From birth to loss of representations in artificial neural networks

Stecher, Philipp[1]

[1] Eberhard-Karls-University Tübingen, Germany

lncs@springer.com

Abstract. "[In] unguarded moments I do think that everything is concepts", stated Murphy in his popular big book of concepts [1] emphasizing the pivotal role of concepts, or "mental representations", in understanding the mind. Inspired by Murphy's unguarded moments, we propose a framework asserting that *in artificial neural networks (ANNs) everything is representations and mechanisms*. Specifically, based on an interdisciplinary literature review, we propose a framework called the representations' lifecycle. The framework consists of two main contributions: first, we propose a template that characterizes representational change along three dimensions: a compositional dimension, a hierarchical dimension, and a temporal dimension. Second, the latter template allows for the characterization and demarcation of six representation-altering processes: abstract primitives' integration, perceptual primitives' integration, assembly, abstraction, differentiation and deletion. Our framework provides the foundation for a more formal description of representational change in neural networks and thus, contributes to the broader efforts towards more transparent and explainable ANNs.

Keywords: Representational development, artificial neural networks, neuro-representationalism.

1 Introduction

Building upon neuro-representationalism and following Hubbard [2], we premise our framework on the understanding that the notion of 'representation' implies two distinct but connected worlds: 'the represented world' and 'the representing world'. Thereby, a representation "is an element in the representing world and reflects, stands for, or signifies some aspect of the represented world" [2]. In this paper, we focus on artificial representations. An artificial representation (hereinafter referred to interchangeably as 'representation') is an aspect of a represented world that is encoded in a particular kind of representing world – an artificial neural network. The represented world can refer to the domain with which the ANN interacts. Specifically, we define representations as "information carrying entities identified in ANNs" [3]. As such, a representation encompasses artificial neural elements that are activated during the prediction of an aspect presented to the input layer of an ANN. With being encoded in ANNs, we assert that

representations are shaped by the ANN's architecture, objective function, and learning rule [4], as well as by the data presented to the ANN during training.

Representations in ANNs change. Thus, advancements in ANNs are regularly attributed to processes involved in representational change. Hence, often deep learning researchers attribute the networks' successes to processes such as generalization [5, 6], abstraction [7, 8], differentiation [9], or association [10, 11]. These processes, each playing a unique role, collectively contribute to nowadays sophistication of ANNs. Generalization, for example, broadens the applicability of learned representations. Abstraction and differentiation are instrumental to form novel representations through distilling and distinguishing complex inputs, respectively. Association, on the other hand, creates representations by connecting existing representations, and integrating their informational content. Despite recognizing individual contributions, however, a cohesive framework that demarcates these processes and integrates them into a uniform model of representational development in ANNs is notably absent. To fill this gap, we develop a framework, we call the representations' lifecycle, consisting of six processes responsible for representational change in ANNs, namely abstract primitives' integration, perceptual primitives' integration, assembly, abstraction, differentiation and deletion. Taking together these processes reflect a 'complete' set capable to describe the development of representations in ANNs 'from their birth to their loss'.

Understanding representational change in the mind has longstanding tradition [12-14]. While the representations' lifecycle draws from research on representational development discussed in the context of (non-human) animals at conceptual levels [1, 15-20] or at neural levels [21-26], we primarily focus on and aim to describe representational development in ANNs. The process-integrative view taken in this paper is complementary to existing ideas that approach representational development in ANNs emphasizing the value of single processes [7, 8, 10], mathematical principles [9], or provide nonneural accounts of representational development [27, 28]. In addition, with its representation-centric perspective, our approach stands in contrast to other non-representational accounts such as computational phenomenology [29]. Furthermore, in contrast to other contributions [3, 29-32], we do not question the nature of representations or discuss their epistemic usefulness for better understanding the inner dynamics of the mind respectively ANNs. Rather, following the neuro-representational approach [29], we axiomatically posit the existence and usefulness of representations in ANNs. Specifically, as we will elaborate in section 2, we posit that representations are hierarchically-related compositions that change in a computational manner through mechanisms that operate on them [33]. Additionally, other than it has been suggested [34], tying in with other deep learning researchers [35, 36], we assume that it is useful to describe representations as entities that differ in their abstractness.

The paper is organized as follows. In section 2, we first introduce the template used to characterize representational change. Latter template provides the basis to understand the in section 3 presented representations' lifecycle consisting of the six representation-altering processes. We conclude with a conclusion, limitations and propositions for future research.

2 Dimensions of representational change in ANNs

Representations change during training. In the representations' lifecycle we model these changes through a template, called the representation-altering process. The template characterizes representational change in ANNs along three dimensions: the compositional dimension, the hierarchical dimension, and the temporal dimension. First, representations are composed of finer representational constituents [30, 35]. Second, these compositions differ depending on their degree of specificity/generality and are related to each other [30, 35-37]. Third, latter hierarchical compositions change during training. In the subsequent section, we will elaborate on each of these dimensions. The notations and illustrations provided in this section form the basis to understand the representations' lifecycle in section 3.

2.1 Compositional dimension

Representations are composed of smaller units [30, 35]. Conceptualizing representations as compositions is a widely applied strategy. Especially, explainable AI researchers, aiming at explaining the inner workings of ANNs mechanistically, regularly draw from such compositional understanding [38-40]. To model compositions of representations, we use conceptual graphs as illustrated in Fig. 1. For example, in the conceptual graph (A) the structure of a representation DOG including its constituents HEAD, BODY, and TAIL is shown. (B) illustrates how the conceptual graph maps to a picture of DOG, presented to an idealized image recognition ANN. Finally, (C) exemplifies how this structure could be represented in the neural elements within the imagine recognition ANN. As such the structure reflects the activated neural elements involved in a prediction. For example, assuming the task of identifying a DOG in a picture through an image recognition ANN, the composition of DOG is displayed by the activated neural elements, or in other words, by the neural correlates of the ANN causing the prediction of DOG in dependence of data presented to the input layer. The constituents, in turn, reflect excerpts of the totality of activated neural elements involved in the DOG prediction, and as such refer to differentiable subunits of the DOG representation such as HEAD, BODY, or TAIL.

