EXPLAINABILITY IN DATA SCIENCE: INTERPRETABILITY, REPRODUCIBILITY, AND REPLICABILITY

Payel Das and Lav R. Varshney

Explaining Artificial Intelligence Generation and Creativity

Human interpretability for novel ideas and artifacts

xxxxxxx

Creativity is often thought of as the pinnacle of human achievement, but artificial intelligence (AI) is now starting to play a central role in creative processes, whether autonomously or in collaboration with people. Widespread deployment is now pushing for explanations on how creative AI is working, whether to engender trust, enable action, provide a basis for evaluation, or for intrinsic reasons. In this article, we review various motivations, algorithms, and methods for explaining either the workings of generative/creative AI algorithms or the generative/creative artifacts they produce.

Introduction

In October 2018, a piece of art, "Edmond de Belamy, from La Famille de Belamy," was sold by Christie's auction house for US$432,500 to an anonymous bidder, hundreds of thousands of dollars more than works by Andy Warhol and Roy Lichtenstein in the same auction. Generated by the French art collective Obvious using a generative adversarial network (GAN), this was putatively the first auctioned artwork generated by AI [1]. In early 2019, the McCormick spice company began widespread sales of the ONE line of seasoning mixes, created using a successor to the IBM Chef Watson culinary creativity system that used stochastic sampling and selection algorithms [2]. By early 2021, de novo AI-generated and experimentally verified antimicrobial peptides were reported in the scientific literature, created using variational autoencoder (VAE) techniques [3], and separately using Transformer-based autoregressive language modeling methods [4]. Image generation from text descriptions, as in OpenAI's Transformer-based DALL-E system from early 2021 [5], is now even said to meet the standards of tests for general creativity [6].

As has been described by the Computer Scientist and Cognitive Scientist Margaret Boden, creativity is often regarded as the pinnacle of intelligence, humanity's crowning glory, and yet, it is not fully understood. Moreover, many regard its unintelligibility as its splendor, not to be sullied by scientific explanation [7]. Due to the widespread deployment of AI technologies in settings involving people, however, there is

Digital Object Identifier 10.1109/MSP.2022.3141365
Date of current version: xxxxxxx

growing interest in explaining their results—whether to engender trust, enable action, provide a basis for evaluation, or for intrinsic reasons [8]. Noting that explanation is infinitely variable and that there can be many valid explanations for given phenomena, Selbst and Barocas have specifically suggested addressing inscrutability and nonintuitiveness. Inscrutability is when AI models available for direct inspection may defy understanding due to their complexity; nonintuitiveness is when AI models are based on statistical relationships that defy human intuition even when they may be understandable.

Much past work in interpretable and explainable AI (XAI) has focused on decisions and predictions, see, e.g., [9]–[11]; here instead, we consider explainability for generative/creative algorithms or for the ideas and artifacts they produce. Interpretability for generative and creative models is critical as they are becoming widely used in many application domains, not only in settings we have noted, but also more broadly in engineering, design, science, and the arts.

Moreover, generative models are being used for nefarious purposes such as deepfakes and fake news. Distinguishing deepfakes/fake news from authentic content becomes more difficult for better generative models, and the state of the field is such that often people cannot tell the difference [12]. Existing hypothesis-testing techniques focus on specific hand-crafted features, domain knowledge, and using machine learning itself to find statistical differences. AI-governance techniques that consider the social positioning of generative AI models [13] may be needed to ensure safety in the face of misinformation. Such governance can be enhanced through a well-documented and explainable understanding of their workings and results.

In artistic domains, explainability is also intertwined with intentionality—a (human) intent, inspiration, or desire to express something [14], [15]. It is said, especially in the Western tradition following Romanticism, that communication of meaning in art is necessary for eliciting an aesthetic experience. For example, considering narration or poetry, (linguistic) meaning is the relationship between a linguistic form and communicative intent, where communicative intents are about things that are outside of language. Communicative intent is distinct from standing meaning, which is constant across all of its possible contexts of use [16]. Recent surveys further indicate that people want not just novelty/quality, but also intentionality and autonomy, to attribute creativity to an artificial system [17].

The specific generative techniques we have described thus far include ones based on deep learning, such as VAEs, GANs, autoregressive models such as Transformers, and normalizing flow models, which, in their original form, are inscrutable and nonintuitive. Later in this article, we discuss posthoc techniques (including visual analytics) that have recently been developed for interpreting the knowledge present within these models, which is used as a part of the generative process and for interpreting the outputs of these kinds of models. The examples include the so-called BERTology approach to interpreting the knowledge in large-scale language models by probing attention mechanisms, and a traveling salesperson path-based explanation of combinatorially creative artifacts, respectively.

Further, we discuss recent extensions of these generative techniques that facilitate interpretability. We also examine emerging generative algorithms that are intrinsically interpretable, such as information lattice learning (ILL), a framework with group- and information-theoretic foundations), disentangled generative models, and computational creativity algorithms based on stochastic sampling and on-case-based reasoning.

Although some of the generative models we evaluate require massive data sets and computational infrastructure (including specialized hardware) for training and creative generation [18], others may work well with tiny data sets of just a few examples [19]. For the ones that do require large-scale data and computations, scalability and speedups can be achieved by using methods such as efficient linear-time attention, adaptive bucketing of batches, open source parallelization, or combinations of these techniques (see, e.g., [20]).

A key part of our synthesis in this review is a discussion of best practices to obtain insights through explainability in generative algorithms. As shown in Figure 1, we believe there can be a virtuous interaction among research in creative AI and XAI, all while interacting with human experts. In particular, this is the case when considering different kinds of creativity—combinatorial (which brings existing ideas together in new ways), exploratory (which expands the conceptual space), or transformational (which requires a completely new representation of conceptual domains) [7]—and when considering a variety of explainability techniques, whether posthoc

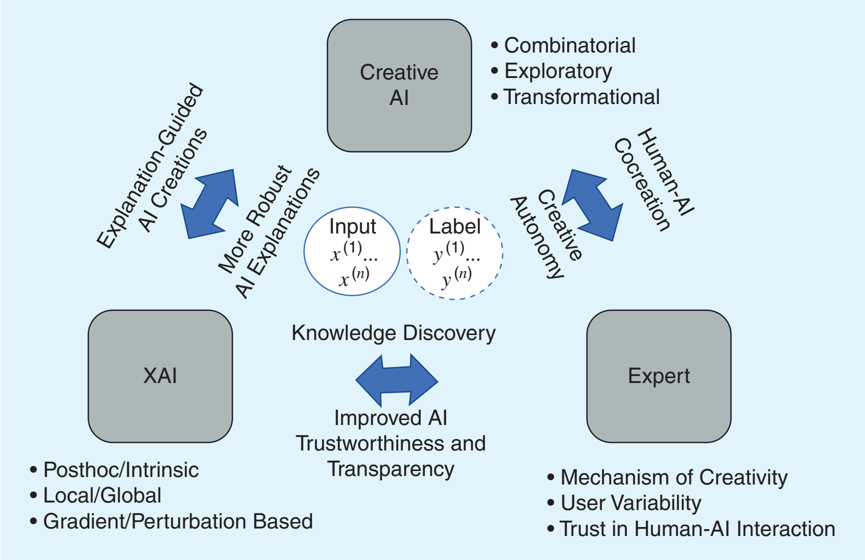


FIGURE 1. The expected improvements due to pursuing research at the intersection of creative AI, XAI, and human–AI interaction.

or intrinsic. Notably, one may develop explanation-guided AI creations as well as more robust AI explanations. Moreover, when interacting with human experts, the result may be better human–AI creative collaborations as well as better human-interpretable and trustworthy knowledge discovery.

