Beyond deep learning: charting the next frontiers of affective computing

Andreas Triantafyllopoulos, Lukas Christ, Alexander Gebhard, Xin Jing, Alexander Kathan, Manuel Milling, Iosif Tsangko, Shahin Amiriparian, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Triantafyllopoulos, Andreas, Lukas Christ, Alexander Gebhard, Xin Jing, Alexander Kathan, Manuel Milling, Iosif Tsangko, Shahin Amiriparian, and Björn W. Schuller. 2024. "Beyond deep learning: charting the next frontiers of affective computing." Intelligent Computing 3: 0089. https://doi.org/10.34133/icomputing.0089.

Intelligent Computing
A SCIENCE PARTNER JOURNAL

REVIEW ARTICLE

Beyond Deep Learning: Charting the Next Frontiers of Affective Computing

Andreas Triantafyllopoulos[1,2,3*], Lukas Christ[1], Alexander Gebhard[1,2,3], Xin Jing[1], Alexander Kathan[1], Manuel Milling[1,2,3], Iosif Tsangko[1], Shahin Amiriparian[1,2], and Björn W. Schuller[1,2,3,4,5]

[1]EIHW—Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany. [2]CHI—Chair of Health Informatics, Technical University of Munich, MRI, Munich, Germany. [3]MCML—Munich Center for Machine Learning, Technical University of Munich, Munich, Germany. [4]MDSI—Munich Data Science Institute, Technical University of Munich, Munich, Germany. [5]GLAM—Group on Language, Audio & Music, Imperial College, London, UK.

*Address correspondence to: andreas.triantafyllopoulos@tum.de

Citation: Triantafyllopoulos A, Christ L, Gebhard A, Jing X, Kathan A, Milling M, Tsangko I, Amiriparian S, Schuller BW. Beyond Deep Learning: Charting the Next Frontiers of Affective Computing. *Intell. Comput.* 2024;3:Article 0089. https://doi.org/10.34133/icomputing.0089

Submitted 4 August 2023
Accepted 12 March 2024
Published 16 September 2024

Affective computing (AC), like most other areas of computational research, has benefited tremendously from advances in deep learning (DL). These advances have opened up new horizons in AC research and practice. Yet, as DL dominates the community's attention, there is a danger of overlooking other emerging trends in artificial intelligence (AI) research. Furthermore, over-reliance on one particular technology may lead to stagnating progress. In an attempt to foster the exploration of complementary directions, we provide a concise, easily digestible overview of emerging trends in AI research that stand to play a vital role in solving some of the remaining challenges in AC research. Our overview is driven by the limitations of the current state of the art as it pertains to AC.

Introduction

Affective computing (AC), like all areas of artificial intelligence (AI), has benefited immensely from the rise of deep learning (DL). The first major advancements came in perception—an area in which DL particularly excels, with the last few years seeing an equal, if not more impressive, progress in the generation of affective behaviors. These advances have sparked widespread enthusiasm in the field but also attracted increasing criticism. Most recently, the European Union's proposed AI regulation casts a critical eye on AC and emotion recognition, in particular. A large share of this criticism is inspired by deeper ethical questions that pertain to whether we should create artificial agents with affective capabilities at all—a question that has been considered elsewhere [1].

However, another cause of worry is the brittleness of DL methods in certain situations and its unexpected failure modes. The latter is in large part a byproduct of the success that DL has sparked. With greater performance come ever increasing user expectations, which raises an interesting question: Can DL overcome its current limitations, or are alternative, complementary intelligence paradigms called for? This is the question we address in our current contribution. Specifically, we attempt to chart out these alternative research directions by drawing inspiration from contemporary advances in the wider AI literature.

We note that DL is not the only paradigm being pursued in contemporary AC research; however, it is drawing a great deal of recent attention. The popularity of DL is reflected in the majority of recent reviews, which tend to focus on DL methodologies, as discussed by Wang et al. [2]. In order to further ascertain the community's interest in this field, we collected all journal articles published in *IEEE Xplore* between 2012 and 2023 containing the terms "affective computing" or "affective computing AND deep learning" (in any of their indexed fields). The trend shown in Fig. 1 illustrates the increasing importance of DL (which is mentioned in almost 30% of published journal articles in recent years). Note that our coarse search potentially missed several articles using DL-related terms such as convolutional neural networks (CNNs)/recurrent neural networks/transformers, and is thus underestimating the number of DL-related papers. Thus, while not being the singular focus point of research, DL is nevertheless one of the dominant computational paradigms, which caused us to ask: What lies beyond?

We begin with an introduction of the state of the art; it covers the progress yielded by DL as well as its most prominent shortcomings. We use those shortcomings as anchors for the research areas we cover in our overview, which constitutes a bottom-up "requirements engineering" as covering all the advances in AI would be an unmanageable feat. Rather, we cover those directions we consider to be most pertinent to where AC currently is. This focus brings, by necessity, a subjective element to our overview. Nevertheless, our work aims to inspire future research in emerging topics that can lead us to the next generation of AC architectures.

A Brief Overview of the State of the Art

As our work is geared toward expected future advances, and existing reviews already cover the existing state of the art (see among others [3–6]), we only provide a brief overview of prior work. We begin with a necessary definition of AC, with emphasis on those aspects relevant to our discussion, and follow up with a synopsis of DL, highlighting its major contributions and, more importantly, its current limitations.

A working definition of AC

AC is the field that concerns itself with the understanding and emulation of human affect. In the present context, we use affect in its broadest connotation, which includes emotions, mood, interpersonal stances, attitudes, and affective personality traits, as per the taxonomy of [7,8]. Furthermore, we consider both the analysis and synthesis of affect, as both aspects play an equally important role in AC. Finally, we do not differentiate among the different modalities (or signals) that have been used to analyze or portray affect; indeed, as we later discuss, one of the defining characteristics of DL is the substitution of customized, task-specific pipelines with powerful, generic learners that can learn any task given sufficient data. While not restricting ourselves to a specific modality makes our scope broader, it allows us to unify trends that are observed across the different strata of AC research.

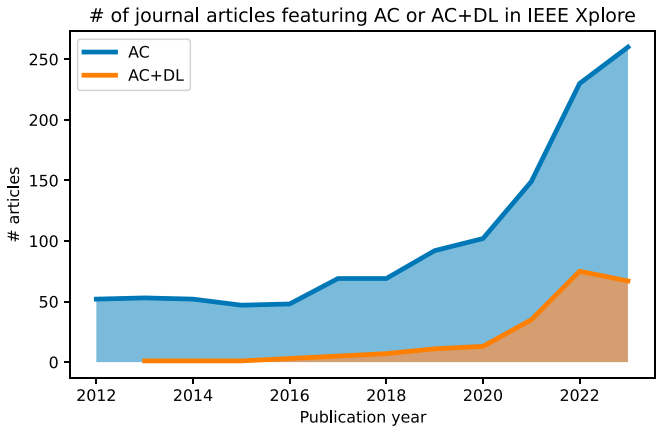

Fig. 1. Number (#) of journal publications appearing in *IEEE Xplore* between 2012 and 2023 featuring the terms "affective computing" (AC) or "affective computing" AND "deep learning" (AC + DL).

Of all the affective states that are of interest to the computational community, emotion has perhaps played the most prominent role. Underlying computational approaches to emotion are the different theoretical models [9]: From Ekman's "big 6" and other categorical models [10], to Russel's circumplex model [11] and emotional dimensions, to appraisal theories and Scherer's emotional component model [12], several different constructs have been proposed and used in AC research. To date, the most dominant one is probably the categorical model, followed by the dimensional one, with alternatives receiving far less attention.

However, categorical descriptions of emotion are not necessarily the most accurate or fitting representations. Contrary to other emotion theory families, appraisal theories seek to explain the whole "emotion process." Therefore, Scherer [13] advocates basing emotional agents on them. A range of architectures and prototypes for such systems have been proposed in recent years, e.g., CAIO [14], ABC-EBDI [15], and Silicon Coppelia [16], the vast majority of them indeed founded on an appraisal framework. For a comprehensive overview of this line of research, see Zall and Kangavari [17].

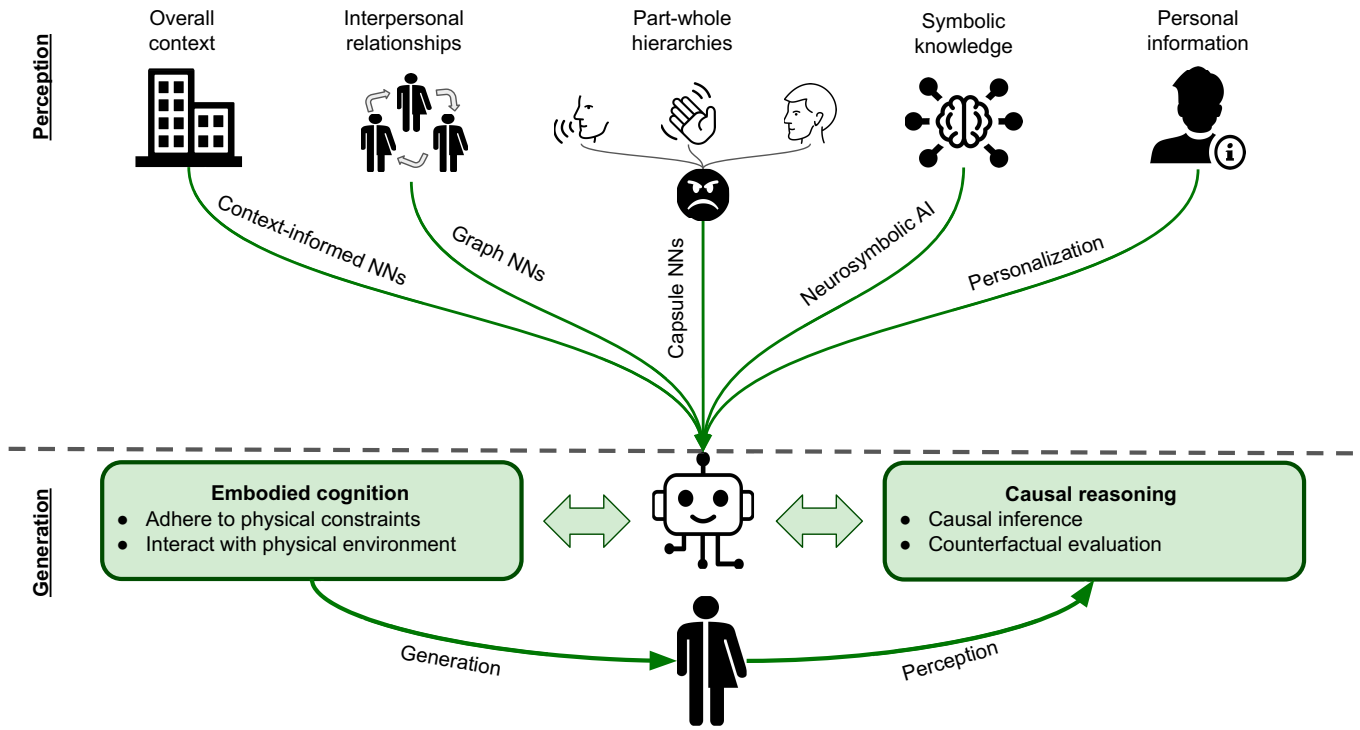

Fig. 2. A schematic diagram of the envisioned system.