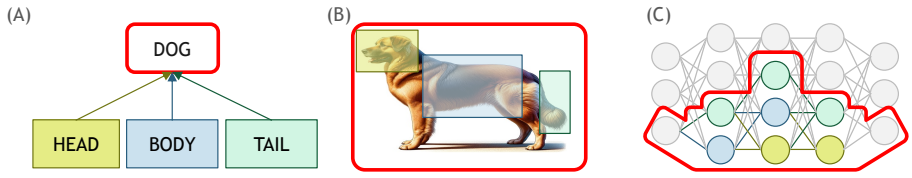

Fig. 1. Conceptual graph and its relation to neural elements in an ANN

2.2 Hierarchical dimension

Representations in ANNs are hierarchically-related [9, 35-37, 41]. Hence, abstract representations form and change in relation to their more perceptual counterparts, and vice

versa [9, 35, 36]. As shown in Fig. 2, following the illustrations provided by Petkov and Petrova [42], we model representational capacities of ANNs' along two axes: composition and hierarchy (A). Particularly, we model the hierarchical relationship of representations in ANNs through (a difference in) shared constituents, whereby, abstract representations share constituents with related perceptual representations but possess less. For example, in (B) the representation CAT is composed of the attributes MEOWS, WARM-BLOODED and NURSING. The CAT representation is hierarchically-related to the superordinate representation, MAMMAL. Both CAT and MAMMAL share attributes like WARM-BLOODED and NURSING. However, CAT has a differentiating attribute such as MEOWS that distinguishes it from MAMMAL and other, subordinate representations associated with MAMMAL. Thus, the abstractness of representations is modelled based on associated constituents, with fewer constituents, or in other words, fewer neural components involved, indicating a higher position in the hierarchy, making it more abstract compared to other representations in the same hierarchy.

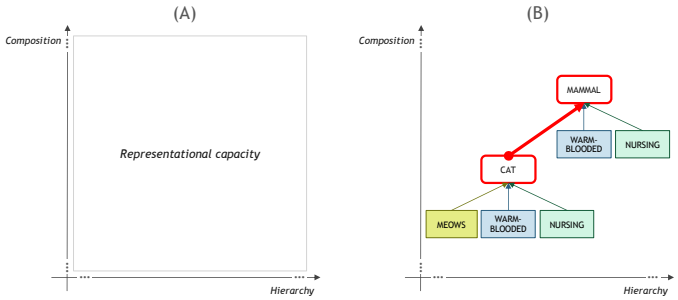

Fig. 2. Excerpt of a representational capacity

2.3 Temporal dimension

As shown in Fig. 3 (A), representation-altering processes describe the changes of representations from state to state or, in other words, "from [an input] structure to [an output] structure" [2] through a transformation mechanism. The transformation mechanism describes how an output structure was reached given an input structure. Specifically, the mechanism provides a causal explanation that demonstrates how and why an input structure produces an output structure given the application of a learning rule responding to (training) data presented to the input layer. As introduced before, the structure of a representation describes its hierarchically organized composition. Hence, ultimately representation-altering processes describe how hierarchically-related constituents of representations are modified, added, or deleted as a result of a mechanism. In (B) the input and output structures of the representation-altering process differentiation are illustrated. Differentiation is characterized by forming more specific representations based on more abstract representations given in the input structure. Accordingly, as illustrated in Fig. 3. (B), the DOG representation was formed integrating constituents

of its hierarchical parent AGENT with additional, differentiating constituents such as BARKS.

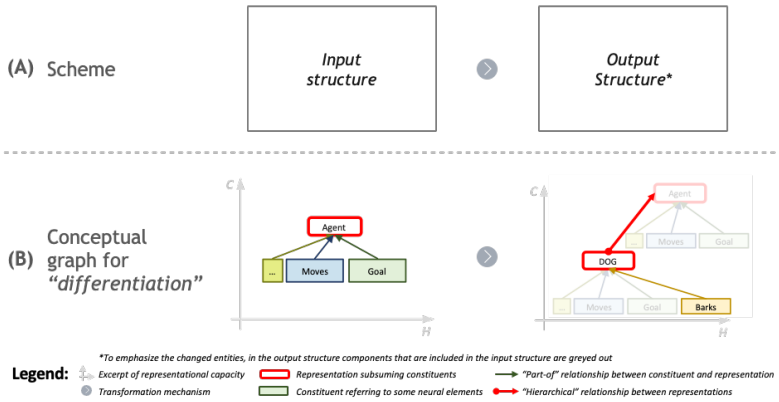

Fig. 3. Representation-altering process „differentiation"

3 From birth to loss: The representations' lifecycle

The representations' lifecycle is composed of six representation-altering processes that are being organized along three phases: innate, form & change, and decay. In the innate phase, the ANN is in a pre-training state, not yet influenced by external data. Processes in this phase concern the integration of primitives into an ANN through developers. Once the incorporation of data starts, the ANN enters the form & change phase. From this point, the system forms new representations or changes existing ones by doing both, incorporating training data, and leveraging already integrated representations. Finally, representations are being deleted. The latter refers to the removal of representations from the ANN's representational capacity.

In the following the representation-altering processes for each of three phases will be presented. Fig. 4 introduces illustrative examples of the representation-altering processes. shows a summary of characteristics of the representation-altering processes as described for (non-human) animals and ANNs, respectively. The representation-altering processes introduced in the sections 3.1 to 3.3 are about descriptions of the processes in terms of input and output structures. Each process presentation starts with a definition, followed by a characterization of the process derived from the cognitive science literature. Hereafter, analogous findings identified in AI literature are summarized.

