr.12502)
88. LeCun Y, Bengio Y. 1995 Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **3361**, 1995.
89. Krizhevsky A, Sutskever I, Hinton GE. 2012 Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105.
90. Le Roux N, Heess N, Shotton J, Winn J. 2011 Learning a generative model of images by factoring appearance and shape. *Neural Comput.* **23**, 593–650. (doi:10.1162/NECO_a_00086)
91. Eslami S, Heess N, Weber T, Tassa Y, Szepesvari D, Kavukcuoglu K, Hinton GE. 2016 Attend, infer, repeat: fast scene understanding with generative models. (http://arxiv.org/abs/1603.08575)
92. Greff K, Rasmus A, Berglund M, Hao TH, Schmidhuber J, Valpola H. 2016 Tagger: deep unsupervised perceptual grouping. (http://arxiv.org/abs/1606.06724)
93. Raposo D, Santoro A, Barrett D, Pascanu R, Lillicrap T, Battaglia P. 2017 Discovering objects and their relations from entangled scene representations. (http://arxiv.org/abs/1702.05068)

94. Van Steenkiste S, Chang M, Greff K, Schmidhuber J. 2018 Relational neural expectation maximization: unsupervised discovery of objects and their interactions. (http://arxiv.org/abs/1802.10353)
95. Kosiorek A, Kim H, Teh YW, Posner I. 2018 Sequential attend, infer, repeat: generative modelling of moving objects. *Adv. Neural Inf. Process. Syst.* **31**, 8606–8616.
96. Engelcke M, Kosiorek AR, Jones OP, Posner I. 2019 Genesis: generative scene inference and sampling with object-centric latent representations. (http://arxiv.org/abs/1907.13052)
97. Burgess CP, Matthey L, Watters N, Kabra R, Higgins I, Botvinick M, Lerchner A. 2019 Monet: unsupervised scene decomposition and representation. (http://arxiv.org/abs/1901.11390)
98. Goyal A, Lamb A, Gampa P, Beaudoin P, Levine S, Blundell C, Bengio Y, Mozer M. 2020 Object files and schemata: factorizing declarative and procedural knowledge in dynamical systems. (http://arxiv.org/abs/2006.16225)
99. Ke NR *et al.* 2021 Systematic evaluation of causal discovery in visual model based reinforcement learning. (http://arxiv.org/abs/2107.00848)
100. Greff K, Kaufman RL, Kabra R, Watters N, Burgess C, Zoran D, Matthey L, Botvinick M, Lerchner A. 2019 Multi-object representation learning with iterative variational inference. In *Int. Conf. on Machine Learning*, pp. 2424–2433.
101. Locatello F, Weissenborn D, Unterthiner T, Mahendran A, Heigold G, Uszkoreit J, Dosovitskiy A, Kipf T. 2020 Object-centric learning with slot attention. (http://arxiv.org/abs/2006.15055)
102. Ahmed O, Träuble F, Goyal A, Neitz A, Wüthrich M, Bengio Y, Schölkopf B, Bauer S. 2020 Causalworld: a robotic manipulation benchmark for causal structure and transfer learning. (http://arxiv.org/abs/2010.04296)
103. Goyal A, Lamb A, Hoffmann J, Sodhani S, Levine S, Bengio Y, Schölkopf B. 2019 Recurrent independent mechanisms. (https://arXiv.1909.10893)
104. Zablotskaia P, Dominici EA, Sigal L, Lehrmann AM. 2020 Unsupervised video decomposition using spatio-temporal iterative inference. (http://arxiv.org/abs/2006.14727)
105. Rahaman N, Goyal A, Gondal MW, Wuthrich M, Bauer S, Sharma Y, Bengio Y, Schölkopf B. 2020 S2RMs: spatially structured recurrent modules. (http://arxiv.org/abs/2007.06533)
106. Du Y, Smith K, Ulman T, Tenenbaum J, Wu J. 2020 Unsupervised discovery of 3D physical objects from video. (http://arxiv.org/abs/2007.12348)
107. Ding D, Hill F, Santoro A, Botvinick M. 2020 Object-based attention for spatio-temporal reasoning: outperforming neuro-symbolic models with flexible distributed architectures. (http://arxiv.org/abs/2012.08508)