Generative and Creative AI

We have given examples thus far of generative and creative AI in visual art, culinary, and protein engineering, but the application domains are seemingly limitless. One can certainly think about developing AI to operate in either autonomous or semi-autonomous mode with human interaction in developing solutions for a broad set of application domains. Systems already exist for application areas that range from molecule design beyond proteins [21]; fragrance formulation design [22]; materials discovery that yields sustainable building materials that are stronger, yet with much less embodied carbon than traditional formulations of concrete [23]; fashion designs that have been worn to the gala at the Metropolitan Museum of Art; artificial weather data generators that can be used to predict the impact of climate change [24]; musical compositions that have been performed on the world's top stages [25]; and language including narrative, poetry, computer programs, and news reports.

Drawing largely on the psychological study of creativity, several ways of assessing creative AI technologies have emerged [26]. Notably, some evaluations are focused on the process of creativity, whereas other evaluations are focused on the product of creativity, such as ideas or artifacts. These two broad notions of how creative systems are evaluated also imply two broad notions of how creative systems should be explainable. Should the entire generative process be explained to people, or should only the final result of the generative process be explained?

In psychology, the definition of creativity also has at least two dimensions. An idea or artifact is said to be creative if it is judged to be novel and also appropriate, useful, or valuable by a knowledgeable social group. These two dimensions of creativity also clarify the distinction between generative AI and AI that is specifically creative. In particular, creative AI is a special case of generative AI where novelty is explicitly important. Many generative AI algorithms are interpolative rather than extrapolative in conceptual spaces, whereas creative algorithms are explicitly trying to extrapolate beyond their training (inspiration) set to yield ideas or artifacts that have never been imagined. For example, a human face image generation system would typically be quite interpolative and try to match the statistical properties of natural face images in a training corpus, whereas a culinary creativity system might be quite extrapolative in combining ingredients that have never been put together before. Explanations for interpolative and extrapolative AI can be quite different and can also be used to support the design of AI algorithms themselves as they move out of distribution while maintaining quality.

For the novelty dimension, scholars often distinguish ex post measures, which make reference to events that happen after an artifact appears, such as citation counts or auction prices from ex ante measures, which refer only to events that led up to its debut, such as the complexity or surprise of an idea. In the context of explanation, ex ante approaches focus on the thing itself, whereas ex post approaches also consider the social embedding in a stronger way to explain how things might be perceived.

As we noted previously, there are several classes of generative and creative AI models. Here we discuss some of them.

A normalizing flow model is a transformation of a simple probability distribution (e.g., a standard Gaussian distribution, typically in high dimensions) into a more complex distribution by a sequence of invertible and differentiable mappings. Neural networks are usually trained to implement these mappings. Then, the probability density of a sample can be evaluated by transforming it back to the original simple (Gaussian) distribution and computing the product of the density of the inverse-transformed sample under this distribution and the associated change in volume induced by the sequence of inverse transformations, measured by determinants of Jacobians of the transformations. This leads to a generative algorithm by sampling from the normalizing flow model.

Rather than working directly with probability densities and their transformations as in normalizing flows, VAEs develop a latent-space representation of conceptual spaces mapped using an encoder and a decoder, which are usually trained neural networks. As a part of the generative process, the encoder is removed and random samples are taken in the latent space and then decoded back into the original space.

GANs approach the problem in a different manner, where two adversarial neural network models are paired: a generator and a discriminator. The training involves each trying to do as well as possible for their respective tasks. The generative part of the system can then be used directly for generative processes. A particular form of GAN that is commonly used for generative applications is StyleGAN [27]. A style-based generator that has the ability to disentangle latent factors of variation into high-level attributes and stochastic features in the generated images in an unsupervised fashion is proposed. This enables an understandable, intuitive way to control the generative process at different scales.

Autoregressive models, such as Transformer neural network architectures that underlie models such as GPT-3 and CTRL [28], have gained recent prominence, not just for natural language generation but also for images and numerous other modalities. The basic idea of training is language modeling, where the goal is to predict the next token given the previous context using self-supervision. Attention mechanisms are leveraged so as to ensure that only the most relevant parts of context are used for prediction. In using these models for generation, the basic idea is just to sequentially make predictions for the next token and then sample from the predicted probability distribution. All of these creative/generative approaches may be difficult to explain in detail, so there is a desire to develop new approaches, which we now describe through several vignettes.

Deep Generative Models and Creativity

Although there has been a rapid emergence of generative AI models, standard, deep generative models discourage out-of-distribution generation to avoid instability. This minimizes spurious sample generation but limits potential in creativity. Indeed, [29] shows that a complete removal of "spurious" samples limits the capacity of the generative model. Nevertheless, a number of works aim for novel, creative generation using deep generative models like VAEs or GANs, formulating the task as a few-/one-shot generation problem [30]–[32] or as a style-ambiguous design [33]. In aiming for novelty and utility at the same time [7], an evaluation along both dimensions may be domain or subject dependent, labor extensive, and may also suffer from high levels of subjectivity. Therefore, new approaches that enhance the creative capacity of generative models, ensure a framework for systematic assessment of creativity by human experts, and provide model-, sample-, or task-level explanations of creativity need to be developed.

In the following, we present a recent case study aimed at developing a principled and label-free framework for enhancing the creativity of a pretrained deep generator [34]. Inspired by neuroimaging studies that show atypical neuronal activation in creative brains, a "creative" decoding algorithm is proposed that generates novel and meaningful samples from the original latent (concept) space, as modeled by a generative neural network. Mimicking the neuronal activation pattern seen in creative brains, the algorithm favors atypical coactivation of high- and low-active neurons during decoding from the pretrained generative model. Specifically, starting with a sample, $z \sim p(z)$, and corresponding neuronal activations for a selected layer k of the decoder, denoted as $d_{jz}^{k} = d_{j}^{k}(z)$, first, the set of active, on neurons and the set of inactive, off neurons are derived. During creative decoding, some fraction ρ of a group of on and/or off neurons in a layer k is flipped either randomly or selectively. For instance, a neuron is selected from the pool of off neurons that have the lowest percentage of activations, a_{j}^{k} (defined using a threshold cutoff). Next, low-active neurons that are most correlated with the selected neuron are also turned on. The method thus modifies $d_{j}^{k}(z)$, and this modified layer output is then passed through the remainder of the decoder to obtain the final generated samples. In a sense, creative decoding allows conceptual leaps by pretrained neural networks without further training or access to additional feedback from a machine learning model or human expert. This exploration process is congruent with the concept of exploratory creativity [7].

Algorithm 1. Creative Decoding Using Low-Active Neuron Activation.

```
Input: Layer output $d_{jz}^{k}$; Percentage activation of neuron $j$ in layer $k = a_{j}^{k}$; percent activation percentile $\kappa$; fraction of neurons to flip on $\rho$
    $t \leftarrow$ $\kappa$-percentile $(\{a_{j}^{k}\})$
    $\mathcal{S} \leftarrow \{j \mid j \in \mathcal{D} \wedge a_{j}^{k} \leq t\}$
    $s \leftarrow$ random select from $\mathcal{S}$
    $\mathcal{S}_{s} \leftarrow \underset{\mathcal{S}' \subset \mathcal{S}, |\mathcal{S}'| = \lfloor \rho |\mathcal{S}| \rfloor}{\operatorname{argmax}} \sum_{h \in \mathcal{S}'} R_{sh}^{k}$
    $d_{jz}^{k} \leftarrow o_{j}^{k}, \forall j \in \mathcal{S}_{s}$.
```

Figure 2 gives a high-level overview of the algorithm as well as some examples of decoded samples. A human assessment of creative decoding outcomes reveals that the proposed decoding method indeed results in samples consistent with human perception of creativity. Importantly, the study evaluates novelty and creativity separately, revealing that the neuro-inspired