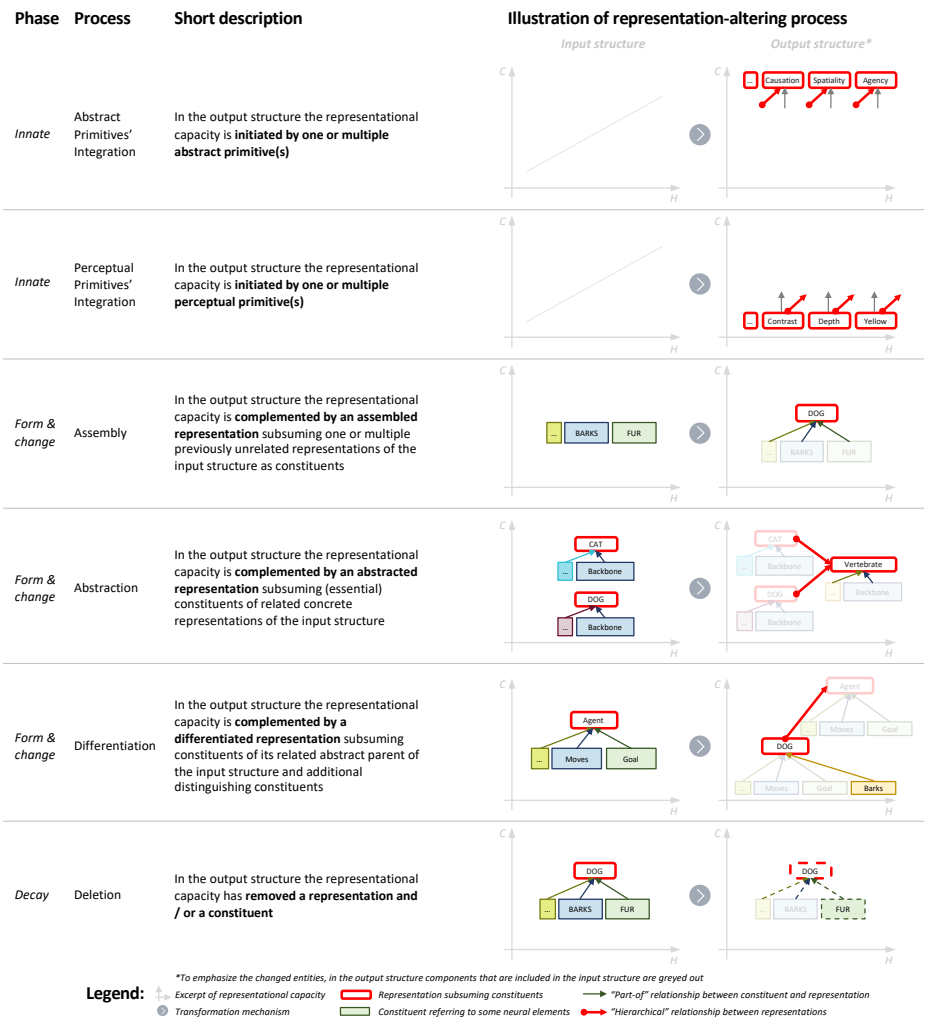

Fig. 4. Overview of representation-altering processes

3.1 Innate

The debate of how much of the mind is innate, and how much is formed by experience is stretching far back in history. Back then, primarily two schools of thought opposed each other [18]: nativism and empiricism. Proponents of the former argued that, to explain the complex behavior of biological cognitive systems right after birth, an innate machinery is required [18]. On the other hand, empiricists have argued that biological cognitive systems learn everything by observation respectively through their experiences, and possess little to no innate representational capacity [43, 44]. Nowadays, it seems to be widely acknowledged [43], however, that "genes and experience" work

together. This interplay stance accepts that both components are important to model cognitive biological systems: innate machinery as well as experience [43]. Thereby, the innate machinery enables the system to learn, while learning, in the terminology of this paper, refers to the process of acquiring novel representations based on sensory data and in the system existing representations.

In AI, the above controverses seem to be mirrored. However, while cognitive scientists mostly aim to discover whether (non-human) animals have innated representational capacities, the AI debate rather revolves around whether and what innate representational structures are useful to develop more capable systems. Thus, empiricists argue that ANNs require little to no primitives to work effectively [9, 45]. On the other hand, scholars leaning toward the nativist viewpoint [18, 35, 46] argue that leveraging innate structures could help building more successful machines. The interplay stance proposes that ANNs should be modelled as a function over innate structures and experience, or in the context of AI, acquired training data [43]. Integrating the nativists and empiricists viewpoint, the stance emphasizes first, that innate structures and training data are relevant to describe the inner workings of ANNs and second, that ANNs differ in their richness of innate structures, but that their innate structure "cannot literally be zero" [43]. Following this stance, in this paper we assume that every ANN starts with an innate set of representational structures which in later stages of development is enriched by additional structures through training. Specifically, we conceptualize the initial representational capacity of an ANN, emerging from its initial architecture, objective function, and learning rule, as a set of primitives, or in other words, a set of innate representations, that both enables and restricts what the ANN can learn over the course of training.

Leaning toward the "nativist" side of the interplay stance, cognitive and computer scientists have tried to identify specific primitives that allow to explain the learning capabilities of biological cognitive systems, or that are useful to design capable AI, respectively. In the following, we systematize these primitives along two angles assuming that primitives can occur in abstract [16, 19, 20, 25, 43, 47, 48] and perceptual varieties [15, 16, 49]. Accordingly, it is assumed that the "life" of some abstract and perceptual representations starts with their integration into the ANN before the network is trained.

Abstract primitives' integration. Abstract primitives' integration refers to the insertion of abstract representations into an ANN, before training. As illustrated in Fig. 4, in the output structure abstract primitives are in the upper right section of a representational capacity. As such, they are considered the most abstract representations and thus, containing the fewest constituents compared to related perceptual representations. The input structure is empty since an ANN's representational capacity comes into existence with having them integrated. Abstract primitives contain constituents which, at later stages of representational development, are inherited by the more perceptual representations formed on their basis.

In (non-human) animals, abstract primitives are often characterized as innate representations on which basis more perceptual representations are formed [15, 16, 47, 49]. A primitive is considered as abstract if, unlike a perceptual one, it allows for subsuming

a wide range of different aspects of the represented world [47, 49] and if it contains rather functional constituents than perceivable ones [15, 49, 50]. According to Mandler [15], contrarily to perceptual primitives, abstract ones include knowledge that is beyond the information provided by the senses and rather describe "what [an aspect] is" instead of "what [an aspect] looks like". For humans, various kinds of abstract primitives have been proposed such as AGENCY [16, 19, 20, 47], CAUSATION [16, 19, 47], NUMERICS [16, 20], SOCIALITY [19, 20] and SPATIALITY [16, 47, 48]. For latter stages of representational development, it is assumed that the initial set of abstract primitives is expanded through top-down processes (see section 3.2) [15, 16, 47, 49]. For example, pointing to experiments with infants, Mandler [15, 47] concludes that "many early [representations] appear to be global, relatively crude, and lacking in detail" [15] and, at latter stages, are differentiated through leveraging perceptual primitives and sensory data [47]. The products of the differentiation are newly formed, more perceptual representations added to the representational capacity.

In AI, abstract primitives increasingly move to the center of current research efforts [7, 8, 43]. In this vein, Mitchell [8] emphasized their importance: "unless we create AI systems that can master [abstract primitives] we have little hope of creating anything like human-level AI". Although considered worthy endeavors [8, 18, 35, 43], efforts to equip modern AI with "[rich] priors, that orientate learning and improve acquisition speed" [18], or with "general priors about the world around us, i.e., priors that are not task-specific but […] useful" [35], seem to remain in an early stage of development [8]. Nonetheless, scholars already suggested abstract primitives that may be worth considering for AI. While pointing to the work of developmentalists, Marcus [43] proposes a set of ten primitives such as SPATIOTEMPORAL CONTIGUITY, CAPACITY FOR COST-BENEFIT ANALYSIS or CAUSALITY. Bengio, Courville [35] proposed general primitives "about the world around us" such as A HIERACHICAL ORGANIZATION OF EXPLANATORY FACTORS, SMOOTHNESS and MANIFOLDS that enable "the learner to discover and disentangle some of the underlying (and a priori unknown) factors of variation that the data may reveal" and which are partially artificially implemented already [35]. However, it remains unknown "how long the list [of abstract primitives] really ought to be" [43]. Generally, identifying a comprehensive list of abstract primitives relevant to AI is significantly complicated by the lack of understanding of how useful abstract representational structures can be artificially generated [7, 8, 51].