108. Goyal A *et al.* 2021 Coordination among neural modules through a shared global workspace. (http://arxiv.org/abs/2103.01197)
109. Bengio Y, Deleu T, Rahaman N, Ke R, Lachapelle S, Bilaniuk O, Goyal A, Pal C. 2019 A meta-transfer objective for learning to disentangle causal mechanisms. (http://arxiv.org/abs/1901.10912)
110. Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. 2012 On causal and anticausal learning. (http://arxiv.org/abs/1206.6471)
111. Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y. 2021 Toward causal representation learning. *Proc. IEEE* **109**, 612–634. (doi:10.1109/JPROC.2021.3058954)
112. Ke NR, Wang J, Mitrovic J, Szummer M, Rezende DJ. 2020 Amortized learning of neural causal representations. (http://arxiv.org/abs/2008.09301)
113. Ke NR, Bilaniuk O, Goyal A, Bauer S, Larochelle H, Pal C, Bengio Y. 2019 Learning neural causal models from unknown interventions. (http://arxiv.org/abs/1910.01075)
114. Alias Parth Goyal AG, Didolkar A, Ke NR, Blundell C, Beaudoin P, Heess N, Mozer MC, Bengio Y. 2021 Neural production systems. *Adv. Neural Inf. Process. Syst.* **34**, 25 673–25 687.
115. Silver T, Chitnis R, Kumar N, McClinton W, Lozano-Perez T, Kaelbling LP, Tenenbaum J. 2022 Inventing relational state and action abstractions for effective and efficient Bilevel planning. (http://arxiv.org/abs/2203.09634)
116. Ke NR, GOYAL AGAP, Bilaniuk O, Binas J, Mozer MC, Pal C, Bengio Y. 2018 Sparse attentive backtracking: temporal credit assignment through reminding. In *Advances in Neural Information Processing Systems*, pp. 7640–7651.
117. Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. 2020 Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**, 335–346. (doi:10.1038/s41583-020-0277-3)
118. Arjona-Medina JA, Gillhofer M, Widrich M, Unterthiner T, Brandstetter J, Hochreiter S. 2019 Rudder: return decomposition for delayed rewards. *Adv. Neural Inf. Process. Syst.* **32**.

119. Patil VP, Hofmarcher M, Dinu MC, Dorfer M, Blies PM, Brandstetter J, Arjona-Medina JA, Hochreiter S. 2020 Align-rudder: learning from few demonstrations by reward redistribution. (http://arxiv.org/abs/2009.14108)
120. Kerg G, Kanuparthi B, Goyal A, Goyette K, Bengio Y, Lajoie G. 2020 Untangling tradeoffs between recurrence and self-attention in neural networks. (http://arxiv.org/abs/2006.09471)
121. Goyal A *et al.* 2022 Retrieval-augmented reinforcement learning. In *Int. Conf. on Machine Learning*, pp. 7740–7765. PMLR.
122. Mittal S, Lamb A, Goyal A, Voleti V, Shanahan M, Lajoie G, Mozer M, Bengio Y. 2020 Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules. (http://arxiv.org/abs/2006.16981)
123. Fan A, Lavril T, Grave E, Joulin A, Sukhbaatar S. 2020 Addressing some limitations of transformers with feedback memory. (http://arxiv.org/abs/2002.09402)
124. McClelland JL, Hill F, Rudolph M, Baldridge J, Schütze H. 2020 Placing language in an integrated understanding system: next steps toward human-level performance in neural language models. *Proc. Natl Acad. Sci. USA* **117**, 25 966–25 974. (doi:10.1073/pnas.1910416117)
125. Pascanu R, Montufar G, Bengio Y. 2013 On the number of inference regions of deep feed forward networks with piece-wise linear activations. (http://arxiv.org/abs/1312.6098), ICLR'2014.
126. Montufar GF, Pascanu R, Cho K, Bengio Y. 2014 On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2924–2932.
127. Bahdanau D, Cho K, Bengio Y. 2014 Neural machine translation by jointly learning to align and translate. (http://arxiv.org/abs/1409.0473).