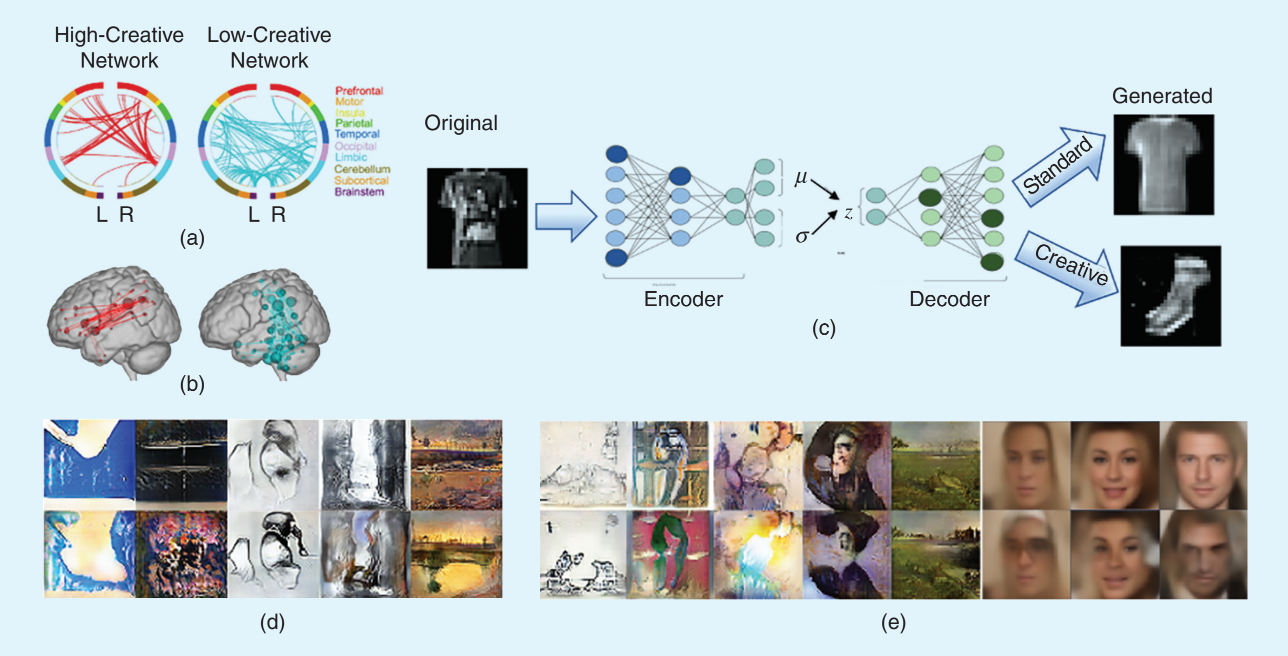


FIGURE 2. (a) and (b) Depictions [(a): circle plots; (b): glass brains] of high- and low-creative networks in human brains with their highest-degree nodes. The circle plot colors correspond to brain lobes. Adapted from [35]. (c) The depiction of a VAE model with our neuroinspired creative decoder. Normally, a small fraction of neurons in each hidden layer are low active (dark color). Inspired by the neural basis of creativity, we activate those "low-active" (task-negative) neurons to induce coupling between task-positive and task-negative neurons during "creative" decoding. (d) and (e) The samples generated by the proposed creative decoding method. The base generative models used were ARTGAN trained on (d) the WikiART data set and VAE trained on (e) the CelebA data set. Adapted from [34]. L: left hemisphere; R: right hemisphere.

decoding method is essential to generate a higher proportion of creative samples. Baselines such as decoding from linear interpolations in the latent space or activation of randomly selected off neurons during decoding fail to achieve this high level of creativity. Moreover, comparing human assessment with out-of-class novelty values (obtained using surrogate machine learning models) of generated samples shows that a linear combination of individual surrogate metrics (e.g., a trained L_1-regularized logistic regressor model) capture the human perception of creativity to a significant extent.

Explaining AI Generations

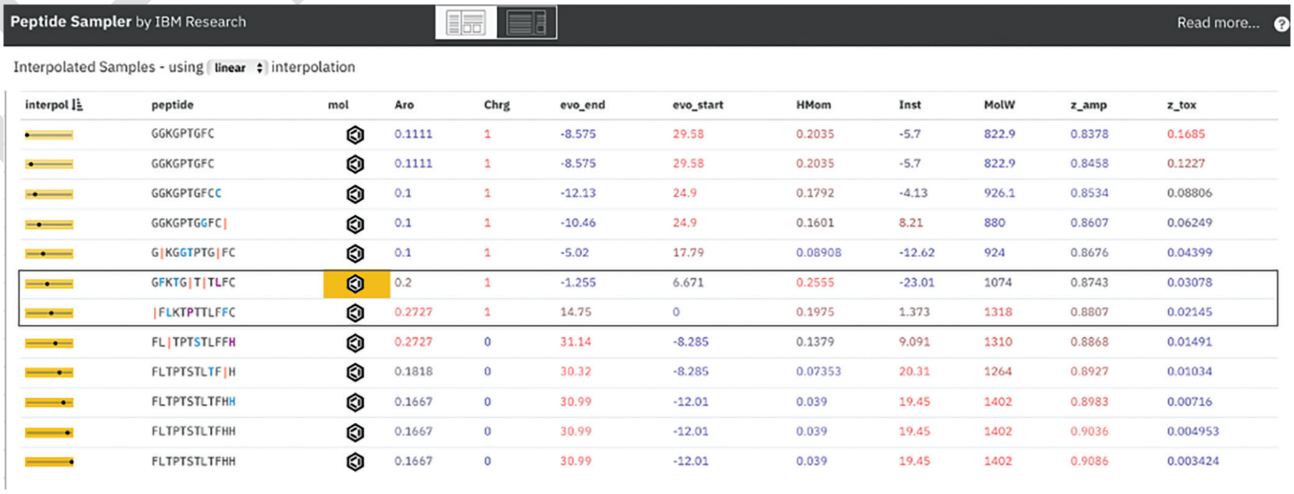
With AI models becoming more complex and therefore increasingly inscrutable, there is growing demand for interpretable or XAI models. This is particularly the case in scientific decision making and accelerating scientific discovery, where AI is seeing significant application. Generating explanations for black-box AI models is important, but assessing those explanations is challenging. For instance, it is not well defined how one should accurately account for the subjective variability in user perception of AI explanations. Further, user perception can vary with the type of explanations provided.

Posthoc techniques (including visualizations) have been developed to interpret the outputs of deep learning-based generative models and have been broadly applied to a variety of input modalities including images, natural and other domain-specific languages, and tabular data. Here we discuss a number of examples where posthoc explanations were studied for interpreting the outputs of deep learning-based generative models of biological sequences that include recurrent neural networks (RNNs), transformers, and generative autoencoders. Figure 3 shows the table generated by a visual analytic platform known as the *Peptide Sampler* (https://peptide-walk.mybluemix.net), which provides a visualization of decoded sequences and their attributes sampled from the latent space of biological peptides modeled by an RNN-based Wasserstein autoencoder (WAE) [3], [36]. In this specific screenshot, the platform illustrates the decoded sequences and their attributes during a linear interpolation between two known and distant antimicrobial peptide sequences in the WAE latent space. The gray box indicates a sampling of novel sequences with low sequence similarity to the endpoint sequences. Those novel sequences are of interest as they show high-antimicrobial class probability and low-toxic class probability as returned by property predictor models trained on the WAE latent embeddings of known peptides. A sequence alignment to the previous row (blue: amino acid addition, violet: amino acid swapped, red: amino acid removed) as well as different attribute changes along the interpolation provides a visual explanation of how peptide sequences are mapped in the WAE latent space.