Perceptual primitives' integration. Perceptual primitives' integration refers to the insertion of perceptual representations into an ANN, before training. As illustrated in Fig. 4, in the output structure perceptual primitives are in the lower left section of a representational capacity. The input structure is empty since an ANNs' representational capacity comes into existence with having them integrated. Perceptual primitives encompass constituents which, at later stages of representational development, are subsumed by assembled representations formed on their basis.

In (non-human) animals, perceptual primitives are often characterized by encompassing sensory, or sensorimotor representations [15, 16, 52] and are likely to be bounded to the senses respectively likely to be modality-specific [53]. For example,

visual perceptual primitives allow the interpretation of accumulated bits of sensory data [16] and can result in perceptions such as EDGES, CONTRAST or DEPTH. Visual perceptual primitives are used for the interpretation of combinations of wavelengths of light that stimulated the visual detectors of a given cognitive system. The "most basic" primitives are likely to have in common that they allow a cognitive system to detect the presence of elements of pattern in sensory data [52]. At a neurological level, perceptual primitives are supposed to be represented as "hard-wired" neuronal networks that allow biological cognitive systems to recognize complex stimuli [53]. These kinds of neuronal networks have been reported across species (e.g., humans [54, 55], reptiles [53], insects [56]) and across different modalities (e.g., lexical [54, 55], visual [55], auditive [54, 56]). The perceptual primitives complement the innate set of abstract primitives and are supposed to be involved in the enrichment of the representational capacity through bottom-up processes (see section 3.2). It is assumed that during the formation of representations, after the system came to life, both, abstract and perceptual primitives are involved simultaneously [15, 50]. For example, while humans learn new representations such as DOG or CAT, they differentiate their abstract primitives such as AGENCY and SPATIALITY and draw from perceptual primitives involved during the pre-interpretation of streams of sensory data which finally lead to an assembly of perceptions such as BROWN, FLUFFY, or BARKS as constituents, in this case, subsumed under the DOG representation.

In AI, artificial analogies that most closely match the scheme of perceptual primitives are representations that help to pre-process incoming raw data, i.e., the data collected by the virtual (e.g., program/data interfaces) or physical sensors (e.g., cameras, microphones) connected to a given ANN. Pre-processing of incoming raw data is supposed to facilitate the learning from or classification of data in AI [57, 58]. Artificial perceptual primitives are domain-specific and thus, help to process data that is coined by a specific modality (e.g., visual/pixels, auditive/tones) and often is collected within a specific application area (e.g., traffic, weather forecast). Same as their biological counterparts, they enable AI to interpret chunks of sensory data. For example, artificial feature detectors are dedicated to enable better predictions through the pre-interpretation of chunks of *visual* sensory data [57]. For the interpretation of images, feature detection methodologies are used that can detect shape-entities, such as edges, contours, corners, or blobs [57] based on e.g., trained representations of those shapes available in ANNs. The detected features can then be combined to assemble new representations that associate the interpreted shape features [57, 59]. We assume that both, abstract and perceptual primitives, are involved when novel representations are formed in ANNs. For example, while an image recognition ANN learns the visual representation of a DOG, it differentiates its abstract primitives and leverages its perceptual primitives, involved in the pre-processing of pixels delivered by the virtual or physical sensors, leading to detected edges, or corners. Latter edges, or corners, are then assembled as constituents of the visual representation of DOG.

3.2 Form and change

During training an ANN forms new representations or changes existing ones by incorporating data and combining already integrated representations. In humans, the processes that form and change representations underpin higher-level cognitive abilities like decision-making, reasoning, language comprehension, and planning [24, 60-62]. In ANNs, given proper training data and effective execution, these processes provide the system with novel and/or refined representations resulting in enhanced predictions. As suggested by Gibson and Gibson [63], there are two theoretical stances to explain the formation of representations: the enrichment theory and the specificity theory. Recasting the two stances in the terminology of this paper, the enrichment theory assumes that an ANN's representational capacity is starting with a set of perceptual primitives and is then, through bottom-up processes, gradually expanded by further abstract representations; on the other hand, the specificity theory argues that a representational capacity starts with abstract primitives on which basis, through top-down processes, more perceptual representations are derived [63]. As emphasized in literature [15, 16, 49] and as described above, both types of processes are mutually dependent and deeply intertwined. Thus, we argue that starting with innate structures a representational capacity can be expanded through bottom-up processes such as assembly and abstraction as well as through top-down processes such as differentiation. Bottom-up processes either assemble representations through associating one or multiple representations, or produce abstract representations based on input structures which are composed of more perceptual representations. On the other hand, top-down processes either produce perceptual representations from abstract ones or specify existing representations.

Assembly. Assembly refers to the process of forming representations along the "Composition" dimension through the association of two or more representations/constituents. As illustrated in Fig. 4, in the output structure the newly assembled representation contains constituents that in the input structure were unrelated. Assembly can take as input representations of any kind, modality, or abstraction and can produce representations of any kind, modality, or abstraction.

In (non-human) animals, assembled representations are often characterized by containing relationships between "separate (i.e., formerly unrelated) […] representations" [64]. In humans, these assembled representations can be multimodal and therefore, can consist of representations from different sensory modalities [65, 66]. Additionally, assembly is characterized as a process that produces representations which are supposed to contain information about the propositional and semantic quality of the relationship in addition to the information that representations are related [24, 67]. The example given in Fig. 4 illustrates how representations of a given input structure derived from sensorimotor and auditive data, i.e., FLUFFY and BARKS, get subsumed as constituents under the newly assembled representation of DOG in the output structure. In this case, the relationship "part of" may describe the shared semantic content of the two associated representations, i.e., FLUFFY and BARKS, with the assembled representation DOG. The latter type of assembly refers to the formation of simple associations,

whereby objects or sensory impressions are classified as associatively related if they tend to spatially or temporally co-appear [24]. Besides the here given entity representation, it is assumed that assembly can result in relational representations that are primarily defined by their relations outside themselves. Consequently, assembly can result in representations such as HUNTING by associating multiple representations such as PREDATOR and BAIT [42, 68] through, for example, extracting the common essences of event representations in which the act of one animal chasing another was observed. Furthermore, assembly also includes the formation of more abstract relations between representations, which are not formed based on co-occurrences but are, nonetheless, "judged as related, or conveying a common concept" [24]. In this sense, it is assumed that the process of assembly can result in all kinds of complex representations such as schema representations. For example, assembly can result in representations that reflect temporal/causal events (e.g., AFTER PUSHING THE BUTTON THE TV TURNS ON) or functional relationships (e.g., A SPOON IS USED TO EAT SOUP) [24]. In summary, the term assembly used in this paper refers to a process in which pre-existing representations of any kind, modality, and abstraction are subsumed as constituents under a newly formed representation of any kind, modality, and abstraction along the "Composition"-dimension, while at the same time propositional respectively semantic content about their association is integrated.