128. Yu W, Tan J, Liu CK, Turk G. 2017 Preparing for the unknown: learning a universal policy with online system identification. (http://arxiv.org/abs/1702.02453)
129. Packer C, Gao K, Kos J, Krähenbühl P, Koltun V, Song D. 2018 Assessing generalization in deep reinforcement learning. (http://arxiv.org/abs/1810.12282)
130. Chevalier-Boisvert M, Bahdanau D, Lahlou S, Willems L, Saharia C, Nguyen TH, Bengio Y. 2018 BabyAI: first steps towards grounded language learning with a human in the loop. (http://arxiv.org/abs/1810.08272)
131. Dulac-Arnold G, Levine N, Mankowitz DJ, Li J, Paduraru C, Gowal S, Hester T. 2020 An empirical investigation of the challenges of real-world reinforcement learning. (http://arxiv.org/abs/2003.11881)
132. Pratt LY, Mostow J, Kamm CA, Kamm AA. 1991 Direct transfer of learned information among neural networks. In *Aaai*, vol. 91, pp. 584–589.
133. Pratt LY. 1993 Discriminability-based transfer between neural networks. In *Advances in Neural Information Processing Systems*, pp. 204–211.
134. Ring MB. 1998 CHILD: a first step towards continual learning. In *Learning to learn*, pp. 261–292. Springer.
135. Bengio Y, Bengio S, Cloutier J. 1990 *Learning a synaptic learning rule*. Citeseer.
136. Schmidhuber J. 1987 *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-· · · hook*. PhD thesis Technische Universität München.
137. Clune J. 2019 Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. (http://arxiv.org/abs/1905.10985)
138. Bengio Y. 2014 Evolving culture versus local minima. In *Growing Adaptive Machines*, pp. 109–138. Springer.
139. Smolensky P. 1988 On the proper treatment of connectionism. *Behav. Brain Sci.* **11**, 1–23. (doi:10.1017/S0140525X00052432)
140. Fodor JA, Pylyshyn ZW. 1988 Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71. (doi:10.1016/0010-0277(88)90031-5)
141. Bahdanau D, de Vries H, O'Donnell TJ, Murty S, Beaudoin P, Bengio Y, Courville A. 2019 CLOSURE: assessing systematic generalization of CLEVR models. (http://arxiv.org/abs/1912.05783)
142. Hinton GE. 1984 Distributed representations.
143. Bengio S, Bengio Y. 2000 Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Trans. Neural Netw.* **11**, 550–557. (doi:10.1109/72.846725)
144. Bengio Y, Ducharme R, Vincent P. 2001 A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, pp. 932–938.

145. Lake B, Baroni M. 2018 Generalization without systematicity: on the compositional skills of sequence-to-sequence recurrent networks. In *Int. Conf. on Machine Learning*, pp. 2879–2888.
146. Ruis L, Andreas J, Baroni M, Bouchacourt D, Lake BM. 2020 A benchmark for systematic generalization in grounded language understanding. *Adv. Neural Inf. Process. Syst.* **33**, 19 861–19 872.
147. Akyürek E, Akyürek AF, Andreas J. 2020 Learning to recombine and resample data for compositional generalization. (http://arxiv.org/abs/2010.03706)
148. Carlson RA, Dulany DE. 1985 Conscious attention and abstraction in concept learning. *J. Exp. Psychol.: Learn. Mem. Cogn.* **11**, 45.
149. Newman J, Baars BJ, Cho SB. 1997 A neural global workspace model for conscious attention. *Neural Netw.* **10**, 1195–1206. (doi:10.1016/S0893-6080(97)00060-9)
150. Schneider W, Dumais ST, Shiffrin RM. 1982 Automatic/control processing and attention. Technical report Illinois Univ Champaign Human Attention Research Lab.