In recent years, attention mechanisms have played a major role in natural language and other domain-specific language modeling and generation. Consequently, intermediate representations offered by these modules are being investigated for the purpose of explaining the reasoning for a model's behavior. For instance, recently, a 3D visualization of the attention of a Transformer protein sequence model has revealed an accurate capture of protein structural and binding properties [37], as shown in Figure 4(a). Similarly, a mapping of the attention differences learned by a long short-term memory protein sequence zoonotic potential classifier on the beta-coronavirus RNA-dependent RNA polymerase structure highlights a protein–protein interaction interface that is known to govern viral replication [Figure 4(b) and (c)] [38]. These visualizations are examples of displaying so-called BERTology analyses [39], which try to delve into what information/knowledge is present in Transformer-type neural architectures after training.

Connecting AI Explanations and AI Creativity

Explaining the generated artifacts or the model itself of a generative AI framework has more challenges than the typical XAI paradigm for explaining predictive models. Notably, the data distribution shift among training samples and novel generated artifacts is often what makes explanations more difficult. In that



Peptide Sampler by IBM Research

Interpolated Samples - using linear interpolation

interpol	peptide	mol	Aro	Chrg	evo_end	evo_start	HMom	Inst	MolW	z_amp	z_tox
	GGKGPTGFC	⬡	0.1111	1	-8.575	29.58	0.2035	-5.7	822.9	0.8378	0.1685
	GGKGPTGFC	⬡	0.1111	1	-8.575	29.58	0.2035	-5.7	822.9	0.8458	0.1227
	GGKGPTGFCC	⬡	0.1	1	-12.13	24.9	0.1792	-4.13	926.1	0.8534	0.08806
	GGKGPTGGFC|	⬡	0.1	1	-10.46	24.9	0.1601	8.21	880	0.8607	0.06249
	G|KGGTPTG|FC	⬡	0.1	1	-5.02	17.79	0.08908	-12.62	924	0.8676	0.04399
	GFKTG|T|TLFC	⬡	0.2	1	-1.255	6.671	0.2555	-23.01	1074	0.8743	0.03078
	|FLKTPTTLFFC	⬡	0.2727	1	14.75	0	0.1975	1.373	1318	0.8807	0.02145
	FL|TPTSTLFFH	⬡	0.2727	0	31.14	-8.285	0.1379	9.091	1310	0.8868	0.01491
	FLTPTSTLTF|H	⬡	0.1818	0	30.32	-8.285	0.07353	20.31	1264	0.8927	0.01034
	FLTPTSTLTFHH	⬡	0.1667	0	30.99	-12.01	0.039	19.45	1402	0.8983	0.00716
	FLTPTSTLTFHH	⬡	0.1667	0	30.99	-12.01	0.039	19.45	1402	0.9036	0.004953
	FLTPTSTLTFHH	⬡	0.1667	0	30.99	-12.01	0.039	19.45	1402	0.9086	0.003424

FIGURE 3. The Peptide Sampler visual analytic platform shows decoded sequences and their attributes during a linear interpolation between two selected sequences in the latent space of a generative autoencoder (a WAE in this example) trained on biological peptides.

direction, [40] employs an anomalous pattern-detection method referred to as *group-based subset scanning* to determine whether a given batch of generated samples contains creative samples, as determined by a human evaluator. Now we provide a short summary of how subset scanning treats the creative quantification and characterization problem as a search for the most anomalous subset of observations in the data.

Consider a set of samples from the latent space $X = \{X_1 \ldots X_M\}$ and nodes $O = \{O_1 \ldots O_J\}$ within the creative decoder. Let $X_S \subseteq X$ and $O_S \subseteq O$. The subsets S under consideration are defined as $S = X_S \times O_S$. The goal is to find the most anomalous subset:

$$S^* = \underset{S}{\operatorname{argmax}}\, F(S). \qquad (1)$$

A nonparametric scan statistic such as the Berk–Jones test statistic is then used on the p values that compare the node activations between the background (regular decoding) and evaluation (creative decoding). An important finding from this analysis is that creative generation requires a larger extent of distinctly anomalous node activation, as depicted in Figure 5. Thus, the anomalous pattern-detection framework provides explanations for creative artifacts, which is based on node activations of the trained models. Such explanation frameworks can provide guidance to generative AI models for promoting creative generation [41] and chart the path toward human–AI cocreation.

Model-Agnostic Posthoc Interpretation

Thus far, we have discussed explainability methods that make use of AI models themselves. One can, however, also perform a model-agnostic posthoc interpretation of creative artifacts without considering how they were produced. Some techniques use knowledge graph-based interpretation; here we describe an example using combinatorial structures within knowledge graphs or embedding spaces for explaining combinatorial creativity.

Combinatorial and compositional creativity—the generation of unfamiliar combinations of familiar ideas—is the typical kind of creativity performed by people and also pursued by computational creativity systems, whether implicitly or through the explicit combining of parts. Even exceptional levels of creativity have a combinatorial character. Several algorithmic techniques now perform combinatorial creativity for different application areas, including language modeling for language, simulated annealing for magic tricks, stochastic sampling + filtering as well as associative algorithms and language modeling for culinary recipes, neural network approaches for music and for building materials, and case-based reasoning for engineering processes. Although products of such creativity are readily interpretable by people in knowing what they are, the process of such creativity may be inscrutable and nonintuitive (in the sense of [8]). Yet, users of computational creativity technologies already anthropomorphize them.

For human understanding, an explanatory process is often just as important as the product; indeed, people want to understand the "theory of mind" of creators. This is especially the case when considering social creativity; understanding others is the most pervasive aspect of successful social interaction. In human-only creativity, the benefits of social interaction for creativity are well known. In this vignette, we describe an inverse problem formulation of going from a combinatorial artifact back to the human-like process that may have created it (even if not the underlying algorithm that actually created it) [42].