In AI, assembly-like processes are often entitled to be essential to the operation of modern ANNs [11, 13, 36]. Indeed, just like the human variant of simple associations, modern AI algorithms recognize patterns in data by analyzing statistical co-occurrences. For example, Wang and Raj [13] argue that machine learning methods in general cluster "samples that are near to each other (under a defined distance) […] in one group" and that they draw more attention to "explanatory variables that frequently occur with response variables". Associations have been created both within and between artificial representations of different modalities [11]. For example, deep learning computer vision algorithms [69-71] have established relationships between visual representations by analyzing their spatial and temporal co-occurrences, resulting in newly assembled modal-specific representations [11, 59]. Furthermore, modern deep learning algorithms have successfully assembled multimodal representations (for comprehensive review, see Guo, Wang [11]), whereby, e.g., representations with lexical and visual formats were combined. As reported by Guo, Wang [11], the resulting representations differed in efficacity depending on the modalities that are combined. For instance, while effective representations combining image and language modalities have been successfully assembled and applied, other representations such as, for example, the ones combining audio and video modalities are in a comparably early stage of development [11]. Furthermore, whereas first artificial associations between representations of different modalities have been established, the question of how to effectively assemble representations which integrate propositional content such as causality remains "largely open" [51].

Abstraction. Abstraction refers to the formation of representations along the "Hierarchy" dimension. As illustrated in Fig. 4, in the output structure the newly abstracted representation is characterized by a reduction in specificity and an expansion in scope.

The reduction in specificity refers to a decrease in constituents of the more abstract representation relative to the related perceptual representation(s). The expansion in scope refers to an increase of subsumed perceptual representations that participate in constituents of the abstracted representations. Abstracted representations include essences distilled from more perceptual representations contained in the input structure.

In (non-human) animals, abstraction is often associated with a reduction in specificity and an expansion of scope [72] of the newly formed representations. The reduction of specificity is characterized by a decreasing number of associated constituents of the newly formed abstract representation in comparison to the original more perceptual representation(s) [72, 73]. The set of reduced constituents of an abstract representation encompasses a set of "invariant central characteristics" [74] of an aspect and thereby, represents "any properties that increase the likelihood of accurately identifying [the aspect] across various contexts". For example, in Fig. 4 the representations CAT and DOG have one or more common constituents such as BACKBONE. In the output structure, this set of common constituents is used to characterize the newly formed, more abstract representation VERTEBRATE. Latter set of constituents describing the abstract representation VERTEBRATE encompasses less associated constituents than CAT or DOG and therefore is less specific. The expansion of scope refers to an increase in the number of subordinate representations associated with the newly formed abstract representation compared to the perceptual representation(s) of the input structure [72, 73]. For example, the output representation VERTEBRATE is not characterized by constituents such as BARKS, or MEAT-EATER which, however, can be attributed to its subordinate DOG. In addition, VERTEBRATE subsumes representations of DOG and CAT and their underlying subordinates (e.g., SHEPHERD and POODLE) as well as other representations that are characterized by having a BACKBONE such as HUMAN or AMPHIBIAN. VERTEBRATE is associated with an expanded scope of subordinate representations and consequently, covers a wider range of aspects. Finally, an abstract representation can encompass constituents from different modalities. Accordingly, it has been argued [63, 75] that abstract representation likely involve perceptual and functional constituents; but in contrast to perceptual ones, abstract representations tend to have more functional constituents attached and are considered to be less associated with sensory impressions.

In AI, a popular explanation for why artificial networks tend to work so well is that they construct "more complex […] representations from simpler and less abstract ones" [76]. In this vein, LeCun, Bengio [36] pointed out that modern deep learning networks operate on the basis of abstraction, whereby higher-level layers of a network contain abstractions formed on the basis of representations from lower-level layers. Thereby, representations in lower-level layers are more likely to be changed by local variations of the input data. Those of higher-level layers are "generally invariant" [35] to most variations of the input. Furthermore, representations contained in higher-level layers re-use representations that were learned by lower-levels, which makes them particularly suited for learning across domains [37] indicating their capability of subsuming wider ranges of aspects. While abstract representations have been artificially produced, there remain major differences to the abstractions that humans can produce and apply. Indeed, unlike humans, who effectively can apply their abstractions to novel situations,

modern AI systems require vast amounts of data to create less generalizable representations that tend to produce mediocre results when applied to new situations/domains [7, 77]. As emphasized by Shanahan and Mitchell [7], although modern AI systems are able to "achieve a certain degree of generalization", the abstract representations formed by modern AI often remain "tied to the domain in which they were acquired". The authors Shanahan and Mitchell [7] conclude that the "shortcomings of contemporary neural network methods such as low sample efficiency, limited transfer ability, and poor out of distribution generalization […] result from an inability to form sufficiently general abstractions". In general, the quest to understand how abstract representations are formed and what qualities they ought to possess to be effectively applied is ongoing [8, 75].

Differentiation. Same as abstraction, differentiation refers to the formation of representations along the "Hierarchy" dimension. Contrarily to abstraction, differentiation results in perceptual representations with an increased specificity and a reduced scope. As illustrated in Fig. 4, the increase in specificity refers to an increase in constituents of the more differentiated representation relative to the related abstract representation; the additional constituents among others differentiate the newly formed representations from other representations participating in the same hierarchy. On the other hand, the decrease in scope refers to a decrease of subsumed perceptual representations relative to its associated more abstract parent.

In (non-human) animals, output structures resulting from differentiation are often characterized by an increased specificity and a reduced scope in comparison to the input structure. The increase in specificity refers to the differentiated representation containing more constituents than its parent [47, 63, 78]. These additional constituents distinguish the resulting representations from other representations related to the parent [78]. For example, as shown in Fig. 4, based on a given representation AGENT the more specified representation DOG is derived which encompasses besides the constituents which were inherited from the superordinate representation (GOAL and ACTS) additional constituents such as BARKS. Thereby, the additional constituents distinguish DOG from other representations that were also considered as AGENT such as CAT. The DOG representation represents a smaller range of aspects and therefore, possesses a reduced scope. Latter allows a cognitive system to separate stimuli that were once indistinguishable [79]. Furthermore, contrarily to their more abstract parents, it is assumed that differentiated representations are likely to have more perceptual and less functional constituents attached than their parent [63].