151. Redgrave P, Rodriguez M, Smith Y, Rodriguez-Oroz MC, Lehericy S, Bergman H, Agid Y, DeLong MR, Obeso JA. 2010 Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nat. Rev. Neurosci.* **11**, 760–772. (doi:10.1038/nrn2915)
152. Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. 2001 Conflict monitoring and cognitive control. *Psychol. Rev.* **108**, 624. (doi:10.1037/0033-295X.108.3.624)
153. Botvinick MM, Braver TS, Carter C, Barch D, Cohen J. 2001 Evaluating the demand for control: anterior cingulate cortex and crosstalk monitoring. *Psychol. Rev.* **108**, 624–652. (doi:10.1037/0033-295X.108.3.624)
154. Mozer MC, Colagrosso M, Huber D. 2001 A rational analysis of cognitive control in a speeded discrimination task. *Adv. Neural Inf. Process. Syst.* **14, 51–57**. (doi:10.1037/e537102012-775)
155. Bargh JA. 1984 Automatic and conscious processing of social information. In *American Psychological Association convention, 1982, Washington, DC, US; Portions of the research discussed in this chapter were presented at the aforementioned conference, and at the 1982 meetings of the Society for Experimental Social Psychology in Nashville, Indiana*. Lawrence Erlbaum Associates Publishers.
156. Kahneman D. 2011 *Thinking, fast and slow*. New York, NY: Macmillan.
157. Baars BJ. 1993 *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.
158. Baars BJ. 1997 In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *J. Conscious. Stud.* **4**, 292–309.
159. Shanahan M. 2006 A cognitive architecture that combines internal simulation with a global workspace. *Conscious Cogn.* **15**, 433–449. (doi:10.1016/j.concog.2005.11.005)
160. Shanahan M. 2010 *Embodiment and the inner life: cognition and consciousness in the space of possible minds*. New York, NY: Oxford University Press.
161. Shanahan M. 2012 The brain's connective core and its role in animal cognition. *Phil. Trans. R. Soc. B* **367**, 2704–2714. (doi:10.1098/rstb.2012.0128)
162. Dehaene S, Changeux JP. 2011 Experimental and theoretical approaches to conscious processing. *Neuron* **70**, 200–227. (doi:10.1016/j.neuron.2011.03.018)
163. Dehaene S, Lau H, Kouider S. 2017 What is consciousness, and could machines have it? *Science* **358**, 486–492. (doi:10.1126/science.aan8871)
164. Dehaene S. 2020 *How we learn: why brains learn better than any machine. For now*. New York, NY: Viking.
165. Shanahan M. 2005 Consciousness, emotion, and imagination. *Approaches to Machine Consciousness*, p. 26.
166. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. 2015 Show, attend and tell: neural image caption generation with visual attention. In *Int. Conf. on Machine Learning*, pp. 2048–2057.
167. Devlin J, Chang M, Lee K, Toutanova K. 2018 BERT: pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* (http://arxiv.org/abs/1810.04805)
168. Graves A, Wayne G, Danihelka I. 2014 Neural turing machines. (http://arxiv.org/abs/1410.5401)
169. Sukhbaatar S, Fergus R. 2015 End-to-end memory networks. In *Advances in Neural Information Processing Systems 28* (eds C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett), pp. 2440–2448. Curran Associates, Inc.

170. Gulcehre C, Chandar S, Cho K, Bengio Y. 2016 Dynamic neural turing machine with soft and hard addressing schemes. (http://arxiv.org/abs/1607.00036)
171. Santoro A *et al*. 2018 Relational recurrent neural networks. *CoRR* (http://arxiv.org/abs/1806.01822)
172. Hochreiter S. 1991 Untersuchungen zu dynamischen neuronalen Netzen [in German] Diploma thesis. *TU Münich*.
173. Bengio Y, Simard P, Frasconi P. 1994 Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **5**, 157–166. (doi:10.1109/72.279181)
174. Cohen L. 1959 Rate of apparent change of a Necker cube as a function of prior stimulation. *Am. J. Psychol.* **72**, 327–344. (doi:10.2307/1420037)
175. Liu D, Lamb AM, Kawaguchi K, Alias Parth Goyal AG, Sun C, Mozer MC, Bengio Y. 2021 Discrete-valued neural communication. *Adv. Neural Inf. Process. Syst.* **34**, 2109–2121.
176. Baddeley A. 1992 Working memory. *Science* **255**, 556–559. (doi:10.1126/science.1736359)
177. Cowan N. 1999 An embedded-processes model of working memory.
178. Zhao M, Liu Z, Luan S, Zhang S, Precup D, Bengio Y. 2021 A consciousness-inspired planning agent for model-based reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34**, 1569–1581.