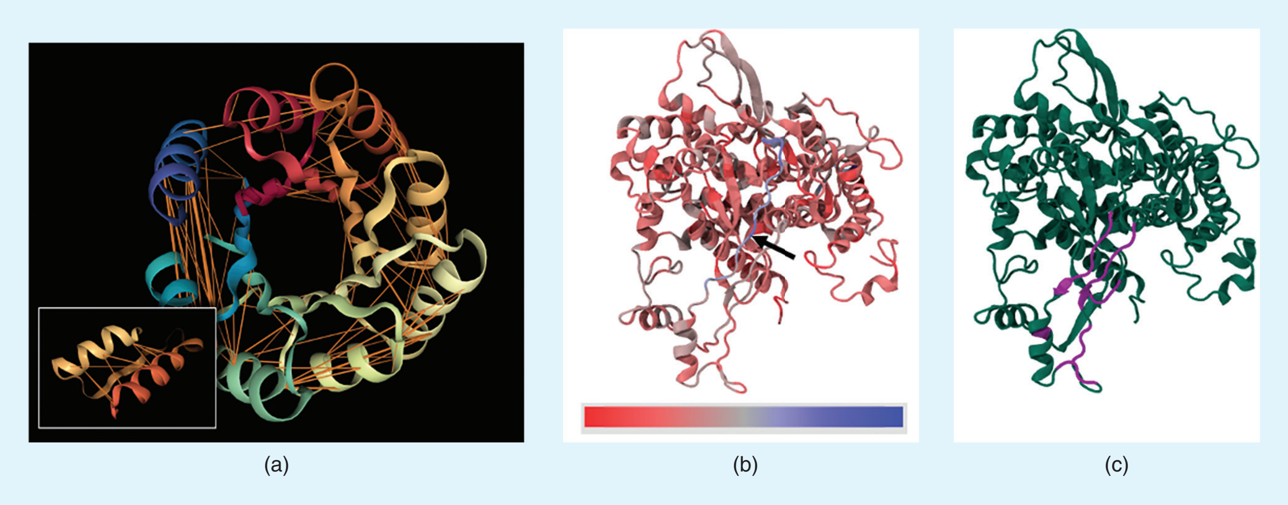


FIGURE 4. (a) An example of how specialized attention heads in a transformer recover protein structure and function, based solely on language model pretraining. The orange lines depict attention among amino acids. This is specifically an attention head that targets amino acid pairs that are close in physical space (see inset subsequence 117 D-157I) but lie apart in the sequence, which is shown for a de-novo-designed triose-phosphate isomerase-barrel (5BVL) with characteristic symmetry (adapted from [37]). (b) The average attention differences between two classes, as learned by a sequence-based zonnotic potential classifier model (adapted from [38]), mapped onto the severe acute respiratory syndrome-associated coronavirus nonstructural protein (nsp)12 structure (6NUR.pdbchain A residue 146–714). A red-gray-blue (low-to-high) color scale is used to visualize the attention heatmap. (c) The regions of RNA-dependent RNA polymerase interacting (cutoff: 6.5 Å) with nsp7 highlighted in magenta (adapted from [38]). This highlights that higher attentions within the positive sequences map to the N-terminal structural regions, which correspond to the nsp12-nsp7 interaction surface.

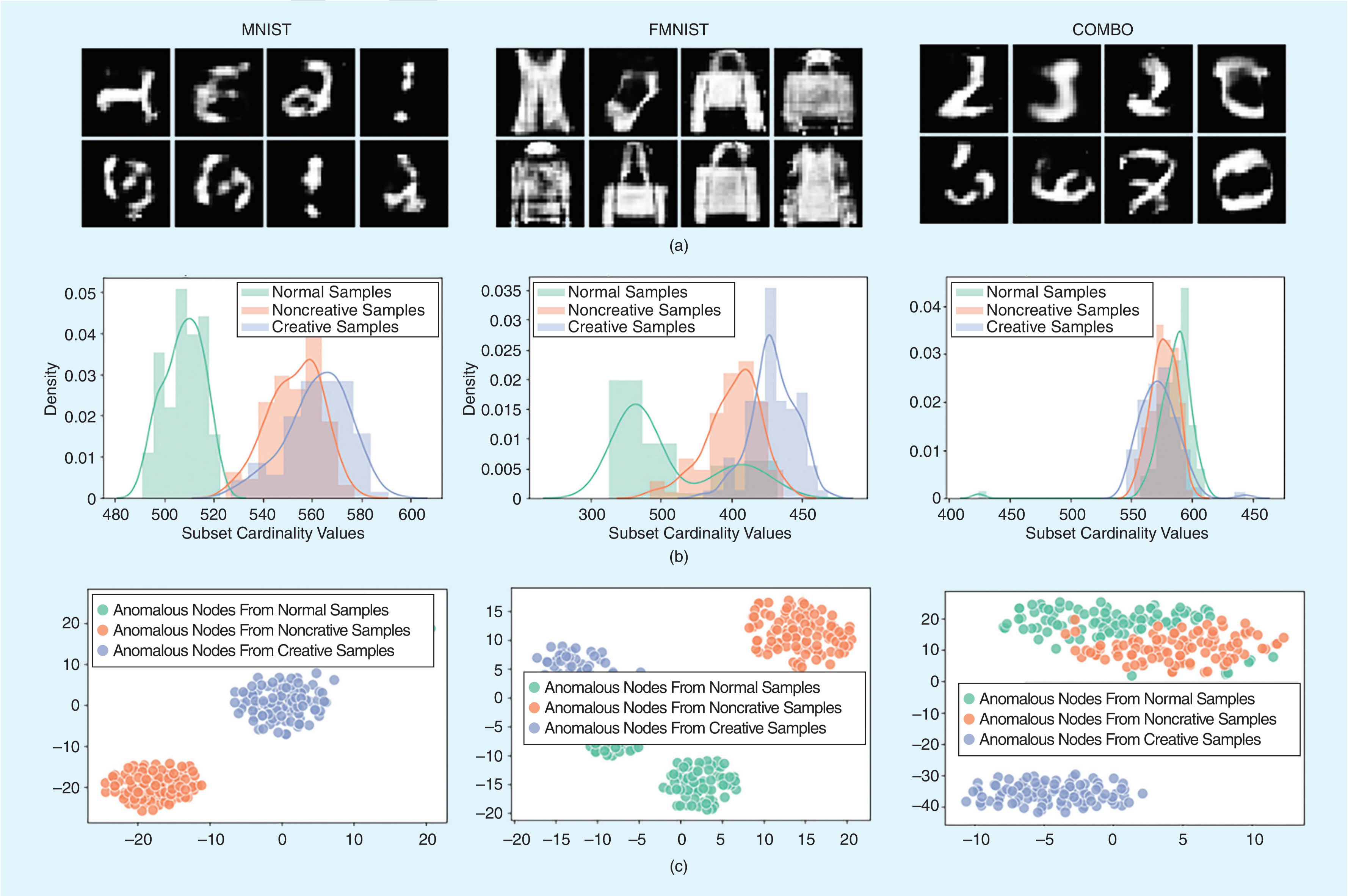


FIGURE 5. (a) Examples of images that are human annotated as creative. The samples were generated by the "creative" decoding of a VAE model trained on a Modified National Institute of Standards and Technology (MNIST), Fashion MNIST (FMNIST), and combined (COMBO) (MNIST + FMIST = COMBO) data set. Adapted from [34]. (b) and (c) The activations characterization. (b) The subset cardinality distributions for anomalous subsets for different types of generated samples. (c) t-distributed stochastic neighbor embedding over anomalous subset nodes for the creative decoder activations under generation of normal, noncreative, and creative samples. Adapted from [40].

Longstanding results in the behavioral sciences show that the way humans think, recall, and process information is largely through associations: an association is the connection between two or more concepts. *Human creativity* is often described as the mental process of combining associative elements into a new form. Indeed, creative processes are often thought of as forming associative elements into new combinations that either meet specified requirements or are in some way useful. Although remote associations are often indicators of creativity, more nearby associations are easier for people to understand and appreciate. Given that the human creative process is largely by association, one may specifically consider explanations using associative chains.

To find associative chains that use nearby associations, one can develop a traveling salesperson problem (TSP) formulation within knowledge graphs, where the nodes are components and edges are associations. Tours, paths, and other combinatorial structures within knowledge graphs are then possible explanations. This is the inverse problem to associative algorithms for computational creativity itself. In this sense, explanation is the opposite process of creativity (under a specific algorithm).

Figure 6 presents the following three examples of such explanations:

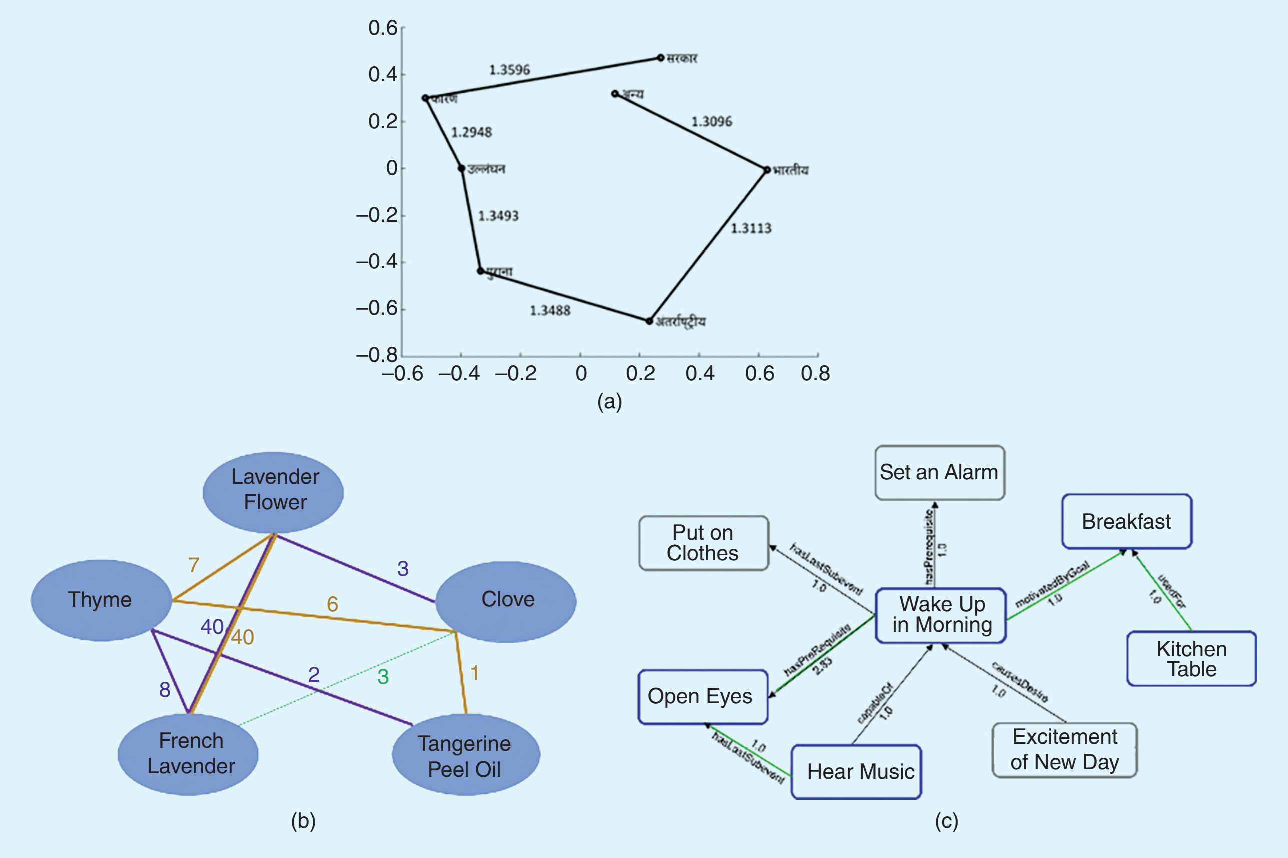
1) The culinary recipe of a new spice mixture that can be used for pastries: thyme, clove, tangerine peel oil, French lavender, and lavender flower.
2) An English sentence: “After hearing the music, I woke up in the morning and opened my eyes, after which I had breakfast at the kitchen table.”
3) A Hindi sentence that translates to “The old Indian government was the cause of other international governments unraveling.”

The first two examples have explicit knowledge graphs that can be used (flavor networks and ConceptNet, respectively), whereas the third example is for a less-resourced setting. A Euclidean word embedding is used as an implicit knowledge graph.

Note that in each of these examples, we are considering a semantic notion of novelty based on explicit or implicit knowledge graphs, rather than a statistical notion of novelty. This can be formalized as follows.

Definition 1

Consider a creative artifact α comprising components $\{x_1, \ldots, x_n\}$ that has a corresponding subgraph, $\tilde{\mathcal{G}}$, of knowledge graph $\mathcal{K}$. Then *the TSP-novelty* of α, $s_{\mathcal{K}}(\alpha)$, is defined to be $\mathbb{TSP}(\tilde{\mathcal{G}})$, where $\mathbb{TSP}(\cdot)$ is an operator for finding the traveling salesperson path length.



(a)

(b)

(c)

FIGURE 6. Examples of explanations via traveling salesperson tours. (a) Explaining a novel spice mixture based on the number of shared flavor compounds (more is a stronger association), where two Hamiltonian paths are highlighted in purple and brown; and an unused edge is in green. (b) Explaining a novel English sentence, where a knowledge graph is denoted by blue nodes and corresponding edges among them, computed using ConceptNet (ignoring directionality, larger values are stronger associations). The gray nodes could have been used for augmentation if needed. The path highlighted in green is a traveling salesperson path. (c) Explaining a novel Hindi sentence, computed using a pretrained word embedding in a 2D principal component analysis basis. As this is a fully connected graph, we omit the unused edges in the traveling salesperson path.

Considering the Euclidean setting, suppose a computational creativity algorithm selects components at random, as in stochastic sampling algorithms that, in principle, encompass all other kinds of computational creativity algorithms. Then we have the following celebrated result in computational geometry for the Euclidean TSP problem.

Theorem 1

Let $\{X_1, \ldots, X_n\}$ be a set of independent identically distributed (i.i.d.) random variables in $\mathbb{R}^d$ with bounded support. Then the length L_n of the shortest TSP tour through the points $\{X_i\}$ satisfies

$$\frac{L_n}{n^{(d-1)/d}} \to \beta_d \int_{\mathbb{R}^d} f(x)^{(d-1)/d} dx$$

with probability 1 as $n \to \infty$, where $f(x)$ is the absolute continuous part of the distribution of the $\{X_i\}$, and β_d is a constant that depends on d but not on $f(x)$.

The intuition from this concentration of measurement theorem is that, asymptotically, the choice of stochastic sampling distribution $f(x)$ in the creativity algorithm can directly control the traveling salesperson tour length in a given d-dimensional conceptual space. Concentrated distributions yield much less novelty than those that are disperse, and this is explicitly computable.

Moreover, note that the TSP tour length L_n is asymptotically, intimately tied to the Renyi entropy $H_\gamma(f(x))$ of the sampling distribution, where for $\gamma \in (0, 1)$,

$$H_\gamma(f) = \frac{1}{1-\gamma} \ln \int f(x)^\gamma (z) dz$$

and approaches the Shannon entropy as $\gamma \to 1$.

Theorem 2

Let $\{X_1, \ldots, X_n\}$ be a set of i.i.d. random variables in $\mathbb{R}^d$ with bounded support. Let L_n be the length of the shortest TSP tour through the points $\{X_i\}$. Then the following estimator for the Renyi entropy

$$\hat{H}_\gamma(X) = \frac{1}{1-\gamma}\left(\frac{\ln L_n}{n^\gamma} - \ln \beta\right),$$

where $\gamma = (d-1)/d$ and β is a fixed constant independent of $f(x)$, is an asymptotically unbiased and almost surely consistent estimator of the Renyi entropy of $f(x)$. That is to say, the TSP tour length is asymptotically a simple function of the Renyi entropy of the stochastic sampling distribution, which approaches the Shannon entropy in high dimensions.

This matches with statistical measures of novelty that are also simple functions of information-theoretic quantities such as Shannon entropy and mutual information; however, here, they are from a measure of novelty that emerges directly from explaining creative artifacts via associative chains.

Thus we see that creativity algorithms can be designed directly using a measurement that emerges from explainability. Various generative algorithms, such as normalizing flows and VAEs, can be thought of as stochastic sampling.

Intrinsically Interpretable Generation

Thus far, we have discussed the processes by which specific creative algorithms work and also the outputs of creative algorithms in a model-agnostic way via associative chains. Both have led to specific design principles for creative AI algorithms and also methods for human–AI interaction. Now we consider an alternative where both the processes and products of creativity are intrinsically interpretable. This vignette focuses on a nonneural network technique called *ILL* [19], [43], which draws upon mathematical techniques from both information and group theory.