Furthermore, emotional reactions do not exist in a vacuum but are interspersed in other functionalities of an agent in its environment. In particular, the agent must be able to sense this environment and produce its responses in a way that accounts for emotional "appropriateness"—a feat indicative of the agent's emotional competence. Scherer [13] characterizes emotion production competence as consisting of 3 aspects: (1) appraisal competence, (2) regulation competence, and (3) communication competence. The appraisal competence of artificial emotional agents is their capability of judging a given situation correctly and updating their emotional state accordingly. Regulation competence is the capability of an agent to appropriately manage and, if necessary, readjust its emotional state. As this capability lies beyond our current scope, we will not discuss it further. Last, communication competence entails both understanding the emotions expressed by others and being capable of expressing emotions in such a way that others can be expected to be able to decode them. In particular, emotion expression has to consider the sociocultural context and aim for appropriateness.

We note that similar models exist for other affective states. For example, personality is typically evaluated with OCEAN [18] or other such models, whereas mood is (self-)evaluated using the Hamilton Depression Scale (HAMD) [19] or similar constructs—although this line of research is most commonly pursued in mental health research. Irrespective of the underlying construct, though, AC is typically tasked with predicting or simulating it.

With respect to the modalities used to model and portray these states, there exists an equally broad spectrum of sensors and signals. The 3 major types are vision, audio, and text, and they are used for both the analysis and synthesis of affect. Visual cues such as facial expressions and gestures are among the most prominent signals investigated by AC researchers [20,21]. Voice is another major signal for the modeling of affective expression [7,8]. Textual analysis is often pursued in isolation from speech as text can be produced by other means (writing) [22]. In addition to those modalities, physiological signals, such as electrodermal activity, heart rate, and electrocardiograms, can be passively and ubiquitously monitored and are highly correlated with affective arousal [23]. Wearable sensors, such as GPS, biomechanical, or tactile sensors, can be used to track mood (mostly in depressed individuals) or well-being [24–26]. However, in these cases, research has focused exclusively on analysis, and not synthesis.

The state of the art in DL

Definition

There are many different interpretations of DL. In fact, as DL is a relatively novel machine learning paradigm, its theoretical foundations are still an area of active research. For our purposes, we focus on its key differentiating characteristic, namely, depth.

Depth arises from stacking together multiple learnable modules, or layers, which are jointly trained to optimize a given objective. The building blocks of DL architectures are relatively simple: linear or convolutional layers, sigmoid or rectified activation functions, different normalizations. Their representation power arises when several are composed in a hierarchical computational graph, with the first layer processing the (raw) input, the second layer processing the output of the first, and so on until the last layer, which produces the output. Each layer then successively generates a higher-level representation, a "deeper" form of abstraction derived over its input, until the last layer, tasked to generate the highest form of abstraction (e.g., the target classes).

A key motivation for DL is the potential to circumvent traditional feature engineering [27]. Allowing a general-purpose learning machine to jointly learn both the features and the predictive "rules" it needs from data is a crucial advantage. Traditional, shallow methods are only as good as their features, which are ultimately as good as the engineers who design them.

Contributions

The advances facilitated by DL are astounding. The most notable one is their superior performance in a vast array of tasks, beginning with computer vision [28,29], speech recognition [30], and natural language processing (NLP) [31], and quickly spreading to encompass most application fields of machine learning [32,33]. AC was also one of the early adopters of this technology, with DL models quickly becoming the new state of the art in facial [21,34] and speech emotion recognition [35] and other application areas.

The biggest improvement came in the form of increased performance; deep neural networks (DNNs) were able to achieve impressive gains over traditional, shallow baselines, especially in those affective dimensions that were previously more challenging, such as valence in the case of speech [33]. Another major improvement came in the form of robustness, with deep models showing better generalization across noise conditions over handcrafted feature-based methods [36]. Although DNNs were later found to be susceptible to "adversarial attacks" [37] (small perturbations that can considerably reduce their performance), their robustness to more realistic degradations that plagued traditional approaches is still remarkable [33].

Another key innovation of DNNs was their conduciveness to transfer learning, and in particular their ability to consume large datasets and produce generalizable representations that can be used to effectively solve downstream tasks [27], an ability that was vastly improved over that of shallower architectures [38]. Further advances in learning without labels allowed the scaling of the amount of available data [39], resulting in training datasets that effectively capture a large part of the entire Internet [40]. Oftentimes, knowledge transfer traversed the boundaries of a single modality, with researchers utilizing the power of learned representations in one modality to improve performance in another, as in the case of image-to-audio transfer learning [41,42].

These trends have accelerated with the advent of transformers [31] and self-supervised learning. Transformers are architectures that rely on a stack of "self-attention" blocks: operators that were originally introduced for NLP and have since found widespread use in computer vision and audition [33,42,43]. The introduction of this new class of models coincided with important advances in self-supervised learning, a form of nonsupervised learning where a network is trained on a set of proxy tasks derived from the data itself. This practice is so widespread that these models have been recently labeled "foundational" and now form the building blocks on which most contemporary machine learning research is founded [44]. Similar advances have also been driving the recent explosion in generative AI [3], which is equally important for the generation of affective behavior.

Limitations

Despite the tremendous performance improvements facilitated by DL, there still remains much ground to be covered before AC applications can meet (ever increasing) user expectations. In the present section, we discuss the missing pieces with an emphasis on those we believe require us to move beyond DL and toward other AI paradigms.

The biggest obstacle is out-of-domain generalization [45]. Now that in-domain performance on several tasks is approaching human levels, researchers are increasingly evaluating their models across a wider set of domains. Their efforts have highlighted the inability of models to generalize even under relatively small distribution shifts that present no challenge to a human. Rather than a shortcoming of DL per se, this particular behavior underlies all statistical learning, as one of its most foundational assumptions is that source data must be identically and independently distributed. However, this assumption rarely holds in practice. Instead, data are often sourced from specific, constrained environments under a certain set of conditions. When these conditions change, a phenomenon corresponding to a distribution or a concept shift [46], models are unable to generalize. While considerable efforts are being currently put into overcoming this problem, we believe that the solution is unlikely to come from more depth; thus, it is something that requires additional research beyond DL.

Beyond simple robustness to noise and different conditions, there is also the issue of generalization across different cultural settings [13]. As different social norms hold in different cultures, established machine learning methods as mentioned above may not yield good results in sociocultural contexts underrepresented in the training data. The desire to effectively apply established methods in new contexts motivates research on cross-cultural transfer, which has received increasing attention recently. The multimodal SEWA [47] database comprises valence/arousal-annotated recordings of persons from 6 different cultures. Hume-VB [48] contains emotional vocal bursts by subjects from 4 different countries. For Hume-Reaction [49,50], individuals from several different countries rated audiovisual recordings of their own emotional reactions to different stimuli. However, cross-cultural performance still lies considerably below within-culture performance.

The identically and independently distributed assumption underpinning DL also stumbles against another fundamental aspect of AC: Oftentimes, data are collected longitudinally from the same individual [25,51]. The individual can be a user of an emotion recognition app, or a patient tracked under a mental health study. In both cases, the data are neither independent nor identical: Each user has unique characteristics that affect how the observed phenomenon manifests in a given signal. Oftentimes, it is also the case that tracking the changes in an individual's state over time is explicitly the goal of an application, for example, in the case of tracking the mood of depressed patients [52]. Enabling adaptation to individual needs requires the introduction of personalized models [53], rather than population-level ones, which are the norm in current literature.

One additional related shortcoming is the inability of the models to perform causal reasoning. Once more, this limitation underpins all statistical learning models, as such models reside by definition at the lowest rung of Pearl's causal ladder [54], namely, the level of understanding "associations" (with interventions and counterfactual reasoning being the upper layers). On the one hand, the inability to account for causal factors exacerbates the out-of-domain generalization issues of the models. This is because small changes in the causal graph of a particular phenomenon can result in large changes in the signal space. To take one example from facial analysis, age has a big influence on the signal space: The face of the same person can change considerably across different periods of that person's life, but the changes in the underlying causal graph are restricted to just a few bits (assuming, for example, that age is represented in decades). Incorporating this level of understanding into DNNs requires a new wave of theoretical advances.

However, the pursuit of causal reasoning does not stop with ameliorating distribution shifts. Rather, there is the much bigger goal of disentangling causal effects. This disentanglement is particularly important for the digital health aspect of AC. Understanding what caused a particular effect that is detected in a signal could unlock a new era for affective applications. For example, knowing what caused a particular emotion, rather than simply identifying what that emotion is, can help an affective agent react properly to a user's behavior. Currently, such a form of reasoning is beyond the capabilities of DL models, and infusing these models with causal understanding is becoming one of the most active areas of current research.

There are 2 facets to understanding the "why" of a decision: One is to find its cause, as discussed above; the other is to interpret why a system made the decision it did. Successful decision analysis will improve transparency and user trustability, which are both necessary preconditions for the acceptance of a new technology. In general, interpretability is one area where DNNs fare worse than their predecessors. Given their increased complexity, which largely arises out of their depth, these models became completely opaque to layman users and researchers alike. While there is a new wave of methods attempting to illuminate the "black box" of DL [55–58], these fall short of the level and detail required for a seamless integration of those models in real-life applications.

Finally, a major downside of DNNs is their increased computational overhead. This overhead translates to much higher training and inference costs, which have a huge environmental impact and form a barrier for the users and developers of that technology who cannot afford the cost. The community is well aware of this issue, and concentrated efforts are being spent on making DNNs smaller and faster to run [59], but this pursuit of efficiency may well include new forms of learning that go beyond depth.

Collectively, we summarize the following affordances that an affective agent needs to provide and are currently missing from DL architectures in Table:

1. Generalizability amounts to performing equally well under distribution and concept shifts (e.g., different environments or cultures).
2. Situatedness highlights the importance of adapting to the specific situation (e.g., to the broader context or the interlocutor)
3. Interactivity encapsulates the ability of an agent to enact changes in its environment.
4. Causal reasoning entails higher cognitive abilities related to causal understanding and planning.
5. Transparency translates to the interpretability of an agent's decisions and reasoning steps (if not necessarily its inner workings).
6. Efficiency in terms of energy use and computational resources is a necessary prerequisite of deploying these agents in real-world settings.