In AI, during learning, abstract primitives and representations get adjusted or refined to fit the presented data. The process of fitting the rather general primitives and representations to data suggests differentiation-like processes [9]. Hereby, the general-purpose primitives serve as a starting point from which a more specific purpose is derived. For example, in their nonlinear ANN, Saxe, McClelland [9] observed "hierarchical, progressive differentiation of structure in its internal hidden representations, in which animals vs. plants are first distinguished, then birds vs. fish and trees vs. flower, and, finally, individual items". Moreover, they claim that "progressive differentiation of hierarchical structure […] is an inevitable consequence of deep-learning dynamics" [9].

Broadly, to accelerate the AI's learning process, AI researchers regularly aim to apply the weights of higher layers (containing more abstract information) collected in a particular domain to other domains. However, as mentioned above, the transfer of gathered abstract representations attained in a specific domain remain often tied to the domain in which they were acquired [7].

3.3 Deletion

Deletion results in the removal of representations and/or constituents from the representational capacity. Deletion is more than a mere biproduct of representation altering. Contrarily, it has been suggested that deletion provides cognitive systems with several benefits [80-83] such as efficient storage management, attention direction, and better decision making. Both, the formation and forgetting of representations, are considered as "complementary processes which construct and maintain useful representations" [84]. The decay of representations can be a product of deleting, overriding, suppressing, or sorting out outdated information [82].

In AI, within deep ANNs learned representations change depending on the data presented to the input layer; this change, in turn, can result in a decay of representations or constituents. For example, considering the case of adapting a pretrained network to a new domain: if trained with data from the new domain, the network's representations change to reflect the contingencies of the new domain. If the new domain's data is significantly different from the original data, the network can "forget" [85]. When the AI is then confronted with input data similar to that of the original domain, it often cannot successfully classify the data because the corresponding representations have been overwritten [86]. While this may be acceptable for ANNs that only works and will work in the new domain, the performance of a network that is expected to switch between the domains suffers significantly. To encounter this issue of co-called "catastrophic forgetting" [85, 86] various strategies have been proposed [85-87] which have in common that they aim to segment the representations in a network and thereby, prioritize the learned representations and constituents that were useful in the original domain. Through this segmentation methods, constituents of representations become more disentangled which, in turn, allows to forget them more systematically respectively allows to forget unrequired parts independent from the required ones.

4 Conclusions, limitations and further research avenues

This paper has introduced a framework to describe the lifecycle of representations in artificial neural networks, from the representations' birth to their loss. We presented a template to characterize representational change across three dimensions—compositional, hierarchical, and temporal—and detailed the processes of abstract primitives' integration, perceptual primitives' integration, abstraction, differentiation, assembly, and deletion. The framework offers a foundation for understanding representational dynamics in ANNs and potentially can serve as a basis to computationally describe the development of representations in ANNs.

However, several limitations exist. One of the implicit assumptions of our framework is that there is a prototypical form of information that can be consistently represented across different ANN architectures. This assumption presents significant challenges. Representational structures may vary widely between architectures due to differences in learning mechanisms, layer configurations, or data inputs. These variations cast doubt on the universality of the processes we outline, as what constitutes a "representation" may differ significantly across networks. Future research should critically investigate whether such a prototypical representation exists or if architecture-specific models are required to better capture the distinct characteristics of different systems.

The current work focuses primarily on the input and output structures of representational transformations, without describing the underlying *transformation mechanisms* that drive these transitions. According to the definition provided in this paper, transformation mechanisms describe the causal operations that explain how an input structure leads to a particular output structure. However, this paper's scope was limited to modelling structural changes in representations, leaving the mechanisms themselves largely unexplored.

The completeness of the representations' lifecycle framework, as illustrated in Fig. 5, derives from its ability to encompass all necessary processes that govern the evolution of representational states across compositional and hierarchical dimensions. The graph explicitly shows how movements along these axes—whether through abstraction, differentiation, assembly, or deletion—can capture any possible state of a representational capacity. The framework assumes that any representational state can be reached starting from an initial set of primitives by applying these processes. While this is merely theoretical completeness, future work must investigate whether it holds across varying neural architectures, where differing mechanisms may affect the actual pathway of representational change.

Finally, formalizing the lifecycle framework is an essential next step. While the current work provides a conceptual and graphical representation of the processes, a mathematical model would enhance precision and allow for empirical validation. Moreover, future studies should map the processes outlined—such as abstraction, differentiation, and assembly—onto specific neural activities observed in ANNs. This would enable a deeper understanding of representational dynamics and offer more universally applicable insights across architectures.

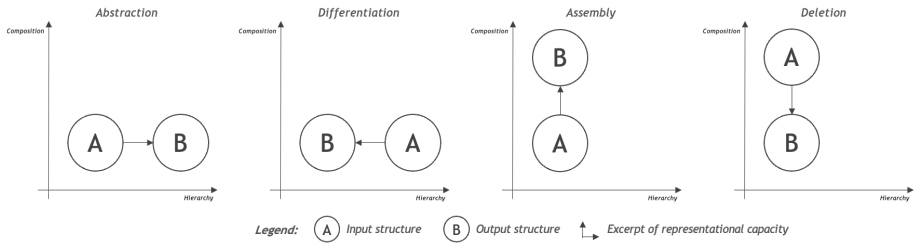

Fig. 5. Minimal set of processes describing representational development "completely" along the compositional and hierarchical dimensions