ILL is a general framework used to learn rules of a signal (e.g., an image or a probability distribution). Here a rule is a coarsened signal used to help people gain one interpretable insight about the original signal. To make full sense of what might govern the signal's intrinsic structure, multiple disentangled rules arranged in a hierarchy, called a *lattice*, are sought. Compared to representation/rule-learning models optimized for a specific task like classification, ILL focuses on explainability: it is designed to mimic human experiential learning and discover rules akin to those humans distill and can apprehend. Thus, ILL addresses the fundamental question "what makes X an X" by creating rule-based explanations designed to help humans understand. By having a human-interpretable way of explaining X, one can then also sample from the space, e.g., in a maximum-entropy manner, while satisfying rules that are themselves interpretable.

The mathematical intuition behind ILL is to break the whole into simple pieces, similar to breaking a signal into a Fourier series. Whereas a Fourier analysis decomposes a signal in a Hilbert space via an inner product and synthesizes it via a weighted sum, ILL decomposes a signal in a hierarchical space called a *lattice*. The goal is to restore human-like, hierarchical rule abstraction and realization via signal decomposition and synthesis in a lattice, called *projection and lifting*, resulting in more than the sum of its parts.

ILL has two phases: lattice construction and learning/searching in the lattice. This is similar to many machine learning models, such as normalizing flows or VAEs comprising function class construction then learning in the function class, e.g., by constructing a neural network architecture then learning/finding optimal network parameters via backpropagation. ILL's construction phase builds on universal priors consistent with human innate cognition, such as a knowledge of symmetries (group-theoretic invariances) that are readily understandable, which then grows a lattice of abstractions. ILL's learning phase operates on "small data" as a signal but searches for rich explanations of the signal via rule learning, wherein abstraction is key to "making small data large." Notably, the construction phase is prior, not data driven—data comes in only at the learning phase.

ILL has been demonstrated to recover 80% of the University of Illinois Urbana-Champaign's music theory curriculum and to discover powerful new concepts of interest to music theorists on the basis of just 370 chorales by Bach in the same human-interpretable form as textbooks [19]. Given its intrinsic interpretability, ILL also supports new modes of human–AI creativity, where music can be decomposed into fragments that corresponds to harmony, melody, rhythm, texture, and so on and then recomposed in a way that yields a high-quality result, even when human participants are not well versed in music theory.

Conclusions

Computational creativity dates back at least to the 1956 Dartmouth meeting that established AI as a research discipline. In this article, we argued that it is important for explainability in AI to not just be considered for decisions and predictions but also for generative and creative algorithms as well. We specifically discussed the intrinsic interpretability and posthoc explainability (both model focused and model agnostic) of AI generations/generative models via several vignettes. Notably, such results provide a virtuous interaction among advances in creative AI, XAI, and human–AI interaction for advancing and understanding creativity. The research we presented in this article also suggests open research questions, such as on hybrid approaches that combine intrinsic and posthoc interpretability.

Beyond human interpretability, explainability is also said to require reproducibility and replicability. Although this holds directly in settings of AI-based decisions and predictions, there is some subtlety to this point in generative/creative settings. Indeed, the purpose of creativity is to generate ideas or artifacts that have never been imagined before, and each run of the algorithm is intended to produce novel ideas or artifacts (often influenced by the specific prompt, seed, or conditioning variable given, but also various sources of randomness). As such, it is important to characterize the sources of randomness in various generative algorithms and how they work. In normalizing flows, randomness is a part of sampling from the simple (Gaussian) distribution, whereas for VAEs and GANs, there is randomness in sampling from the latent space. In language models, there can be randomness in the various neural decoding algorithms that are used on top of the trained models. In ILL, again, there can be randomness in maximum-entropy sampling from within valid rule sets. Moreover, the provenance of data that is used for training neural network-based generative algorithms or for inspiring other kinds of generative algorithms is of critical importance, whether in the governance of AI safety, or in questions that arise in intellectual property law and free speech law. Self-awareness and self-evaluation are also highly relevant as creative AI algorithms should know "when to stop" [44]—as a path toward gaining some low-level autonomy. The recent advances in steerable or controllable generative models [3] are highly relevant in this context as they not only can allow control over those AI frameworks but can also aid in bridging explanations with creations.

Finally, we note that discussions of interpretable AI, especially in settings of decisions/predictions, often start with the misconception that there is a fundamental tradeoff between interpretability and accuracy, but numerous examples show the contrary. In the context of generative algorithms, our vignettes demonstrated that not only is there no tradeoff between understanding and novelty/quality of results, but in fact, ensuring interpretability/explainability can even enhance performance on creativity metrics.

Acknowledgment

This work was supported, in part, by the IBM-Illinois Center for Cognitive Computing Systems Research, a research collaboration as a part of the IBM AI Horizons Network, and in part by the IBM-Illinois Discovery Accelerator Institute.

Authors

Payel Das (daspa@us.ibm.com) received her Ph.D. degree from Rice University in 2007. Currently, she is a principal research staff member and a manager at IBM Research AI, IBM Thomas J. Watson Research Center, New York, USA. She is also an IBM Master Inventor. Her research interest includes statistical physics, trustworthy machine learning, neuro- and physics-inspired artificial intelligence, and machine creativity. She is a Member of IEEE.

Lav R. Varshney (varshney@illinois.edu) received his Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology. He is an associate professor of electrical and computer engineering at the University of Illinois Urbana-Champaign, Champaign, Illinois, USA, with further affiliations in computer science, neuroscience, industrial engineering, digital agriculture, and personalized nutrition as well as the Discovery Partners Institute. He is a founder of Kocree, Inc., Champaign, IL, USA, a start-up company focused on social cocreativity platforms. His research interests include information theory, artificial intelligence, and creativity. He is a Senior Member of IEEE.

References

[1] D. Somaya and L. R. Varshney, "Ownership dilemmas in an age of creative machines," *Issues Sci. Technol.*, vol. 36, no. 2, pp. 79–85, Winter 2020.

[2] L. R. Varshney, F. Pinel, K. R. Varshney, D. Bhattacharjya, A. Schörgendorfer, and Y.-M. Chee, "A big data approach to computational creativity: The curious case of Chef Watson," *IBM J. Res. Develop.*, vol. 63, no. 1, pp. 7:1–7:18, 2019, doi: 10.1147/JRD.2019.2893905.

[3] P. Das, T. Sercu, K. Wadhawan, I. Padhi, S. Gehrmann, F. Cipcigan, V. Chenthamarakshan, H. Strobelt *et al.*, "Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations," *Nat. Biomed. Eng.*, vol. 5, no. 6, pp. 613–623, 2021, doi: 10.1038/s41551-021-00689-x.

[4] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, Jr., C. Xiong *et al.*, "Deep neural language modeling enables functional protein generation across families," Jul. 2021, bioRxiv 2021.07.18.452833.

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," Feb. 2021, arXiv:2102.12092 [cs.CV].

[6] M. O. Riedl, "The Lovelace 2.0 test of artificial intelligence and creativity," in *Proc. 29th AAAI Conf. Artif. Intell. Workshops*, Jan. 2015.

[7] M. A. Boden, *The Creative Mind: Myths and Mechanisms*, 2nd ed. London, U.K.: Routledge, 2004.

[8] A. D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," *Fordham Law Rev.*, vol. 87, no. 3, pp. 1085–1139, 2018.

[9] M. Hind, D. Wei, M. Campbell, N. C. F. Codella, A. Dhurandhar, A. Mojsilovic, K. Natesan Ramamurthy, and K. R. Varshney, "TED: Teaching AI to explain its decisions," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES '19)*, Jan. 2019. pp. 123–129.

[10] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in *Proc. Conf. Fairness, Accountability, Transparency (FAT* '19)*, Jan. 2019, pp. 279–288.

[11] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, and F. Doshi-Velez, “Human evaluation of models built for interpretability,” in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing (HCOMP)*, Oct. 2020, vol. 7, no. 1, pp. 59–67.

[12] S. Agarwal and L. R. Varshney, “Limits of deepfake detection: A robust estimation viewpoint,” in *Proc. ICML Workshop on Deep Learning Detecting AudioVisual Fakes,* Jun. 2019.