Table. Overview of the different research areas included in our overview and how each one maps to the intersection between the limitations of DL and the affordances required by affective agents

	Capsules (section "Capsule networks")	Geometric DL (section "Geometric DL")	Spiking DNNs (section "Spiking neural networks")	Neurosymbolic AI (section "Neurosymbolic intelligence")	Context (section "Context-informed neural networks")	Embodiment (section "Embodied cognition")	Generation (section "Generation")	Personalization (section "Personalization")	Causality (section "Causality")
Generalizability	✓	✓	✗	✓	✗	✗	✗	✗	✓
Situatedness	✓	✗	✗	✗	✓	✓	✗	✓	✓
Interactivity	✗	✗	✗	✗	✗	✓	✓	✗	✗
Causal reasoning	✗	✗	✗	✗	✗	✗	✗	✗	✓
Transparency	✗	✗	✗	✓	✗	✗	✗	✗	✓
Efficiency	✗	✗	✓	✗	✗	✗	✗	✗	✗

Table maps the theoretical advances included in our overview to these limitations of DL. This formed the starting point for the rest of our work.

Next-Generation Neural Networks

We begin our overview of next-generation AI paradigms by considering new network architectures. While depth might be a common feature with current DL architectures, the next generation must introduce some fundamental difference other than depth. The most promising paradigms are capsule networks, which aim to overcome the inability of DNNs to learn complex part-whole hierarchies, spiking neural networks, which constitute a more "natural" approximation of biological neural systems and come with additional computational benefits, and geometric DL, which guides the design of a new class of models using invariance properties in different topological spaces. All 3 address specific shortcomings of DNNs and thus hold great promise for future AC applications. While they have seen important advances in the broader AI field, their uptake within the AC community is rather limited—a trend we expect to change in the coming years. In the following sections, we proceed to discuss them in more detail.

Capsule networks

CNNs have become indispensable to several modern computer vision applications [29], ranging from image recognition over object detection to image segmentation. Moreover, they not only are limited to the computer vision domain but also can be found in a variety of other areas, such as computer audition and NLP [60]. Nevertheless, despite their achievements over the recent years, they suffer some considerable drawbacks. For example, CNNs typically implement pooling layers, which tend to lose some of the features and spatial information of the image. This loss makes them invariant to translation, which allows them to detect features detached from the location in an image. However, it also means that they are not able to recognize the pose or deformations of an entity [61]. Accordingly, they cannot capture part-whole hierarchies, i.e., interlink the different parts that constitute a separate entity.

Hinton et al. [62] proposed capsule networks (CapsNets) as a solution to these issues for the field of computer vision. The primary novelty of CapsNets is the substitution of layers with "capsules." Each capsule outputs the probability of the existence as well as the instantiation parameters of an entity. For example, these instantiation parameters may comprise the pose or deformation of the entity. Thus, whenever an entity is deformed, its instantiation parameters change as well, which makes them equivariant to transformations. Using this mechanism, CapsNets can preserve information about the location and pose of an object throughout the network.

More importantly, CapsNets first capture the parts of an entity before they recognize the entity as a whole. That is, if the predictions of capsules in one layer are "agreeing," they are assumed to have the correct spatial relationship and activate a higher-level capsule in the next layer. For instance, the parts of a human body include arms, feet, and the torso. In order to form a human body, these parts have to be at the correct locations and have the right spatial relationships to each other. When each capsule represents a body part, the predictions of the capsules have to agree in order for the higher-level capsule to be activated and recognize the human body as a whole. This ability of capturing part-whole hierarchies is a key contribution of CapsNets.

Several distinct implementations of capsules have been proposed, starting in 2011 with transforming autoencoders [62], followed by capsules with vector outputs that employ routing by agreement (dynamic routing) [63], and matrix capsules using expectation–maximization routing (EM routing) [64]. The first successful approach that accelerated the uptake of CapsNets was the one introduced by Sabour et al. [63], where routing by agreement is utilized and the capsules have vectors as input and output. Each activation vector represents the instantiation parameters of an entity, and its length determines the probability of the existence of that entity. The proposed architecture is widely used in the literature and serves as the basis for modifications and further improvements. The

rough CapsNet structure is as follows: First, there is a convolutional layer that extracts initial features. These features are then fed to a primary (lower-level) capsule layer. Afterward, there can be several higher-level capsule layers with a final class capsule layer as the classification layer to obtain the class predictions. The routing is only conducted between 2 consecutive capsule layers [63]. Finally, there is a decoder to reconstruct the input based on the final activation vectors from the last capsules. Following this initial implementation, there have been several modifications and improvements, as well as widespread application of CapsNets in different areas, such as healthcare [65–69], autonomous driving [70,71], and NLP [72,73].

CapsNets have also been successfully applied to AC, in the context of different modalities. For instance, Liu et al. [74] and Li et al. [75] both utilize a CapsNet to recognize emotions based on multichannel electroencephalogram (EEG) signals. Their motivation was to learn the intrinsic relationships among various EEG channels. In addition to EEG signals, several works have explored the effectiveness of CapsNets for speech emotion recognition (SER). For example, they have been used to capture the spatial information in input features like spectrograms, e.g., the positional and relationship information of low-level features including pitch and formant frequencies [76,77]. In this regard, sequential capsules have been introduced in order to best handle the sequential nature of input feature frames in SER and thus optimize on the whole sequence [76,78].

Finally, as CapsNets were initially developed for the computer vision domain, they have an obvious application to visual AC, and in particular facial expression recognition. A human face comprises spatial relationships among its different parts (such as eyes, mouth, and nose), which is important for recognizing a human face. Since CapsNets were developed to capture exactly these kinds of information and incorporate them into a part-whole hierarchy, they are highly suited to this task. Accordingly, employing them for facial expression recognition results in both performance improvements and increased robustness compared to CNNs [79,80].

Overall, CapsNets are a relative newcomer to the field of AI (although traces of the idea can be found in much earlier works of Hinton [81,82]). Crucially, they have been introduced to address the problem of capturing part-whole hierarchies, which remains a fundamental issue in "vanilla" DNNs. Capturing part-whole hierarchies is a critical component of AC systems, especially systems designed around the idea of appraisal theories, which require the understanding of the different parts that led to the emergence of an emotion (see the "A working definition of AC" section). Incorporating CapsNets into AC will improve the generalizability of systems, as it enables compositionality over different parts, as well as facilitate improved situatedness by means of capturing these part-whole hierarchies. As we later discuss, capturing part-whole hierarchies is connected to the more comprehensive modeling of the surrounding context; indeed, if emotions can be seen as reactions to external events, then incorporating those events as "parts" of some part-whole hierarchy should enable a better understanding of affect in naturalistic environments.

Geometric DL

The notion of geometric DL was initially introduced by Bronstein et al. [83], where the authors tackled the challenge of DL in non-Euclidean geometric structures like graph neural networks (GNNs) and manifolds. The authors introduced a constructive method for classifying the known DL architectures based on the symmetries of their domain spaces. This constructive approach allows for the design of more modular architectures, which are adapted to the idiosyncrasies of different signals and tasks (i.e., the introduction of suitable inductive biases).

These efforts culminated in the Geometric Deep Learning Blueprint [84], a scheme for constructing functions that exhibit desirable invariance—or equivariance—properties. A typical instance of this setup is that of a function defined on a graph $\mathcal{G}$. Such a function should be permutation-invariant, i.e., the output of the function should be identical irrespective of the order in which the inputs are given. In AC, such permutation invariance can be useful in the case of multiparty interactions, where the order in which the different agents are given should leave the outcome unaffected.

In the case of MeshCNN [85], the domain consists instead of a manifold where the pertaining functions should preserve the distances—i.e., be isometry-invariant. The key idea is to explicitly introduce invariance into certain transformations, thus making the system less prone to small variations in the data. However, even though this scheme can provide useful insights for the creation of robust models, learning invariances is challenging depending on the underlying modality [86]; thus, there is growing research interest in this direction [87–91].

Research has so far largely focused on the 2 most prominent non-Euclidean structures: graphs and manifolds. With applications ranging from tasks such as node classification in social media [92] to tasks such as the prediction of microstructures in materials science [93], the design of a GNN consists of node update operations

$$x_u = \phi\left(x_u, \bigoplus_{v \in \mathcal{N}_u} \psi\left(x_u, x_v\right)\right)$$

such as summation or averaging. The success of GNN frameworks, such as the graph convolutional network [94], can be traced back to this structure, which may comprise distinct methods of information aggregation over neighborhood nodes and edges [94,95]. Essentially, this aggregation allows information to propagate among interconnected entities; by carefully configuring the edges and nodes of the graph, one can incorporate prior knowledge about the downstream task. GNNs are typically shallow, and increasing their depth can produce adverse effects such as over-smoothing [96]: nodes collapsing to a single feature vector. Nevertheless, more research is ongoing toward adding depth to GNNs [97,98] to alleviate emerging problems.

GNNs have already been successfully applied in AC applications, most often in the field of sentiment analysis [99,100], but also in facial analysis [101,102]. They have also been used for affect-aware recommendation, where affective information is taken into account when providing personalized suggestions [103]. In general, the malleability of GNNs makes them ideal for modeling the relationships and interactions between different entities, such as users, items, and events. These multiway interactions can form the basis for the holistic comprehension of the surrounding context, and in turn lead to an improved understanding of the factors leading to an affective expression by one of the users in an environment.

Geometric DL also goes beyond graphs and allows for more complex constructs, such as manifolds. A manifold is mathematically defined as an n-dimensional, "well-behaved" topological space, which is locally homeomorphic to Euclidean

space $\mathbb{R}^n$. The prototypical example of a manifold is the sphere $\mathbb{S}^2$, a subspace of $\mathbb{R}^3$ that can be locally perceived as $\mathbb{R}^2$. This modeling is convenient when considering 3D shapes in the fields of computer graphics and computer vision, and its applications encompass meshes, point clouds, and convexes [104,105]. However, current implementations still suffer from high memory consumption and slow training [106]. In addition to the classical manifold learning methods traditionally used for high dimensional data, recent research demonstrates that manifold learning combined with a graph convolutional network can considerably improve performance on image classification tasks [107]. In the context of AC, manifold learning can provide a toolkit for analyzing emotions in 3-dimensional (3D) environments, e.g., by means of body gestures or facial expressions [20,108]. We expect that integrating better inductive biases into our architectures will vastly improve their generalizability.

Spiking neural networks

A promising candidate to tackle the problem of the high computational costs of training large-scale DNNs are spiking neural networks (SNNs), which have been recently seeing increased usage in AC [109–111]. The development of spiking neurons goes back to the 1950s [112], but has been overshadowed by the overwhelming success of DNNs since the end of the most recent AI winter. The design of spiking neurons can be considered a step closer to biological neurons compared to the artificial neurons encapsulated in the most commonly used DL architectures. The design of a standard neuron is well known in the research community, and its output is commonly denoted as

$$o = f(\mathbf{W}\mathbf{x} + b), \tag{1}$$

with the inputs $\mathbf{x}$, a weight vector $\mathbf{W}$, a bias b, and a nonlinear activation function f. In DL architectures, multiple layers of such neurons are stacked in a sequential manner and the outputs of the network are calculated in a feed-forward manner, with information propagating from the input layer to the output layer in a sequential fashion. In biological neurons, this propagation corresponds to the constant-rate firing of impulses throughout the entire network. Such networks thus lack at their core the ability to simultaneously process neuronal output at different layers with a continuous flow of time [113].