179. Shenhav A *et al*. 2017 Toward a rational and mechanistic account of mental effort. *Annu. Rev. Neurosci.* **40**, 99–124. (doi:10.1146/annurev-neuro-072116-031526)
180. Kool W, Botvinick M. 2018 Mental labour. *Nat. Hum. Behav.* **2**, 899–908. (doi:10.1038/s41562-018-0401-9)
181. Koch C. 2004 The quest for consciousness a neurobiological approach. Boston, MA: Roberts and Co.
182. Watts DJ, Strogatz SH. 1998 Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442. (doi:10.1038/30918)
183. Latora V, Marchiori M. 2001 Efficient behavior of small-world networks. *Phys. Rev. Lett.* **87**, 198701. (doi:10.1103/PhysRevLett.87.198701)
184. Colagrosso M, Mozer MC. 2004 Theories of access consciousness. *Adv. Neural Inf. Process. Syst.* **17**.
185. Winograd T. 1972 Understanding natural language. *Cognit. Psychol.* **3**, 1–191. (doi:10.1016/0010-0285(72)90002-3)
186. Hermann KM *et al*. 2017 Grounded language learning in a simulated 3D world. (http://arxiv.org/abs/1706.06551)
187. Hill F, Lampinen A, Schneider R, Clark S, Botvinick M, McClelland JL, Santoro A. 2019 Environmental drivers of systematicity and generalization in a situated agent. (http://arxiv.org/abs/1910.00571)
188. Binz M, Schulz E. 2022 Using cognitive psychology to understand GPT-3. (http://arxiv.org/abs/2206.14576)
189. Kosoy E, Chan DM, Liu A, Collins J, Kaufmann B, Huang SH, Hamrick JB, Canny J, Ke NR, Gopnik A. 2022 Towards understanding how machines can learn causal overhypotheses. (http://arxiv.org/abs/2206.08353)
190. Russell PN. 2010 Artificial intelligence: a modern approach by stuart. *Russell and Peter Norvig contributing writers, Ernest Davis···[et al.]*.
191. Cohen JD, Botvinick M, Carter CS. 2000 Anterior cingulate and prefrontal cortex: who's in control?. *Nat. Neurosci.* **3**, 421–423. (doi:10.1038/74783)
192. Kschischang FR, Frey BJ, Loeliger HA. 2001 Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory* **47**, 498–519. (doi:10.1109/18.910572)
193. Frey BJ. 2012 Extending factor graphs so as to unify directed and undirected graphical models. (http://arxiv.org/abs/1212.2486)
194. Kok S, Domingos P. 2005 Learning the structure of Markov logic networks. In *Proc. of the 22nd Int. Conf. on Machine Learning*, pp. 441–448.
195. Frey BJ, Koetter R, Petrovic N. 2001 Very loopy belief propagation for unwrapping phase images. *Adv. Neural Inf. Process. Syst.* **14**. (doi:10.7551/mitpress/1120.003.0099)
196. Murphy K, Weiss Y, Jordan MI. 2013 Loopy belief propagation for approximate inference: an empirical study. (http://arxiv.org/abs/1301.6725)
197. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A. 2016 beta-VAE: learning basic visual concepts with a constrained variational framework.
198. Burgess CP, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, Lerchner A. 2018 Understanding disentangling in \beta-VAE. (http://arxiv.org/abs/1804.03599)

199. Chen RT, Li X, Grosse RB, Duvenaud DK. 2018 Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620.
200. Kim H, Mnih A. 2018 Disentangling by factorising. (http://arxiv.org/abs/1802.05983)
201. Locatello F, Bauer S, Lucic M, Raetsch G, Gelly S, Schölkopf B, Bachem O. 2019 Challenging common assumptions in the unsupervised learning of disentangled representations. In *Int. Conf. on Machine Learning*, pp. 4114–4124.
202. Bau D, Zhu JY, Strobelt H, Zhou B, Tenenbaum JB, Freeman WT, Torralba A. 2018 Gan dissection: visualizing and understanding generative adversarial networks. (http://arxiv.org/abs/1811.10597), ICLR'2019.