[13] L. R. Varshney, N. S. Keskar, and R. Socher, “Pretrained AI models: Performativity, mobility, and change,” Sep. 2019, arXiv:1909.03290 [cs.CY].

[14] R. G. Collingwood, *The Principles of Art*. Oxford, U.K.: Clarendon, 1938.

[15] L. R. Varshney, “Limits theorems for creativity with intentionality,” in *Proc. 11th Int. Conf. Comput. Creativity (ICCC)*, Sep. 2020, pp. 390–393.

[16] E. M. Bender and A. Koller, “Climbing towards NLU: On meaning, form, and understanding in the age of data,” in *Proc. Annu. Conf. Assoc. Comput. Linguistics (ACL),* Jul. 2020, pp. 5185–5198.

[17] D. Ventura, “Autonomous intentionality in computationally creative systems,” in *Computational Creativity*, T. Veale and F. A. Cardoso, Eds. Cham, Switzerland: Springer-Verlag, 2019, pp. 49–69.

[18] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg *et al*., “On the opportunities and risks of foundation models,” Aug. 2021, arXiv:2108.07258 [cs.LG].

[19] H. Yu, J. A. Evans, and L. R. Varshney, “Information lattice learning,” unpublished.

[20] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, “Do large scale molecular language representations capture important structural information?” Jun. 2021, arXiv:2106.09553 [cs.LG].

[21] V. Chenthamarakshan, P. Das, S. Hoffman, H. Strobelt, I. Padhi, K. W. Lim, B. Hoover, M. Manica *et al*., “CogMol: Target-specific and selective drug design for COVID-19 using deep generative models,” in *Advances in Neural Information Processing Systems 33*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, pp. 4320–4332.

[22] R. Goodwin, R. Segal, J. Maria, P. Das, R. Horesh, J. Fu, and C. Harris, “AI for fragrance design,” in *Proc. Machine Learning for Creativity and Design NIPS 2017 Workshop*, Dec. 2017.

[23] X. Ge, R. T. Goodwin, J. R. Gregory, R. E. Kirchain, J. Maria, and L. R. Varshney, “Accelerated discovery of sustainable building materials,” in *Proc. AAAI Spring Symp. Towards AI Collaborative Open Sci.,* Mar. 2019.

[24] A. Jain, D. Oliveira, A. Sharma, L. R. Varshney, C. Watson, K. Weldemariam, D. Wuebbles, and B. Zadrozny, “Toward an AI-based framework for accelerated discovery of climate impacts on agriculture,” in *Proc. AAAI Fall Symp. AI Meets Food Security,* Nov. 2021.

[25] C.-Z. A. Huang, H. V. Koops, E. Newton-Rex, M. Dinculescu, and C. J. Cai, “AI song contest: Human-AI co-creation in songwriting,” in *Proc. 21st Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Jul. 2020, pp. 708–716.

[26] T. B. Hashimoto, H. Zhang, and P. Liang, “Unifying human and statistical evaluation for natural language generation,” in *Proc. North Amer. Assoc. Comput. Linguistics (NAACL-HLT)*, Jun. 2019, pp. 1689–1701.

[27] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proc. IEEE Conf. Comput. Vision Pattern Recogn. (CVPR),* Jun. 2019, pp. 4401–4410.

[28] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “CTRL: A conditional transformer language model for controllable generation,” Sep. 2019, arXiv:1909.05858 [cs.CL].

[29] B. Kégl, M. Cherti, and A. Kazakçi, “Spurious samples in deep generative models: Bug or feature?” Oct. 2018, arXiv:1810.01876 [cs.LG].

[30] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “Human-level concept learning through probabilistic program induction,” *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015, doi: 10.1126/science.aab3050.

[31] D. J. Rezende, S. Mohamed, I. Danihelka, K. Gregor, and D. Wierstra, “One-shot generalization in deep generative models,” Mar. 2016, arXiv:1603.05106 [stat.ML].

[32] L. Clouâtre and M. Demers, “FIGR: Few-shot image generation with reptile,” Jan. 2019, arXiv:1901.02199 [cs.LG].

[33] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, “CAN: Creative adversarial networks, generating ‘art’ by learning about styles and deviating from style norms,” in *Proc. 8th Int. Conf. Comput. Creativity*, Jun. 2017, pp. 96–103.

[34] P. Das, B. Quanz, P.-Y. Chen, J. wook Ahn, and D. Shah, “Toward a neuro-inspired creative decoder,” in *Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2020, pp. 2746–2753.

[35] R. E. Beaty, Y. N. Kenett, A. P. Christensen, M. D. Rosenberg, M. Benedek, Q. Chen, A. Fink, J. Qiu *et al*., “Robust prediction of individual creative ability from brain functional connectivity,” in *Proc. Nat. Acad. Sci.*, 2018, pp. 1087–1092, doi: 10.1073/pnas.1713532115.

[36] T. Sercu, S. Gehrmann, H. Strobelt, P. Das, I. Padhi, C. Dos Santos, K. Wadhawan, and V. Chenthamarakshan, “Interactive visual exploration of latent space (IVELS) for peptide auto-encoder model selection,” in *Proc. ICLR 2019 Workshop on Deep Generative Models Highly Struct. Data*, May 2019.

[37] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, “BERTology meets biology: Interpreting attention in protein language models,” in *Proc. 9th Int. Conf. Learning Representations (ICLR)*, May 2021.

[38] K. Wadhawan, P. Das, B. A. Han, I. R. Fischhoff, A. C. Castellanos, A. Varsani, and K. R. Varshney, “Towards interpreting zoonotic potential of beta-coronavirus sequences with attention,” in *Proc. ICLR 2021 Workshop: Machine Learning Preventing Combating Pandemics*, May 2021.

[39] A. Rogers, O. Kovaleva, and A. Rumshisky, “A primer in BERTology: What we know about how BERT works,” *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 842–866, Jan. 2021, doi: 10.1162/tacl_a_00349.

[40] C. Cintas, P. Das, B. Quanz, S. Speakman, V. Akinwande, and P.-Y. Chen, “Towards creativity characterization of generative models via group-based subset scanning,” in *Proc. Synthetic Data Generation ICLR 2021*, May 2021.

[41] V. Nagisetty, L. Graves, J. Scott, and V. Ganesh, “xAI-GAN: Enhancing generative adversarial networks via explainable AI systems,” in *Proc. AAAI Explainable Agency Artif. Intell. Workshop*, Feb. 2021.

[42] L. R. Varshney, N. F. Rajani, and R. Socher, “Explaining creative artifacts,” in *Proc. ICML 2020 Workshop on Hum. Interpretability Machine Learning (WHI)*, Jul. 2020.

[43] H. Yu, H. Taube, J. A. Evans, and L. R. Varshney, “Human evaluation of interpretability: The case of AI-generated music knowledge,” in *Proc. ACM CHI 2020 Workshop on Artif. Intell. HCI: A Modern Approach*, Apr. 2020.

[44] C. Moruzzi, “Measuring creativity: An account of natural and artificial creativity,” *European J. Philosophy Sci.*, vol. 11, no. 1, pp. 1–20, 2021, doi: 10.1007/s13194-020-00313-w.

If an AI system can be explained, its intention becomes much more clear.

Creative AI is a special case of generative AI where novelty is explicitly important.

Generating explanations for black-box AI models is important, but assessing those explanations is challenging.

For human understanding, an explanatory process is often just as important as the product; indeed, people want to understand the "theory of mind" of creators.

Various generative algorithms, such as normalizing flows and VAEs, can be thought of as stochastic sampling.

ILL's learning phase operates on "small data" as a signal but searches for rich explanations of the signal via rule learning, wherein abstraction is key to "making small data large."