Modern SNNs, referred to as the third generation of neurons, incorporate simultaneous processing via an internal state, which is dynamically changing with respect to its inputs and also with respect to time. Once a critical threshold of the internal state is reached, a spiking neuron fires. Among spiking neurons, there exist a variety of designs, which, in general, follow time-dependent differential equations inspired by nature. The dynamic nature of spiking neurons can be seen in the exponential decay over time in the commonly applied leaky integrate-and-fire (LIF) model [114,115]

$$V_{\mathrm{mp}}(t_p) = V_{\mathrm{mp}}(t_{p-1})\exp\frac{t_{p-1}-t_p}{\tau_{\mathrm{mp}}} + w_i^{(p)} w_{\mathrm{dyn}}, \tag{2}$$

with the membrane potential V_{mp} and the membrane time constant τ_{mp}, the timings of previous input spikes t_p and t_{p-1}, as well as synaptic and dynamic weights $w_i^{(p)}$ and w_{dyn}. The membrane potential of the LIF neuron thus has an exponential decay over time and is raised proportionally to a connection-specific weight when a new spike arrives, with the dynamic w_{dyn} weight hindering the impact of new inputs for a specified duration of time. The LIF model is thus characteristic of an SNN architecture, where the output potential of a neuron depends on the timing of its inputs.

Alternative spiking neuron models are explored in other works, such as Yamazaki et al. [113]. Architectures of the SNNs, i.e., connected systems of spiking neurons, are often similar to those of artificial neural networks (ANNs), but with a different processing unit at their core. Consequentially, adjustments to the backpropagation algorithm commonly applied to ANNs can be made to train SNNs in a supervised fashion [115,116].

Crucially, as each neuron fires only when a threshold is reached, an SNN comes with a greatly improved energy efficiency compared to traditional ANNs. This is because the default state of a neuron in hardware can be denoted as a "0" state (no voltage) that is only set to "1" when firing. In contrast, an ANN would "fire" all its neurons during a forward pass, resulting in much higher energy consumption. This improved efficiency allows the deployment of ANNs on low-resource computational units that can vastly extend the outreach of affective intelligence. It also reduces the computational costs for GNNs [117] (thus helping overcome an obstacle to increasing their depth) and enables the deployment of deep SNNs in real-world applications with low latency, thus substantially improving their efficiency. Real-world efficiency is especially important in connection with the ideas discussed in the next section, where we consider embodied affective agents that are able to sense and interact with their surrounding environment.

Resurgent Themes

DL itself is to some extent a resuscitation of an old idea using modern tools. The rise of DL illustrates how the process of "rediscovering" old ideas and adapting them to a new context is a promising research avenue. While other old ideas might have failed in the past, they might have done so based on criteria on which DL excels. However, their strengths might complement the shortcomings of DNNs, and other network architectures in general. Therefore, their introduction as additional components in AC applications can help supply the deficiencies discussed in the "Limitations" section. In particular, we expect hybrid neurosymbolic architectures to play an important role in integrating expert knowledge and helping overcome the lack of causal understanding, and embodied and context-informed agents to become mainstays of next-generation AC.

Neurosymbolic intelligence

DL adheres to the connectionist paradigm of cognition, which contrasts with the "traditional" symbolic paradigm pursued in the earlier stages of AI research. Indeed, a key distinction in the design of neural network architectures is their ability to learn distributed representations: Information is represented in multiple neurons within a neural network, and each neuron may focus on different facets of a given input. This type of representation is in opposition to "classic" AI, which emphasizes explicit symbolic manipulation, with each symbol having one concrete interpretation. While the former has proven more conducive to learning from (large amounts of) data, the latter has clear advantages in the form of robustness (evaluating a symbolic expression with the same inputs always returning the same result), out-of-domain generalization (plugging in a previously unseen value for a symbol in an equation will not

cause the model to break), and explainability (assuming that the symbols themselves are interpretable). Recent years have seen a renaissance in attempts to bridge the gap between these 2 approaches in order to overcome these particular limitations of DNNs [118–120].

The simplest way to integrate the 2 paradigms uses DNNs to process symbolic inputs and generate symbolic outputs; the DNN then is trained—as usual—to learn a mapping between input and output. Alternatively, the output of one or several DNNs that predict certain statistical variables may be propagated to a (cascade) symbolic solver for further processing. However, both of these approaches are somewhat "shallow" in nature, since the first uses a DNN to perform symbolic reasoning and the other reasons over the outputs of a neural architecture, but none of them integrates symbolic reasoning into a DNN—which is the key distinction of a neurosymbolic system.

In his 2023 AAII keynote speech, Kautz [120] outlined 6 different architecture designs for neurosymbolic systems:

1. Symbolic Neuro symbolic systems, which are currently the standard in DL-based AI, pass symbolic inputs to DL models (e.g., in the form of words), which subsequently output symbols in return (i.e., the predicted categories at the output layer).

2. Symbolic[Neuro] systems combine symbolic problem solvers with DNNs for pattern recognition. A prime example of such an architecture is AlphaGo, which uses Monte-Carlo Tree Search for planning its next moves while relying on DNNs for evaluating the game state.

3. Neuro|Symbolic methods rely on the conversion of non-symbolic inputs (e.g., pixels of an image) to symbols that can be manipulated by a symbolic engine.

4. Neuro: Symbolic → Neuro paradigms rely on DNNs that are trained on outputs generated by symbolic rules and that subsequently learn to emulate these symbolic functions without explicit logic operations.

5. Neuro_{Symbolic} approaches implement symbolic rules as neural operations and embed them within a DNN as specialized, trainable subnetworks.

6. Neuro[Symbolic] perspectives pursue a tighter integration of neural and symbolic architectures in a setup inspired by Kahneman's "fast and slow thinking." The neural components handle the "faster" thinking tasks at which they excel (e.g., pattern recognition and action-taking), while the symbolic parts are responsible for longer-term planning and reasoning.

In the AC domain, neurosymbolic systems can be useful in multiple ways: They allow the integration of expert background knowledge in the form of symbolic equations. Thus, expert knowledge can be used to constrain the outputs of a neural network to a meaningful domain (as seen in the "Context-informed neural networks" section). In addition, they facilitate a human-in-the-loop paradigm where a user continuously guides an affective agent through the definition of certain symbolic variables. As seen in the "Personalisation" section, this paradigm can be used to adapt to (new) individuals that interact with the agent. More importantly, such capabilities allow the translation of AC technologies to new domains. For instance, a key issue is the modeling and synthesis of affective states in cultures differing substantially from the cultures a DNN system was based on. Affect may both be expressed differently and be enshrouded in different cultural norms, thus hampering the generalization of the system. Providing a way to configure and convey the systems for these new domains can be critical for the fair application of AC, as it will enhance the transparency, and thus the trustability, of those systems and make them more generalizable to new contexts.

Context-informed neural networks

As AC applications increasingly find their way into the real world, there is a pressing need to accommodate ever-growing user expectations. Going beyond traditional performance metrics such as accuracy and correlation, such accommodation entails the impetus of conforming to social norms that underlie human behavior. Conforming to social norms is particularly important for the synthesis of affective states: Rather than being correct according to some predefined measure of success, the requirement in the real-world is to be appropriate—as dictated by the enveloping context. Context is determined by both the external environment and the agent's role in it, and may include the location, the application in which the agent is embedded, the interlocutor or receiver of the agent's affective message, etc. Achieving the objective of contextual appropriateness requires accounting for these externalities.

Appraisal-based affective systems, in particular, are often dependent on context [17]. Existing methods typically rely on rule-based approaches to model appraisal processes and the resulting emotional expressions. As a notable exception, Hoorn et al. [16] utilize a machine learning model to predict the attractiveness of a person as one appraisal variable. Moreover, due to the considerable manual effort of designing appraisal rules, these approaches have only been evaluated for specific and simplified proof-of-concept applications, e.g., a doctor telling bad news to a patient [15] and a simple dating scenario [16].

Understanding the overarching context is also crucial to achieving the goal of communication competence [13]. As for the understanding of others' emotions, a plethora of studies have shown that emotion recognition techniques benefit from taking contextual information into account. As an example, the task of emotion recognition in both text-only (e.g. [121]) and multimodal dialogue settings (e.g. [122]) can be considered. Here, numerous methods, e.g. [123–125], have been proposed that predict the emotionality of an utterance considering the utterances preceding it, leading to improved performance compared to emotion recognition on isolated utterances only.

When it comes to emotionally responding to other agents, an artificial agent should therefore conform to social rules and expectations. Part of a solution to this problem may be to leverage manually curated databases of common-sense knowledge to infer such information. For example, ATOMIC [126] provides about 900,000 crowdsourced If-Event-Then triples in the form (<event>, <relation>, <event>). The types of relations between events include the expected reactions by others and the possible intentions motivating someone to cause an event. More context-oriented, CICERO [127] comprises about 50,000 common-sense inferences from dialogues, among them motivations and reactions. Ghosal et al. [128] demonstrate that textual transformer models trained on such databases are capable of inferring common-sense rules that are not explicitly part of the database. One major challenge here is that existing data are likely not applicable to multiple sociocultural settings. Another research question is how an agent should handle these kinds of norms once they are available.

Inspiration for a principled way to adhere to additional requirements other than correctness may be found in the use of DNNs in physics. There is a recent wave of approaches falling under the umbrella of physics-informed neural networks [129], which constrain the output of a DNN so that it conforms to the laws of physics that apply to the problem it attempts to solve. Translating that to AC requires us to satisfy context-based "laws": These can be the norms of the culture the agent is embedded in, the interlocutor's identity and personal history, or generally anything that can be considered relevant context. These constraints may be provided in the form of symbolic expressions, thus providing a bridge between context-informed neural networks and the neurosymbolic systems discussed before. Including more context is a necessary aspect of situatedness for an affective agent, and this goal can be achieved by introducing these constraints during system training and inference.

Embodied cognition

Perhaps the most extreme version of context is embodiment, with all the implications it entails for an affective agent. Embodied cognition can be viewed as the polar opposite of pure symbolic reasoning [130,131]; instead of abstract symbols that are manipulated according to some predefined set of (logic) rules and a knowledge base acquired through experience, embodied cognition posits an agent inextricably grounded in physical reality [132]. This form of cognition does not deny the need for good (world) representations; it merely subjugates them to the practical needs arising from an agent embedded in its environment and the concrete goals it pursues in it [130]. We note that embodied intelligence does not presuppose an actual physical body in the "real" world: The agent's environment may be entirely or partially digital, its physical sensors may be constrained to a few modalities (e.g., it may be blind or deaf), and it may be unable to move. What is important is that the agent may purposefully pursue its goals under a set of physical constraints.

The importance of embodiment for AC is most evident in the case of synthesis: The portrayal of affective states must be appropriate for the state of the environment in which an agent finds itself [133]. This requirement is particularly strong for gestures, or more generally "body language," which needs to conform to the physical constraints of the agent. However, other modalities may benefit from embodiment as well. Using facial expressions to intentionally convey an affective state relies on direct line of sight; without it, such expressions are meaningless. The agent may thus choose to reposition itself before enacting an expression. Speech, too, may be adapted to the surrounding physical space. Loudness, for example, takes on different meanings depending on the distance between 2 interlocutors; being loud when the distance is large can be interpreted as an attempt to boost communication, whereas it may be interpreted as a signal of aggression if the distance is small. Accounting for these constraints can therefore lead to more natural affective expressions. Crucially, the agent must ensure a degree of congruity across all different media of expression; otherwise, the (human) receiver of a message will struggle to decode it [134].