203. Spirtes P, Glymour CN, Scheines R, Heckerman D. 2000 *Causation, prediction, and search*. Cambridge, MA: MIT Press.
204. Sowa JF. 1987 Semantic networks.
205. Shanahan M, Baars B. 2005 Applying global workspace theory to the frame problem. *Cognition* **98**, 157–176. (doi:10.1016/j.cognition.2004.11.007)
206. Noles NS, Scholl BJ, Mitroff SR. 2005 The persistence of object file representations. *Percept. Psychophys.* **67**, 324–334. (doi:10.3758/BF03206495)
207. Gordon RD, Irwin DE. 1996 What's in an object file? Evidence from priming studies. *Percept. Psychophys.* **58**, 1260–1277. (doi:10.3758/BF03207558)
208. Kahneman D, Treisman A, Gibbs BJ. 1992 The reviewing of object files: object-specific integration of information. *Cognit. Psychol.* **24**, 175–219. (doi:10.1016/0010-0285(92)90007-O)
209. Cowles MK, Carlin BP. 1996 Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.* **91**, 883–904. (doi:10.1080/01621459.1996.10476956)
210. Gilks WR, Richardson S, Spiegelhalter D. 1995 *Markov chain Monte Carlo in practice*. Boca Raton, FL: CRC press.
211. Gigerenzer G, Goldstein DG. 1996 Reasoning the fast and frugal way: models of bounded rationality. *Psychol. Rev.* **103**, 650. (doi:10.1037/0033-295X.103.4.650)
212. Alvarez P, Squire LR. 1994 Memory consolidation and the medial temporal lobe: a simple network model. *Proc. Natl Acad. Sci. USA* **91**, 7041–7045. (doi:10.1073/pnas.91.15.7041)
213. Hassabis D, Kumaran D, Maguire EA. 2007 Using imagination to understand the neural basis of episodic memory. *J. Neurosci.* **27**, 14 365–14 374. (doi:10.1523/JNEUROSCI.4549-07.2007)
214. Hihi S, Bengio Y. 1995 Hierarchical recurrent neural networks for long-term dependencies. *Adv. Neural Inf. Process. Syst.* **8**.
215. Chung J, Ahn S, Bengio Y. 2016 Hierarchical multiscale recurrent neural networks. (http://arxiv.org/abs/1609.01704)
216. Koutnik J, Greff K, Gomez F, Schmidhuber J. 2014 A clockwork RNN. In *Int. Conf. on Machine Learning*, pp. 1863–1871. PMLR.
217. Mozer MC, Miller D. 1997 Parsing the stream of time: the value of event-based segmentation in a complex real-world control problem. *Int. School on Neural Networks, Initiated by IIASS and EMFCSC*, pp. 370–388.
218. Suddendorf T, Corballis MC. 2007 The evolution of foresight: what is mental time travel, and is it unique to humans? *Behav. Brain Sci.* **30**, 299–313. (doi:10.1017/S0140525X07001975)
219. Ciaramelli E, Grady CL, Moscovitch M. 2008 Top-down and bottom-up attention to memory: a hypothesis (AtoM) on the role of the posterior parietal cortex in memory retrieval. *Neuropsychologia* **46**, 1828–1851. (doi:10.1016/j.neuropsychologia.2008.03.022)
220. Berntsen D, Staugaard SR, Sørensen LMT. 2013 Why am I remembering this now? Predicting the occurrence of involuntary (spontaneous) episodic memories. *J. Exp. Psychol.: General* **142**, 426. (doi:10.1037/a0029128)
221. Dreyfus HL. 1985 From Socrates to expert systems: the limits and dangers of calculative rationality. In *Philosophy and Technology II: Information Technology and Computers in Theory and Practice* (eds C Mitcham, A Huning). Dordrecht, the Netherlands: Reidel.