References

1. Murphy, G.L., *The big book of concepts*. The big book of concepts. 2002, Cambridge, MA, US: Boston Review. 555-555.
2. Hubbard, T.L., *What is mental representation? And how does it relate to consciousness?* Journal of Consciousness Studies, 2007. **14**(1): p. 37-61.
3. Anderson, M.L. and H. Champion, *Some dilemmas for an account of neural representation: A reply to Poldrack.* Synthese, 2022. **200**(2).
4. Richards, B.A., et al., *A deep learning framework for neuroscience.* Nat Neurosci, 2019. **22**(11): p. 1761-1770.
5. Bengio, Y., L. Kaelbling, and K. Kawaguchi, *Generalization in Deep Learning*, in *Mathematical Aspects of Deep Learning*, P. Grohs and G. Kutyniok, Editors. 2022, Cambridge University Press: Cambridge. p. 112-148.
6. Zhang, C., et al., *Understanding deep learning (still) requires rethinking generalization.* Communications of the ACM, 2021. **64**(3): p. 107-115.
7. Shanahan, M. and M. Mitchell, *Abstraction for deep reinforcement learning*, in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. 2022: Vienna, Austria.
8. Mitchell, M., *Abstraction and analogy-making in artificial intelligence.* Ann N Y Acad Sci, 2021. **1505**(1): p. 79-101.
9. Saxe, A.M., J.L. McClelland, and S. Ganguli, *A mathematical theory of semantic development in deep neural networks.* Proceedings of the National Academy of Sciences, 2019. **116**(23): p. 11537-11546.
10. Meng, K., et al., *Locating and Editing Factual Associations in GPT*, in *Advances in Neural Information Processing Systems*, M.I. Jordan, Y. LeCun, and S.A. Solla, Editors. 2022, The MIT Press: New Orleans, USA.
11. Guo, W., J. Wang, and S. Wang, *Deep Multimodal Representation Learning: A Survey.* IEEE Access, 2019. **7**: p. 63373-63394.
12. Nimtz, C. and J. Langkau, *Concepts in philosophy - a rough geography.* Grazer Philosophische Studien, 2010. **81**: p. 1-11.
13. Wang, H. and B. Raj, *On the Origin of Deep Learning.* 2017.
14. Margolis, E. and S. Laurence. *Concepts*. Fall 2023 Edition 2019; Available from: https://plato.stanford.edu/archives/fall2023/entries/concepts/.
15. Mandler, J.M., *Perceptual and Conceptual Processes in Infancy.* Journal of Cognition and Development, 2000. **1**(1): p. 3-36.
16. Carey, S., *Precis of 'The Origin of Concepts'.* Behav Brain Sci, 2011. **34**(3): p. 113-24; discussion 124-62.
17. Sloutsky, V.M. and W. Sophia Deng, *Categories, Concepts, and Conceptual Development.* Lang Cogn Neurosci, 2019. **34**(10): p. 1284-1297.
18. Versace, E., et al., *Priors in Animal and Artificial Intelligence: Where Does Learning Begin?* Trends Cogn Sci, 2018. **22**(11): p. 963-965.
19. Liu, S., N.B. Brooks, and E.S. Spelke, *Origins of the concepts cause, cost, and goal in prereaching infants.* Proc Natl Acad Sci U S A, 2019. **116**(36): p. 17747-17752.
20. Spelke, E.S. and K.D. Kinzler, *Core knowledge.* Dev Sci, 2007. **10**(1): p. 89-96.
21. Martin, A., *The representation of object concepts in the brain.* Annu Rev Psychol, 2007. **58**: p. 25-45.
22. Ebitz, R.B. and B.Y. Hayden, *The population doctrine in cognitive neuroscience.* Neuron, 2021. **109**(19): p. 3055-3068.
23. Chung, S. and L.F. Abbott, *Neural population geometry: An approach for understanding biological and artificial neural networks.* Curr Opin Neurobiol, 2021. **70**: p. 137-144.

24. Zucker, L. and L. Mudrik, *Understanding associative vs. abstract pictorial relations: An ERP study.* Neuropsychologia, 2019. **133**: p. 107127.
25. Kiefer, M. and F. Pulvermüller, *Conceptual representations in mind and brain: theoretical developments, current evidence and future directions.* Cortex, 2012. **48**(7): p. 805-25.
26. Pulvermüller, F., *How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics.* Trends Cogn Sci, 2013. **17**(9): p. 458-70.
27. Parthemore, J., *The Unified Conceptual Space Theory: an enactive theory of concepts.* Adaptive Behavior, 2013. **21**(3): p. 168-177.
28. Gärdenfors, P., *Conceptual spaces: The geometry of thought*. Conceptual spaces: The geometry of thought. 2000, Cambridge, MA, US: The MIT Press. x, 307-x, 307.
29. Beckmann, P., G. Köstner, and I. Hipólito, *An Alternative to Cognitivism: Computational Phenomenology for Deep Learning.* Minds and Machines, 2023. **33**(3): p. 397-427.
30. Poldrack, R.A., *The physics of representation.* Synthese, 2020. **199**(1-2): p. 1307-1325.
31. Thomson, E. and G. Piccinini, *Neural Representations Observed.* Minds and Machines, 2018. **28**(1): p. 191-235.
32. Morris, M., *Why there are no mental representations.* Minds and Machines, 1991. **1**: p. 1-30.
33. Thagard, P. *Cognitive science*. Winter 2023 Edition 2023; Available from: https://plato.stanford.edu/archives/win2023/entries/cognitive-science/.
34. Barsalou, L.W., L. Dutriaux, and C. Scheepers, *Moving beyond the distinction between concrete and abstract concepts.* Philos Trans R Soc Lond B Biol Sci, 2018. **373**(1752).
35. Bengio, Y., A. Courville, and P. Vincent, *Representation Learning: A Review and New Perspectives.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013. **35**(8): p. 1798-1828.
36. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.
37. Bengio, Y. and O. Delalleau, *On the Expressive Power of Deep Architectures*, in *Algorithmic Learning Theory*. 2011. p. 18-36.
38. Kästner, L. and B. Crook, *Explaining AI through mechanistic interpretability*. 2023.
39. Nanda, N., et al., *Progress measures for grokking via mechanistic interpretability*. 2023.
40. Olah, C., *Mechanistic interpretability, variables, and the importance of interpretable bases*. 2022: Transformer Circuits Thread.
41. Matsuo, Y., et al., *Deep learning, reinforcement learning, and world models.* Neural Netw, 2022. **152**: p. 267-275.
42. Petkov, G. and Y. Petrova, *Relation-Based Categorization and Category Learning as a Result From Structural Alignment. The RoleMap Model.* Front Psychol, 2019. **10**: p. 563.
43. Marcus, G., *Innateness, AlphaZero, and Artificial Intelligence*. 2018.
44. Locke, J., *An essay concerning human understanding*. Vol. 3. 1689: Oxford University Press. 601-605.
45. Silver, D., et al., *Mastering the game of Go without human knowledge.* Nature, 2017. **550**(7676): p. 354-359.
46. Barabasi, D.L., et al., *Complex computation from developmental priors.* Nat Commun, 2023. **14**(1): p. 2226.
47. Mandler, J.M., *On the Birth and Growth of Concepts.* Philosophical Psychology, 2008. **21**(2): p. 207-230.