Exhibiting empathy through mimicry is one of the affective affordances that stand to gain the most from embodied cognition, given the fact that humans appear to be experiencing emotions in an embodied fashion [134,135]. Mimicking the body language, facial expressions, and speaking style of their interlocutors might be necessary in order for agents to build rapport with them. Mimicry requires the agents to monitor and mirror the affective expressions of their interlocutors. Importantly, given the unavoidable physical differences between humans and robotic agents, they should not attempt a naive reproduction, but rather map human expressions to their own characteristics—a feat that requires them to be aware of these characteristics.

The understanding of affect may benefit as well from situating the agent in a physical environment. In particular, it may facilitate the understanding of natural, contextualized affective responses that result from an interaction with the environment. A loud bang, for instance, is expected to cause sudden (surprise) fear in a human. Learning those associations can help the agent reason about the causes, and expected effects, of such external events on the humans it interacts with. Such understanding may help determine both if the response is relevant to the agent and the appropriate response to it.

Recent advances in embodied AI can be readily assimilated to AC. One of the most exciting avenues of current research is the simulation of realistic environments for the purpose of training (multi)agent systems [136]. These environments are used for the training of agents via reinforcement learning (RL). Video games, in particular, have become a focal point of experimentation in RL agents [137]. These environments form a natural training ground for affective agents as well, as they foster natural interaction between multiple entities and, crucially, entail working toward solving real-world tasks under physical constraints.

RL allows a move away from labeled data and facilitates interactive learning, which can overcome the downsides of the standard supervised learning paradigm. For AC, the tasks differ from the standard goals pursued by most RL agents (i.e., navigation). The focus is instead on objectives where AC is relevant: collaboration with humans in education, healthcare, etc.

In general, embodiment is a crucial step toward achieving better situatedness and interactivity for affective agents. It will allow them to better understand their environment and overarching context while simultaneously providing them with numerous interfaces that can be exploited to interact with that environment and other entities inside it.

New Frontiers for AC

In this last section, we outline ideas that have seen substantial progress in the last "few" years (i.e., in the 21st century). While their roots go back decades, or even centuries, we believe that recent advances have set these ideas apart from the crowd of the resurgent themes we discussed above. The first such frontier is generation; while it has been considered a holy grail of AI for decades, the recent advances of large generative models, largely based on diffusion processes, mark a new era in the field. The next one is personalization, which encompasses an array of methods that move away from generic, population-level models and toward models that account for the characteristics of an individual subject. Finally, we provide an overview of causality, whose mathematical foundations have been pioneered by Pearl [54]—an achievement that led him to win the 2011 Turing Award. We note that due to the relatively recent emergence of these topics, our estimated importance of their future impact has a higher subjective element than the other

more established research directions outlined in the previous segments.

Generation

Given our emphasis on embodied agents that are able to interact with their environment, the ability to generate appropriate affective expressions becomes a key prerequisite for the next generation of AC [3]. The recent developments in generative models have shown the remarkable aptitude of AI to use simple prompts to create stunning images [138–141], 3D models [142], videos [143], and audio [144,145]—feats marking the beginning of a new era in AI generation.

Spearheading those advances are diffusion models. Relying on breakthrough work from a theoretical perspective [139,146] and improvement from empirical application and optimization [141], diffusion-based models have already beaten the state-of-the-art generative adversarial networks (GANs) on image synthesis [140] and have also been shown to be more stable and easier to train while generating high-quality images with the desired properties. The original diffusion model was first proposed by Sohl-Dickstein et al. [138] and was inspired by non-equilibrium thermodynamics and probabilistic methods such as Markov chains. It is based on the idea that a real data distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x})$ can be gradually transformed into a simple isotropic Gaussian distribution $\boldsymbol{x}_T \sim \mathcal{N}(\mu,\sigma)$ by iteratively adding a small amount of Gaussian noise to the sample in T steps, and that this forward process can be reversed by sampling from the reverse diffusion process $q(x_{t-1}|x_t)$. However, the reverse diffusion process $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ requires access to the entire dataset for accurate estimation. Therefore, neural networks can be utilized to perform modeling of a series of noise distributions (denoising diffusion probabilistic modeling) [139] or to perform modeling of the gradient of the log probability density function, also known as the (Stein) score function (score matching with Langevin dynamics) [147], to effectively approximate the reverse diffusion process and generate high-quality samples.

Given these exciting advances in related fields, we are just now beginning to scratch the surface of affect generation. Recent works are predominantly occupied with "transmitting the message": generating high-fidelity samples suitable to the "prompt" given by a human user. Yet, to facilitate more natural interactions, we require agents that are able to go beyond that and decide for themselves the appropriate response. Agents thus need an underlying decision process that assigns a probability distribution to a limited set of "actions" that can be taken given system "states." Inspiration for such a process can be drawn from the RL literature, which describes the decision process as a policy. RL is commonly described as an agent-based framework that has to learn the policy based on an unfolding history of state–action pairs that are assigned a reward based on the impact of the actions according to the agent's goals. RL is applied in a plethora of dialogue systems [148] to generate adequate responses, e.g., politeness [149], empathy [150], or emotionality [151], and can be co-opted to create affective agents that can decide for themselves the most suitable affective response.

Personalization

In practice, we encounter individual differences among different subjects. These differences can be reflected in several data modalities. Individual characteristics can manifest themselves in the audio modality in the form of the speech and voice used (e.g., higher/lower "normal" fundamental frequency), in the video modality, e.g., on the basis of facial expressions and gestures, and in wearable sensors as general physical condition and individual movement behavior. Nevertheless, most of the models used in research today pay too little attention to these individual differences and are based on a "one model fits all" approach. To cope with these personalized characteristics, there is a need for models that are not only trained on an entire population but also tailored to one specific subject.

Personalization is an area of research that has become increasingly important in recent years and is based on a fundamental difference compared to population-level models: In many cases, longitudinal data on individuals are available, e.g., from patients undergoing medical treatments or users of an app that tracks emotions. These data are not independent and identically distributed; therefore, their potential is not fully exploited by general models. Thus, personalized models are needed to adapt to individual particularities.

In recent years, several major categories of personalization methods have emerged. Each of these groups has strengths and weaknesses and can be distinguished by the amount of data that is necessary for implementing them, ranging from zero-shot personalization to fine-tuning on longitudinal data. Thus, depending on the use case, different methods may be more suitable.

1. Similarity-based approaches are based on the assumption that a model can be tailored to one person even if no longitudinal data are available and can therefore also be referred to as user-independent [152,153]. Sridhar and Busso [154] recently introduced a method of this category for personalizing an SER system using an unsupervised approach. In their approach, an adaptation set is formed, consisting of the speakers in the training set with acoustic patterns most similar to those of the subject in the test set. Subsequently, this adaptation set is utilized for fine-tuning the SER system, resulting in improved performance, even if the speaker in the test set was never present in the training set. Another group of methods within this category is based on dividing all subjects into smaller subgroups. For example, Kathan et al. [24,25] experimented with sex-based group-level model adaptation for personalized depression forecasting and exertion prediction. Dividing people into subgroups, followed by a separate training for each group, enables the models to learn specific characteristics, which would not be possible for models trained on an entire population.

2. Enrolment-based methods also belong to the group of user-independent approaches as they do not need any longitudinal data. Instead, only a limited number of samples and labels are used for adapting to a new user. In contrast to similarity-based methods, this approach offers the advantage that real data from one subject (not just data from similar subjects) are taken into account to tailor the model to a specific user without the need for long-term data. For example, Triantafyllopoulos et al. [155] introduced a framework for SER that is capable of adapting to a new user using a single neutral utterance, which makes this personalization method suitable even for relatively short interactions.

3. User-specific approaches, unlike the other 2 types, which focus on the need for little to no data, exploit the full potential of longitudinal data by using some of it to customize models. This group therefore represents user-dependent methods,

which means that they can only be applied if several data points of a user are available. A popular method of user-specific personalization is to use a common backbone model that is trained on an entire population and then enriched with personalized layers for each subject [51,53,156,157]. Another user-dependent personalization method consists of using subject-dependent feature normalization [52,158]. For example, Busso et al. [158] proposed a normalization approach for speaker adaptation in speech applications.

Despite recent progress, there are also several open challenges. One of the hardest to face is data privacy, which of course conflicts with personalized models. However, federated learning (FL) presents a promising approach to counteract this concern. FL employs a decentralized approach. Following this paradigm, models are collaboratively trained on data that never leaves the device on which it was collected (e.g., a smartphone or smartwatch). In doing so, methods in all of the categories described above can be incorporated into FL (e.g., similarity-based methods such as clustering or user-specific ones such as a common backbone model in combination with personalized layers) [159,160].

In addition to personalizing the recognition of affect, a similar approach can be taken to the modeling of the receiver. Humans not only generate but also perceive affect differently, and a system must accordingly adapt to these differences. On the one hand, this can help the system to "see the world from the user's side": An AC system might be required to have the same understanding of affect as its user. Beyond that, modeling how individuals understand affect is crucial in the generation phase. So far, comparatively little effort has been placed on the personalized generation of affect in order to match the expectations of the receiver, but, given individual differences in perception, this is a key open issue for bringing AC to real-world applications.

Ultimately, personalized machine learning is expected to play a central role in the next generation of AC, given its unique ability to adapt to new users as they interact with an affective agent, thus embedding them more deeply into the context of the current interaction. As such, personalization is a key enabling technology for bridging a crucial gap in the performance of contemporary systems: the gap arising from the individual idiosyncrasies in the expression, and understanding, of affect.

Causality

We end our perspective on exciting new research directions with a discussion of causality. In addition to being one the most fascinating theoretical innovations in recent times—leading to the 2011 Turing Award—it can also act as the "binding sauce" that can bring together several of the research areas we outlined above.

Pearl's "Ladder of Causation" describes 3 levels of causal ability [54]: (a) association, defined as the ability to reason from observations; (b) intervention, denoting the skill to intervene in results and reason about potential outcomes; and (c) counterfactuals, encompassing the facilities of imagination, retrospection, and understanding, which enable reasoning about alternative outcomes that have not yet been observed.

The current generation of AI methods already excels at reasoning from (passive) observations. Indeed, DL has been the cornerstone of recent advances and has vastly improved the predictive performance of human states. However, the next 2 levels of causality require different approaches. For an agent to successfully perform an "intervention," which in the case of AC is tantamount to generating a response that alters the affective state of the interlocutor, it is necessary to distill its knowledge of affect in a set of possible causes and manipulate one or more of them at a time. For example, an agent should learn that a positive reaction to good news is different from a person reacting positively to a humorous interaction with their digital assistant; failing to do so might result in an affective agent that floods the user with a constant stream of (irrelevant) good news to achieve its goal of eliciting a positive response.