222. Richmond LL, Zacks JM. 2017 Constructing experience: event models from perception to action. *Trends Cogn. Sci.* **21**, 962–980. (doi:10.1016/j.tics.2017.08.005)
223. Zacks JM, Speer NK, Swallow KM, Braver TS, Reynolds JR. 2007 Event perception: a mind-brain perspective. *Psychol. Bull.* **133**, 273. (doi:10.1037/0033-2909.133.2.273)
224. Radvansky GA, Zacks JM. 2017 Event boundaries in memory and cognition. *Curr. Opin. Behav. Sci.* **17**, 133–140. (doi:10.1016/j.cobeha.2017.08.006)

225. Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA. 2017 Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721. (doi:10.1016/j.neuron.2017.06.041)
226. Stocco A, Lebiere C, Anderson JR. 2010 Conditional routing of information to the cortex: a model of the basal ganglia's role in cognitive coordination. *Psychol. Rev.* **117**, 541. (doi:10.1037/a0019077)
227. Ben-Yakov A, Henson RN. 2018 The hippocampal film editor: sensitivity and specificity to event boundaries in continuous experience. *J. Neurosci.* **38**, 10 057–10 068. (doi:10.1523/JNEUROSCI.0524-18.2018)
228. Bonasia K, Sekeres MJ, Gilboa A, Grady CL, Winocur G, Moscovitch M. 2018 Prior knowledge modulates the neural substrates of encoding and retrieving naturalistic events at short and long delays. *Neurobiol. Learn. Mem.* **153**, 26–39. (doi:10.1016/j.nlm.2018.02.017)
229. Buschman TJ, Miller EK. 2007 Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* **315**, 1860–1862. (doi:10.1126/science.1138071)
230. Beck DM, Kastner S. 2009 Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Res.* **49**, 1154–1165. (doi:10.1016/j.visres.2008.07.012)
231. McMains S, Kastner S. 2011 Interactions of top-down and bottom-up mechanisms in human visual cortex. *J. Neurosci.* **31**, 587–597. (doi:10.1523/JNEUROSCI.3766-10.2011)
232. Kinchla RA, Wolfe JM. 1979 The order of visual processing: 'Top-down, ' 'bottom-up, ' or 'middle-out'. *Percept. Psychophys.* **25**, 225–231. (doi:10.3758/BF03202991)
233. Rauss K, Pourtois G. 2013 What is bottom-up and what is top-down in predictive coding?. *Front. Psychol.* **4**, 276. (doi:10.3389/fpsyg.2013.00276)
234. McClelland JL, Rumelhart DE. 1981 An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol. Rev.* **88**, 375. (doi:10.1037/0033-295X.88.5.375)
235. Deleu T, Góis A, Emezue CC, Rankawat M, Lacoste-Julien S, Bauer S, Bengio Y. 2022 Bayesian Structure Learning with Generative Flow Networks. In *The 38th Conf. on Uncertainty in Artificial Intelligence*.
236. Eberhardt F, Glymour C, Scheines R. 2012 On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. (http://arxiv.org/abs/1207.1389)
237. Peters J, Janzing D, Schölkopf B. 2017 *Elements of causal inference - foundations and learning algorithms*. Cambridge, MA, USA: MIT Press.
238. Bottou L. 2019 Learning representations using causal invariance. *ICLR Keynote Talk*.
239. Priol RL, Harikandeh RB, Bengio Y, Lacoste-Julien S. 2020 An analysis of the adaptation speed of causal models. (http://arxiv.org/abs/2005.09136)
240. Dasgupta I *et al.* 2019 Causal reasoning from meta-reinforcement learning. (http://arxiv.org/abs/1901.08162)
241. Gibson JJ. 1977 *The theory of affordances*. Hilldale, MI: Hilldale Educational Publishers.
242. Cisek P. 2007 Cortical mechanisms of action selection: the affordance competition hypothesis. *Phil. Trans. R. Soc. B* **362**, 1585–1599. (doi:10.1098/rstb.2007.2054)
243. Pezzulo G, Cisek P. 2016 Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends Cogn. Sci.* **20**, 414–424. (doi:10.1016/j.tics.2016.03.013)
244. Bengio E, Thomas V, Pineau J, Precup D, Bengio Y. 2017 Independently controllable features. *CoRR* (http://arxiv.org/abs/1703.07718)
245. Thomas V, Pondard J, Bengio E, Sarfati M, Beaudoin P, Meurs MJ, Pineau J, Precup D, Bengio Y. 2017 Independently controllable features. (http://arxiv.org/abs/1708.01289)
246. Sutton R. 2019 The bitter lesson. *Incomplete Ideas (blog)* **13**, 12.