48. Mandler, J.M., *The spatial foundations of the conceptual system.* Language and Cognition, 2014. **2**(1): p. 21-44.
49. Carey, S., *The Origin of Concepts.* Journal of Cognition and Development, 2000. **1**(1): p. 37-41.
50. Chalmers, D.J., R.M. French, and D.R. Hofstadter, *High-level perception, representation, and analogy: A critique of artificial intelligence methodology.* Journal of Experimental & Theoretical Artificial Intelligence, 1992. **4**(3): p. 185-211.
51. Scholkopf, B., et al., *Toward Causal Representation Learning.* Proceedings of the IEEE, 2021. **109**(5): p. 612-634.
52. Aslin, R.N. and L.B. Smith, *Perceptual development.* Annu Rev Psychol, 1988. **39**: p. 435-73.
53. Martin, K.A., *A brief history of the "feature detector".* Cereb Cortex, 1994. **4**(1): p. 1-7.
54. Eimas, P.D. and J.D. Corbit, *Selective Adaptation of Linguistic Feature Detectors.* Cognitive Psychology, 1973. **4**: p. 99-109.
55. Pelli, D.G., et al., *Feature detection and letter identification.* Vision Res, 2006. **46**(28): p. 4646-74.
56. Ronacher, B., *Innate releasing mechanisms and fixed action patterns: basic ethological concepts as drivers for neuroethological studies on acoustic communication in Orthoptera.* J Comp Physiol A Neuroethol Sens Neural Behav Physiol, 2019. **205**(1): p. 33-50.
57. Li, Y., et al., *A survey of recent advances in visual feature detection.* Neurocomputing, 2015. **149**: p. 736-751.
58. Yu, L. and H. Liu, *Efficient Feature Selection via Analysis of Relevance and Redundancy.* Journal of Machine Learning Research, 2004. **5**: p. 1205–1224.
59. Higgins, I., et al., *SCAN: Learning hierachical compositional visual concepts*, in *International Conference on Learning Representations*. 2018: Vancouver, Canada.
60. Wasserman, E.A. and R.R. Miller, *What's elementary about associative learning?* Annu. Rev. Psychology, 1997. **48**: p. 573-607.
61. Solomon, K., D. Medin, and E. Lynch, *Concepts do more than categorize.* Trends Cogn Sci, 1999. **3**(3): p. 99-105.
62. Welling, H., *Four Mental Operations in Creative Cognition: The Importance of Abstraction.* Creativity Research Journal - CREATIVITY RES J, 2007. **19**.
63. Gibson, J. and E. Gibson, *Perceptual learning: differentiation or enrichment?* Psychological Review, 1955. **62**(1): p. 32-41.
64. Caviezel, M.P., et al., *The Neural Mechanisms of Associative Memory Revisited: fMRI Evidence from Implicit Contingency Learning.* Front Psychiatry, 2019. **10**: p. 1002.
65. Kiefer, M. and L.W. Barsalou, *Grounding the Human Conceptual System in Perception, Action, and Internal States*, in *Action Science*. 2013. p. 381-407.
66. Barsalou, L.W., *Grounded cognition.* Annu Rev Psychol, 2008. **59**: p. 617-45.
67. Mitchell, C.J., J. De Houwer, and P.F. Lovibond, *The propositional nature of human associative learning.* Behav Brain Sci, 2009. **32**(2): p. 183-98; discussion 198-246.
68. Asmuth, J. and D. Gentner, *Relational categories are more mutable than entity categories.* Q J Exp Psychol (Hove), 2017. **70**(10): p. 2007-2025.
69. Ullman, S., et al., *Atoms of recognition in human and computer vision.* Proc Natl Acad Sci U S A, 2016. **113**(10): p. 2744-9.
70. Voulodimos, A., et al., *Deep Learning for Computer Vision: A Brief Review.* Comput Intell Neurosci, 2018. **2018**: p. 7068349.
71. Li, D., et al., *Visual Feature Learning on Video Object and Human Action Detection: A Systematic Review.* Micromachines (Basel), 2021. **13**(1).

72. Gentner, D. and C. Hoyos, *Analogy and Abstraction.* Top Cogn Sci, 2017. **9**(3): p. 672-693.
73. Frankland, S.M. and J.D. Greene, *Concepts and Compositionality: In Search of the Brain's Language of Thought.* Annu Rev Psychol, 2020. **71**: p. 273-303.
74. Burgoon, E.M., M.D. Henderson, and A.B. Markman, *There Are Many Ways to See the Forest for the Trees: A Tour Guide for Abstraction.* Perspect Psychol Sci, 2013. **8**(5): p. 501-20.
75. Borghi, A.M., et al., *The challenge of abstract concepts.* Psychol Bull, 2017. **143**(3): p. 263-292.
76. Buckner, C., *Deep learning: A philosophical introduction.* Philosophy Compass, 2019. **14**(10).
77. Voudouris, K., et al., *Direct Human-AI Comparison in the Animal-AI Environment.* Front Psychol, 2022. **13**: p. 711821.
78. Smith, C., S. Carey, and M. Wiser, *On differentiation: A case study of the development of the concepts of size, weight and density.* Cognition, 1985. **21**: p. 177-237.
79. Goldstone, R.L., *Perceptual learning.* Annu. Rev. Psychol., 1998. **49**: p. 585-612.
80. Bjotk, E., R.A. Bjork, and M. Anderson, *Varieties of goal directed forgetting*, in *Intentional forgetting: Interdisciplinary approaches*, J.M. Golding and C.M. MacLeod, Editors. 1998, Lawrence Erlbaum Associates Publishers. p. 103–137.
81. Williams, M., et al., *The benefit of forgetting.* Psychon Bull Rev, 2013. **20**(2): p. 348-55.
82. Timm, I.J., et al., *Intentional Forgetting in Artificial Intelligence Systems: Perspectives and Challenges*, in *KI 2018: Advances in Artificial Intelligence*. 2018. p. 357-365.
83. Ellwart, T. and A. Kluge, *Psychological Perspectives on Intentional Forgetting: An Overview of Concepts and Literature.* KI - Künstliche Intelligenz, 2018. **33**(1): p. 79-84.
84. Markovitch, S. and P.D. Scott, *The role of forgetting in learning*, in *Proceedings of the fifth international conference on machine learning*. 1998: Ann Arbor, Michigan.
85. Jung, H., et al., *Less-forgetful learning for domain expansion in deep neural networks*, in *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
86. Kirkpatrick, J., et al., *Overcoming catastrophic forgetting in neural networks.* Proc Natl Acad Sci U S A, 2017. **114**(13): p. 3521-3526.
87. Ebrahimi, S., et al., *Remembering for the right reasons: Explanations reduce catastrophic forgetting.* Applied AI Letters, 2021. **2**(4).