Similarly, reasoning about counterfactuals requires taking yet another step up Pearl's ladder. Being able to systematically evaluate "alternative realities" is a crucial component for an affective agent that needs to adjust its strategy on the fly. This entails being able to answer "what if" questions about the past, such as "What if I had greeted the user more formally?", or, crucially, about the future ("What if I now said X?").

The formal tool for performing causal reasoning is the do-calculus. This axiomatic system allows for performing all 3 types of reasoning. Do-calculus has matured rapidly in past years and is used widely in different fields. However, it requires the use of a graph model (or a structural equation model) representing the different variables and their associations. In general, providing such a model is a hard and domain-specific problem and has been receiving increasing interest from the community. Recent efforts have focused on learning a causal model through interactions with the environment, while GNNs have shown great promise in representing them. These causal representations are learned iteratively, with an agent using its prior knowledge to drive future interactions, while in turn learning from them and updating its causal model of the world. Thus, causal reasoning becomes the missing link that allows embodied, context-informed agents to interact with their environment and learn effective strategies for generating appropriate affective responses.

Overall, we see this research field playing a crucial role in future AC systems. We expect it to not only result in more generalizable performance but also imbue systems with the necessary causal reasoning capabilities that will allow them to become more transparent to downstream users while better accounting for the changes in their environment.

A Blueprint for the Next Generation of AC

We have discussed various strategies for moving beyond DL and creating affective agents that can satisfy contemporary needs in an ethical fashion. We end with a discussion of how the different components presented in the previous sections can interact with one another and work together to jointly tackle the challenges in AC research (Fig. 2). In doing so, we will work on the assumption that the desired goal is to build an embodied affective agent (embodied in the sense that it exists in a physical or digital environment with which it is relatively free to interact) whose goal is to communicate with different, potentially multiple human users in a broad context and help them to achieve their goals. Therefore, it needs to understand the needs of different individuals (contrary, for example, to a personal assistant who only ever interacts with one person). A prototypical example would be a robot assistant permanently situated in a public space (e.g., a hospital).

When starting each new interaction with a user or group of users, the first critical requirement is for the agent to assess (a) their goal(s), mostly in terms of semantics, (b) their mental states, and (c) their interrelationships. Assessing goals falls primarily under the auspices of natural language understanding, although the process can be supplemented by an accurate characterization of mental states (e.g., by imparting a sense of urgency). Gauging those mental states is already working at a sufficient level following the advancements sparked by DL.

The advances we have described are mostly geared toward facilitating longer interactions. As the agent converses with its interlocutors, it can adapt to their individual characteristics using personalization, follow the conversation by accounting for multiparty relations, and emanate an appropriate set of social signals befitting the situation (i.e., accomplish generation). Importantly, deciding what signal to procure and when requires an understanding of both context and causation so that a response that is conducive to the agent's goal can be selected. Context, as well as the flow of conversation, can be represented in graphs, which allow for symbolic manipulation according to the identified causal factors driving the interaction. Crucially, these processes need to be implemented on the fly with minimal latency and low energy requirements to minimize the off-time of the agent, something that can be achieved with next-generation spiking neural networks.

In total, the 9 components we have presented in our overview can be used in the following order:

1. Graphs allow for mapping user–user relations and external context to an interpretable representation.
2. Capsules can further facilitate the modeling of part-whole hierarchies, which drive the understanding of an affective interaction in terms of a set of "affective primitives."
3. Manipulating these primitives with a (neuro)symbolic engine allows the agent to reason about counterfactuals and plan its response.
4. Symbols further allow for the specification of common knowledge, rules, or human feedback, which constrain the interaction.
5. Embodiment provides an indirect path to learning, via the pursuit of objectives that necessitate the understanding and portrayal of affect (e.g., collaborative learning) in a constrained (digital or physical) environment.
6. Personalization is foundational to adapting to user characteristics and thus moving away from a "one size fits all" solution toward more adaptive agents.
7. The advances in generative AI across multiple modalities can then be tasked with creating the agent's response.
8. Causal models enable the principled disentangling of causes from effects and facilitate higher-order reasoning.
9. Finally, spiking neural networks have shown great promise in enabling the deployment of DNNs in physical agents with limited computational resources.

Limitations

Our review of recent advances in AI research that can catalyze the deployment of affective agents in the real world has an unavoidable element of subjectivity. AI is a rapidly growing field that has been drawing tremendous academic and industrial interest in the last decade. Therefore, it is impossible to map out all advances that may be relevant for the AC community. For example, we have decided to omit the exciting advances taking place in the use of large language models (LLMs) [161], as their major improvement over previous models is an increase in their depth. Nevertheless, we expect the areas we have chosen to be among those that play a vital role in the upcoming years.

Conclusion

We have presented an overview of 9 (re-)emerging themes in AI research that seem poised to play a pivotal role in AC. Our starting point was the limitations of DL—the driving force behind much of the recent progress. The success of DL has opened up new paths toward making human–computer interaction more affective. Indeed, the overwhelming effectiveness of DL has caused many to claim that it is enough to achieve the goals of the community. While this may well prove true, it is important not to overlook additional complementary research directions. Our article serves as a primer on some of the most prominent areas in the broader AI field.

Acknowledgments

Funding: This work was partially funded by EU H2020 project no. 101135556 (INDUX-R).
Author contributions: B.W.S. proposed the current contribution and outlined the general premise. A.T. conceptualized the structure of the manuscript, identified the areas to be reviewed, and coordinated the initial draft. S.A. and B.W.S. assisted with the structuring of the manuscript and the areas to be reviewed. A.T., L.C., A.G., X.J., A.K., M.M., and I.T. conducted literature reviews on individual sections. All authors read and contributed to the final manuscript.
Competing interests: The authors declare that they have no competing interests.

References

1. Batliner A, Hantke S, Schuller B. Ethics and good practice in computational paralinguistics. *IEEE Trans Affect Comput*. 2020;13(3):1236–1253.
2. Wang Y, Song W, Tao W, Liotta A, Yang D, Li X, Gao S, Sun Y, Ge W, Zhang W, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf Fusion*. 2022;83:19–52.
3. Triantafyllopoulos A, Schuller BW, İymen G, Sezgin M, He X, Yang Z, Tzirakis P, Liu S, Mertes S, André E, et al. An overview of affective speech synthesis and conversion in the deep learning era. *Proc IEEE*. 2023;111(10):1355–1381.
4. Khalil RA, Jones E, Babar MI, Jan T, Zafar MH, Alhussain T. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*. 2019;7:117327–117345.
5. Canal FZ, Müller TR, Matias JC, Scotton GG, de Sa Junior AR, Pozzebon E, Sobieranski AC. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Inf Sci*. 2022;582:593–617.
6. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf Fusion*. 2017;37:98–125.
7. Schuller B, Batliner A. *Computational paralinguistics: Emotion, affect and personality in speech and language processing*. Hoboken (NJ): John Wiley & Sons; 2013.

8. Scherer KR. Vocal communication of emotion: A review of research paradigms. *Speech Comm*. 2003;40(1–2):227–256.
9. Scherer KR. Psychological models of emotion. In: *The neuropsychology of emotion*. Oxford (UK): Oxford Univ. Press; 2000. p. 137–162.
10. Ekman P. An argument for basic emotions. *Cognit Emot*. 1992;6(3–4):169–200.
11. Russell JA. A circumplex model of affect. *J Pers Soc Psychol*. 1980;39(6):1161–1178.
12. Scherer KR. Towards a prediction and data driven computational process model of emotion. *IEEE Trans Affect Comput*. 2021;12(2):279–292.
13. Scherer KR, Emotion and emotional competence: Conceptual and theoretical issues for modelling agents. In: *Blueprint for affective computing: A sourcebook*. Oxford (UK): Oxford Univ. Press; 2010. p. 3–20.
14. Adam C, Johal W, Pellier D, Fiorino H, and Pesty S. Social human-robot interaction: A new cognitive and affective interaction-oriented architecture. In: *International conference on social robotics*. New York City (NY): Springer; 2016. p. 253–263.
15. Sanchez Y, Coma T, Aguelo A, Cerezo E. ABC-EBDI: An affective framework for BDI agents. *Cogn Syst Res*. 2019;58:195–216.
16. Hoorn JF, Baier T, Van Maanen JA, Wester J. Silicon Coppélia and the formalization of the affective process. *IEEE Trans Affect Comput*. 2021;14(1):255–278.
17. Zall R, Kangavari MR. Comparative analytical survey on cognitive agents with emotional intelligence. *Cogn Comput*. 2022;14:1223–1246.
18. Goldberg LR. An alternative "description of personality": The big-five factor structure. *J Pers Soc Psychol*. 1990;59(6): 1216–1229.
19. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23(1):56–62.
20. Verma B, Choudhary A. Affective state recognition from hand gestures and facial expressions using Grassmann manifolds. *Multimed Tools Appl*. 2021;80:14019–14040.
21. Ben X, Ren Y, Zhang J, Wang S-J, Kpalma K, Meng W, Liu Y-J. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Trans Pattern Anal Mach Intell*. 2021;44(9):5826–5846.
22. Alswaidan N, Menai MEB. A survey of state-of-the-art approaches for emotion recognition in text. *Knowl Inf Syst*. 2020;62:2937–2987.
23. Kim KH, Bang SW, Kim SR. Emotion recognition system using short-term monitoring of physiological signals. *Med Biol Eng Comput*. 2004;42:419–427.
24. Kathan A, Triantafyllopoulos A, Amiriparian S, Gebhard A, Ottl S, Gerczuk M, Jaumann M, Hildner D, Dieter V, Schneeweiss P, et al. Investigating individual-and group-level model adaptation for self-reported runner exertion prediction from biomechanics. In: *Proceedings of the E-Health and Bioengineering Conference (EHB)*. Iași (Romania): IEEE; 2022. p. 1–4.
25. Kathan A, Harrer M, Küster L, Triantafyllopoulos A, He X, Milling M, Gerczuk M, Yan T, Rajamani ST, Heber E, et al. Personalised depression forecasting using mobile sensor data and ecological momentary assessment. *Front Digit Health*. 2022;4:964582.
26. Gunes H, Pantic M. Automatic, dimensional and continuous emotion recognition. *Int J Synth Emot*. 2010;1(1):68–99.
27. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
28. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Paper presented at: Advances in Neural Information Processing Systems 25 (NIPS 2012); 2012; Lake Tahoe, USA.
29. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: A brief review. *Comput Intell Neurosci*. 2018;2018:7068349.
30. Graves A, Mohamed Ar, Hinton G. Speech recognition with deep recurrent neural networks. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver (Canada): IEEE; 2013. p. 6645–6649.
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. *Adv Neural Inf Proces Syst*. 2017;30:5998–6008.
32. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, Zafeiriou S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai (China): IEEE; 2016. p. 5200–5204.
33. Wagner J, Triantafyllopoulos A, Wierstorf H, Schmitt M, Burkhardt F, Eyben F, Schuller BW. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Trans Pattern Anal Mach Intell*. 2022;45(9):10745–10759.
34. Matsugu M, Mori K, Mitari Y, Kaneda Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Netw*. 2003;16(5–6): 555–559.
35. Wöllmer M, Kaiser M, Eyben F, Schuller B, Rigoll G. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image Vis Comput*. 2013;31(2): 153–163.
36. Qian Y, Bi M, Tan T, Yu K. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Trans Audio Speech Lang Process*. 2016;24(12):2263–2276.
37. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv. 2014. https://doi.org/10.48550/arXiv.1312.6199.
38. Le Q and Mikolov T. Distributed representations of sentences and documents. In: *International Conference on Machine Learning*. PMLR; 2014. p. 1188–1196.
39. Liu S, Mallol-Ragolta A, Parada-Cabaleiro E, Qian K, Jing X, Kathan A, Hu B, Schuller BW. Audio self-supervised learning: A survey. *Patterns*. 2022;3(12):100616.
40. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. *Adv Neural Inf Proces Syst*. 2020;33:1877–1901.
41. Amiriparian S, Gerczuk M, Ottl S, Cummins N, Freitag M, Pugachevskiy S, Baird A, Schuller B. Snore sound classification using image-based deep spectrum features. Paper presented at: INTERSPEECH 2017; 2017 Aug 20–24; Stockholm, Sweden.
42. Gong Y, Chung YA, Glass J. AST: Audio spectrogram transformer. arXiv. 2021. https://doi.org/10.48550/arXiv.2104.01778.
43. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell*. 2022;45(1):87–110.
44. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, Arx S von, Bernstein MS, Bohg J, Bosselut A, Brunskill E, et al.

On the opportunities and risks of foundation models. arXiv. 2021. https://doi.org/10.48550/arXiv.2108.07258.
45. Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, Dorundo E, Desai R, Zhu T, Parajuli S, Guo M, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. Paper presented at: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021.
46. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern Recognit*. 2012;45(2):521–530.
47. Kossaifi J, Walecki R, Panagakis Y, Shen J, Schmitt M, Ringeval F, Han J, Pandit V, Toisoul A, Schuller B, et al. Sewa DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans Pattern Anal Mach Intell*. 2019;43(3):1022–1040.
48. Baird A, Tzirakis P, Brooks JA, Gregory CB, Schuller B, Batliner A, Keltner D, Cowen A. The ACII 2022 Affective Vocal Bursts Workshop & Competition. In: *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. Nara, Japan: IEEE; 2022. p. 1–5.
49. Cowen A, Prasad G, Tanaka M, et al. How emotion is experienced and expressed in multiple cultures: A large-scale experiment. PsyArXiv. 2021. doi:10.31234/osf.
50. Christ L, Amiriparian S, Baird A, Tzirakis P, Kathan A, Müller N, Stappen L, Meßner E-M, König A, Cowen A, et al. *The MuSe 2022 multimodal sentiment analysis challenge: Humor, emotional reactions, and stress*. Lisboa (Portugal): ACM; 2022. p. 5–14.
51. Song M, Triantafyllopoulos A, Yang Z, Takeuchi H, Nakamura T, Kishi A, Ishizawa T, Yoshiuchi K, Jing X, Karas V, et al. Daily mental health monitoring from speech: A real-world Japanese dataset and multitask learning analysis. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodos (Greece): IEEE; 2023. p. 1–5.
52. Gerczuk M, Triantafyllopoulos A, Amiriparian S, Kathan A, Bauer J, Berking M, Schuller B. Personalised deep learning for monitoring depressed mood from speech. In: *Proceedings of the E-Health and Bioengineering Conference (EHB)*. Iași (Romania): IEEE; 2022. p. 1–5.
53. Rudovic O, Lee J, Dai M, Schuller B, Picard RW. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Sci Robot*. 2018;3(19):eaao6760.
54. Pearl J. *Causality.* Cambridge (UK): Cambridge Univ. Press; 2009.
55. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Proces Syst*. 2017;30.
56. Samek W, Binder A, Montavon G, Lapuschkin S, Müller K-R. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Netw Learn Syst*. 2016;28:2660–2673.
57. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R. Layer-wise relevance propagation: An overview. In: *Explainable AI: Interpreting, explaining and visualizing deep learning*. New York City (NY): Springer; 2019. p. 193–209.
58. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. Paper presented at: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, CA, USA.
59. Blalock D, Gonzalez Ortiz JJ, Frankle J, Guttag J. What is the state of neural network pruning? *Proc Mach Learn Syst*. 2020;2:129–146.
60. Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. *IEEE Trans Neural Netw Learn Syst*. 2020;32(2):604–624.
61. Patrick MK, Adekoya AF, Mighty AA, Edward BY. Capsule networks—A survey. *J King Saud Univ Comput Inf Sci*. 2022;34(1):1295–1310.
62. Hinton GE, Krizhevsky A, Wang SD. Transforming auto-encoders. In: *Artificial Neural Networks and Machine Learning—ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*. New York City (NY): Springer; 2011. p. 44–51.
63. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. *Adv Neural Inf Proces Syst*. 2017;30.
64. Hinton GE, Sabour S, Frosst N. Matrix capsules with EM routing. In: *International Conference on Learning Representations*. Vancouver (Canada): PMLR; (2018).
65. Amiriparian S, Awad A, Gerczuk M, Stappen L, Baird A, Ottl S, Schuller B. Audio-based recognition of bipolar disorder utilising capsule networks. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE; 2019. p. 1–7.
66. Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A. COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from x-ray images. *Pattern Recogn Lett*. 2020;138:638–643.
67. Gupta P, Siddiqui MK, Huang X, Morales-Menendez R, Panwar H, Terashima-Marin H, Wajid MS. COVID-WideNet–a capsule network for COVID-19 detection. *Appl Soft Comput*. 2022;122:Article 108780.
68. Iesmantas T, Alzbutas R. Convolutional capsule network for classification of breast cancer histology images. In: *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*. New York City (NY): Springer; 2018. p. 853–860.
69. Zhang X, Zhao S-G. Fluorescence microscopy image classification of 2D HeLa cells based on the CapsNet neural network. *Med Biol Eng Comput*. 2019;57(6):1187–1198.
70. Kim Y, Wang P, Zhu Y, Mihaylova L. A capsule network for traffic speed prediction in complex road networks. In: *2018 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. New York City (NY): IEEE; 2018. p. 1–6.
71. Kim M, Chi S. Detection of centerline crossing in abnormal driving using CapsNet. *J Supercomput*. 2019;75:189–196.
72. Renkens V, van Hamme H. Capsule networks for low resource spoken language understanding. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA; 2018.
73. Ren H, Lu H. Compositional coding capsule network with k-means routing for text clasification. *Pattern Recogn Lett*. 2022;160:1–8.
74. Liu Y, Ding Y, Li C, Cheng J, Song R, Wan F, Chen X. Multi-channel EEG-based emotion recognition via a multi-level features guided capsule network. *Comput Biol Med*. 2020;123:Article 103927.
75. Li C, Wang B, Zhang S, Liu Y, Song R, Cheng J, Chen X. Emotion recognition from EEG based on multi-task learning with capsule network and attention mechanism. *Comput Biol Med*. 2022;143:Article 105303.
76. Wu X, Liu S, Cao Y, et al. Speech emotion recognition using capsule networks. In: *ICASSP 2019-2019 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP). New York City (NY): IEEE; 2019. p. 6695–6699.
77. Wu X, Cao Y, Lu H, Liu S, Wang D, Wu Z, Liu X, Meng H. Speech emotion recognition using sequential capsule networks. *IEEE/ACM Trans Audio Speech Lang Process*. 2021;29:3280–3291.
78. Shahin I, Hindawi N, Nassif AB, Alhudhaif A, Polat K. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Syst Appl*. 2022;188:Article 116080.
79. Hosseini S and Cho NI. GF-CapsNet: Using gabor jet and capsule networks for facial age, gender, and expression recognition. In: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. New York City (NY): IEEE; 2019. p. 1–8.
80. Li D, Zhao X, Yuan G, Liu Y, Liu G. Robustness comparison between the capsule network and the convolutional network for facial expression recognition. *Appl Intell*. 2021;51: 2269–2278.
81. Hinton G. Some demonstrations of the effects of structural descriptions in mental imagery. *Cogn Sci*. 1979;3:231–250.
82. Hinton GE. Mapping part-whole hierarchies into connectionist networks. *Artif Intell*. 1990;46:47–75.
83. Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Process Mag*. 2017;34:18–42.
84. Bronstein MM, Bruna J, Cohen T, Veličković P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv. 2021. https://doi.org/10.48550/arXiv.2104.13478.
85. Hanocka R, Hertz A, Fish N, Giryes R, Fleishman S, Cohen-Or D. Meshcnn: A network with an edge. *ACM Trans Graph*. 2019;38:1–12.
86. Maron H, Ben-Hamu H, Shamir N, Lipman Y. Invariant and equivariant graph networks. arXiv. 2018. https://doi.org/10.48550/arXiv.1812.09902.
87. Bulusu S, Favoni M, Ipp A, Müller DI, Schuh D. Equivariance and generalization in neural networks. In: *EPJ Web of Conferences*. Les Ulls (France): EDP Sciences; 2022. p. 09001.
88. Favoni M, Ipp A, Müller DI, Schuh D. Lattice gauge symmetry in neural networks. arXiv. 2021. https://doi.org/10.48550/arXiv.2111.04389.
89. Cao W, Yan Z, He Z, He Z. A comprehensive survey on geometric deep learning. *IEEE Access*. 2020;8:35929–35949.
90. Benton G, Finzi M, Izmailov P, Wilson AG. Learning invariances in neural networks from training data. *Adv Neural Inf Proces Syst*. 2020;33:17605–17616.
91. Wu YX, Wang X, Zhang A, He X, Chua TS. Discovering invariant rationales for graph neural networks. arXiv. 2022. https://doi.org/10.48550/arXiv.2201.12872.
92. Bhagat S, Cormode G, Muthukrishnan S. Node classification in social networks. In: *Social network data analytics*. New York City (NY): Springer; 2011.
93. Dai M, Demirel MF, Liang Y, Hu J-M. Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials. *npj Computational Materials*. 2021;7:103.
94. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv. 2016. https://doi.org/10.48550/arXiv.1609.02907.
95. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv. 2017. https://doi.org/10.48550/arXiv.1710.10903.
96. Chen D, Lin Y, Li W, Li P, Zhou J, Sun X. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. New York (NY): AAAI; 2020. p. 3438–3445.
97. Liu M, Gao H, Ji S. Towards deeper graph neural networks. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York City (NY): ACM; 2020. p. 338–348.
98. Loukas A. What graph neural networks cannot learn: Depth vs width. arXiv. 2019. https://doi.org/10.48550/arXiv.1907.03199.
99. Liao W, Zeng B, Liu J, Wei P, Cheng X, Zhang W. Multi-level graph neural network for text sentiment analysis. *Comput Electr Eng*. 2021;92:Article 107096.
100. Wang M, Hu G. A novel method for twitter sentiment analysis based on attentional-graph neural network. *Information*. 2020;11(2):92.
101. Song T, Chen L, Zheng W, Ji Q. Uncertain graph neural networks for facial action unit detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Washington (DC): AAAI; 2021. p. 5993–6001.
102. Song T, Cui Z, Zheng W, Ji Q. Hybrid message passing with performance-driven structures for facial action unit detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York City (NY): IEEE; 2021. p. 6267–6276.
103. Wu J, He X, Wang X, Wang Q, Chen W, Lian J, Xie X. Graph convolution machine for context-aware recommender system. *Front Comp Sci*. 2022;16:Article 166614.
104. Qi CR, Su H, Mo K, Guibas LJ. PointNet: Deep learning on point sets for 3d classification and segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New York City (NY): IEEE; 2017. p. 652–660.
105. Monti F, Boscaini D, Masci J, Rodola E, Svoboda J, Bronstein MM. Geometric deep learning on graphs and manifolds using mixture model CNNs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2017. p. 5115–5124.
106. Cao W, Zheng C, Yan Z, Xie W. Geometric deep learning: Progress, applications and challenges. *Science China Inf Sci*. 2022;65:Article 126101.
107. Valem LP, Pedronette DCG, Latecki LJ. Graph convolutional networks based on manifold learning for semi-supervised image classification. *Comput Vis Image Underst*. 2023;227:Article 103618.
108. Weber R, Barrielle V, Soladié C, Séguier R. Unsupervised adaptation of a person-specific manifold of facial expressions. *IEEE Trans Affect Comput*. 2018;11:419–432.
109. Wang B, Dong G, Zhao Y, Li R, Yang H, Yin W, Liang L. *Spiking emotions: Dynamic vision emotion recognition using spiking neural networks*. Aachen (Germany): AHPCAI; 2022.
110. Tan C, Šarlija M, Kasabov N. Neurosense: Short-term emotion recognition and understanding based on spiking neural network modelling of spatio-temporal EEG patterns. *Neurocomputing*. 2021;434:137–148.
111. Tan C, Ceballos G, Kasabov N, Puthanmadam Subramaniyam N. FusionSense: Emotion classification using feature fusion of multimodal data and deep learning in a brain-inspired spiking neural network. *Sensors*. 2020;20(18):5328.
112. Ghosh-Dastidar S, Adeli H. Spiking neural networks. *Int J Neural Syst*. 2009;19:295–308.

113. Yamazaki K, Vo-Ho V-K, Bulsara D, Le N. Spiking neural networks and their applications: A review. *Brain Sci*. 2022;12(7):863.
114. Gerstner W, Kistler WM. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge (UK): Cambridge Univ. Press; 2002.
115. Lee JH, Delbruck T, Pfeiffer M. Training deep spiking neural networks using backpropagation. *Front Neurosci*. 2016;10:508.
116. Bohte SM, Kok JN, La Poutré JA. SpikeProp: Backpropagation for networks of spiking neurons. In: *ESANN*. Bruges (Belgium); 2000. p. 419–424.
117. Aimone JB, Ho Y, Parekh O, Phillips CA, Pinar A, Severa W, Wang Y. Provable advantages for graph algorithms in spiking neural networks. In: *Proc. ACM Symposium on Parallelism in Algorithms and Architectures*. New York City (NY): ACM; 2021. p. 35–47.
118. Sarker MK, Zhou L, Eberhart A, Hitzler P. Neuro-symbolic artificial intelligence. *AI Commun*. 2021;34(3):197–209.
119. d' Avila Garcez A, Lamb LC. Neurosymbolic AI: The 3rd wave. *Artif Intell Rev*. 2023:1–20.
120. Kautz H. The third AI summer: AAAI Robert S. Engelmore memorial lecture. *AI Mag*. 2022;43:105–125.
121. Li Y, Su H, Shen X, Li W, Cao Z, Niu S. DailyDialog: A manually labelled multi-turn dialogue dataset. arXiv. 2017. https://doi.org/10.48550/arXiv.1710.03957.
122. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. arXiv. 2019. https://doi.org/10.48550/arXiv.1810.02508.
123. Hazarika D, Poria S, Mihalcea R, Cambria E, Zimmermann R. ICON: Interactive conversational memory network for multimodal emotion detection. In: *Proc. EMNLP*. Brussels (Belgium): ACL; 2018. p. 2594–2604.
124. Wang Y, Zhang J, Ma J, Wang S, and Xiao J. Contextualized emotion recognition in conversation as sequence tagging. In: *Proc. SIGDIAL*. Virtual: ACL; 2020. p. 186–195.
125. Tu G, Wen J, Liu C, Jiang D, Cambria E. Context- and sentiment-aware networks for emotion recognition in conversation. *IEEE Trans Artif Intell*. 2022;3:699–708.
126. Sap M, Bras RL, Allaway E, Bhagavatula C, Lourie N, Rashkin H, Roof B, Smith NA, Choi Y. ATOMIC: An atlas of machine commonsense for if-then reasoning. In: *Proc. AAAI Conference on Artificial Intelligence*. Honolulu (HI): AAAI Press; 2019. p. 3027–3035.
127. Ghosal D, Shen S, Majumder N, Mihalcea R, Poria S. CICERO: A dataset for contextualized commonsense inference in dialogues. In: *Proc. ACL*. Dublin (Ireland): ACL; 2022. p. 5010–5028.
128. Ghosal D, Majumder N, Gelbukh A, Mihalcea R, Poria S. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In: *Findings of the ACL: EMNLP 2020*. Virtual: ACL; 2020. p. 2470–2481.
129. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J Comput Phys*. 2019;378:686–707.
130. Anderson ML. Embodied cognition: A field guide. *Artif Intell*. 2003;149:91–130.
131. Chrisley R. Embodied artificial intelligence. *Artif Intell*. 2003;149:131–150.
132. Shapiro L. *The Routledge handbook of embodied cognition*. New York (NY): Routledge; 2014.
133. Semin GR, Smith ER. Interfaces of social psychology with situated and embodied cognition. *Cogn Syst Res*. 2002;3(3):385–396.
134. Niedenthal PM. Embodying emotion. *Science*. 2007;316:1002–1005.
135. Barrett LF and Lindquist KA. The embodiment of emotion. Embodied grounding: Social, cognitive, affective, and neuroscientific approaches. Cambridge (UK): Cambridge University Press; 2008. p. 237–262.
136. Savva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B, Straub J, Liu J, Koltun V, Malik J, et al. Habitat: A platform for embodied AI research. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. New York City (NY): IEEE; 2019. p. 9339–9347.
137. Kempka M, Wydmuch M, Runc G, Toczek J, Jaśkowski W. Vizdoom: A doom-based AI research platform for visual reinforcement learning. In: *Proc. IEEE Conference on Computational Intelligence and Games (CIG)*. New York City (NY): IEEE; 2016. p. 1–8.
138. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: *International Conference on Machine Learning*. PMLR; 2015. p. 2256–2265.
139. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Proces Syst*. 2020;33:6840–6851.
140. Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. *Adv Neural Inf Proces Syst*. 2021;34:8780–8794.
141. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York City (NY): IEEE; 2022. p. 10684–10695.
142. Zhou L, Du Y, Wu J. 3D shape generation and completion through point-voxel diffusion. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. New York City (NY): IEEE; 2021. p. 5826–5835.
143. Ho J, Chan W, Saharia C, Whang J, Gao R, Gritsenko A, Kingma DP, Poole B, Norouzi M, Fleet DJ, et al. Imagen video: High definition video generation with diffusion models. arXiv. 2022. https://doi.org/10.48550/arXiv.2210.02303.
144. Jeong M, Kim H, Cheon SJ, Choi BJ, Kim NS. Diff-TTS: A denoising diffusion model for text-to-speech. arXiv. 2021. https://doi.org/10.48550/arXiv.2104.01409.
145. Popov V, Vovk I, Gogoryan V, Sadekova T, Kudinov M. Grad-TTS: A diffusion probabilistic model for text-to-speech. In: *International Conference on Machine Learning*. PMLR; 2021. p. 8599–8608.
146. Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. arXiv. 2020. https://doi.org/10.48550/arXiv.2011.13456.
147. Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. *Adv Neural Inf Proces Syst*. 2019;32.
148. Ni J, Young T, Pandelea V, Xue F, Cambria E. Recent advances in deep learning based dialogue systems: A systematic survey. *Artif Intell Rev*. 2022;56:3055–3155.
149. Niu T, Bansal M. Polite dialogue generation without parallel data, *Transactions of the Association for. Comput Linguist*. 2018;6:373–389.

150. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In: *Proc. Web Conference*. Ljubljana (Slovenia): ACM; 2021. p. 194–205.
151. Zhou X, Wang WY. MojiTalk: Generating emotional responses at scale. In: *Proc. ACL*. Melbourne (Australia): Stroudsburg (PA): ACL; 2018. p. 1128–1137.
152. Xu X, Chikersal P, Dutcher JM, Sefidgar YS, Seo W, Tumminia MJ, Villalba DK, Cohen S, Creswell KG, David Creswell JD, et al. Leveraging collaborative-filtering for personalized behavior modeling: A case study of depression detection among college students. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. New York City (NY): ACM; 2021. p. 1–27.
153. Li B, Sano A. Early versus late modality fusion of deep wearable sensor features for personalized prediction of tomorrow's mood, health, and stress. In: *Proc. EMBC*. Montreal (Canada): New York City (NY): IEEE; 2020. p. 5896–5899.
154. Sridhar K, Busso C. Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech. arXiv. 2022. https://doi.org/10.48550/arXiv.2201.07876.
155. Triantafyllopoulos A, Liu S, Schuller BW. Deep speaker conditioning for speech emotion recognition. In: *Proceedings of the International Conference on Multimedia and Expo (ICME)*. Shenzhen (China): IEEE; 2021. p. 1–6.
156. Taylor S, Jaques N, Nosakhare E, Sano A, Picard R. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Trans Affect Comput*. 2017;11(2):200–213.
157. Kathan A, Amiriparian S, Christ L, Triantafyllopoulos A, Müller N, König A, Schuller BW. A personalised approach to audiovisual humour recognition and its individual-level fairness. In: *Proceedings of the 3rd International Multimodal Sentiment Analysis Workshop and Challenge*. Lisboa (Portugal): ACM; 2022. p. 29–36.
158. Busso C, Mariooryad S, Metallinou A, Narayanan S. Iterative feature normalization scheme for automatic emotion detection from speech. *IEEE Trans Affect Comput*. 2013;4(4):386–397.
159. Kulkarni V, Kulkarni M, Pant A. Survey of personalization techniques for federated learning. In: *Proceedings of the Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. London (UK): IEEE; 2020. p. 794–797.
160. Tan AZ, Yu H, Cui L, Yang Q. Towards personalized federated learning. In: *IEEE Transactions on Neural Networks and Learning Systems*. New York City (NY): IEEE; 2022.
161. Amin MM, Cambria E, Schuller BW. Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intell Syst*. 2023;38(2):15–23.