Operationalizing Essential Characteristics of Creativity in a Computational System for Music Composition

Paul M. Bodily[1], **Dan Ventura**[2]

[1]Department of Computer Science, Idaho State University, 921 South 8th Avenue, Pocatello, Idaho, 83209
[2]Department of Computer Science, Brigham Young University, 3361 Talmage Building, Provo, Utah, 84604
bodipaul@isu.edu, ventura@cs.byu.edu

Abstract

We address the problem of building and evaluating a computational system whose primary objective is creativity. We illustrate seven characteristics of computational creativity in the context of a system that autonomously composes Western lyrical music. We conduct an external evaluation of the system in which respondents rated the system with regard to each characteristic as well as with regard to overall creativity. Average scores for overall creativity exceeded the ratings for any single characteristic, suggesting that creativity may be an emergent property and that unique research opportunities exist for building CC systems whose design attempts to comprehend all known characteristics of creativity.

Introduction

AI has been outperforming humans on certain types of narrow intelligence tasks for quite some time [e.g., (Silver et al. 2016)]; however, this is still not the case for other types of tasks, including especially those requiring creativity, leading researchers to characterize *computational creativity* (CC) as "the final frontier" of AI (Colton and Wiggins 2012). Building systems that are computationally creative requires both the explicit identification of and the intentional operationalization of necessary characteristics of creativity.

State-of-the-art foundation models exhibit autonomy and generativity, producing impressive results, most notably language and image artifacts. As a result, one might be tempted to claim such systems are creative as a side-effect of their primary objective (e.g., next-token prediction). However, while autonomy and generativity are certainly necessary for creativity, they are not sufficient, particularly in highly structured and specialized domains (Ventura 2016). Notably, current artificial neural-network-focused implementations (including highly-successful transformer and large language models) have demonstrated weakness when highly-structured or constrained generation is needed, e.g., in certain forms of poetry and music (Hadjeres and Nielsen 2018; Glines 2022).

While the concept of creativity is likely inherently contestable, multiple characteristics have been identified as necessary for computational creativity (Colton et al. 2015; Ventura 2017; Jordanous 2014) with many, if not all, ultimately derived from research on human creativity [cf. (Csikszentmihalyi 1997)]. In what follows, we examine a set of seven such characteristics of creativity: *generation*, *knowledge representation*, *intentionality*, *aesthetic*, *domain knowledge*, *autonomy* and *self-evaluation*, and we do so in the context of a highly-structured domain of creativity, namely structured lyrical lead sheet generation for Western-popular music.

We present an autonomous system for music composition *explicitly designed to be creative*. The system incorporates a variety of machine learning models aimed at achieving (many of) the characteristics that distinguish creativity, and other work details these models and how they address aspects of the problem of creativity (in the context of pop music composition). Here, we present the complete system for the first time. We evaluate the system to assess the extent to which observers perceive these characteristics individually compared to their perception of creativity overall. Results show that the perception of creativity overall generally exceeds the perception of any of its constituent characteristics individually, emphasizing that creativity emerges from, and is greater than, the sum of its parts.

Characteristics of Creative Systems

A high-level overview of the system—named Pop* (Pop Star) —is shown in Figure 1. Pop* searches Twitter for a posting that appeals to its aesthetic. This tweet serves as an originating idea from which Pop* formulates its own intention—a theme that the system will communicate in the form of a novel music composition. Pop* then starts a learning phase in which lyric and sheet music databases are filtered based on relevance to Pop*'s chosen intention. Pop* uses these filtered training sets to train Markov models for chords, rhythm, pitch, and lyrics. Besides learning models for local structure, Pop* also learns global structural patterns (e.g., musical motifs, verse-chorus structure, rhyme schemes, etc.) from existing sheet music. The generative process utilizes constrained Markov models to produce a novel composition that is scored and filtered according to an intention-driven self-evaluation function. Compositions that pass the self-evaluation phase are rendered as audio and sheet music.

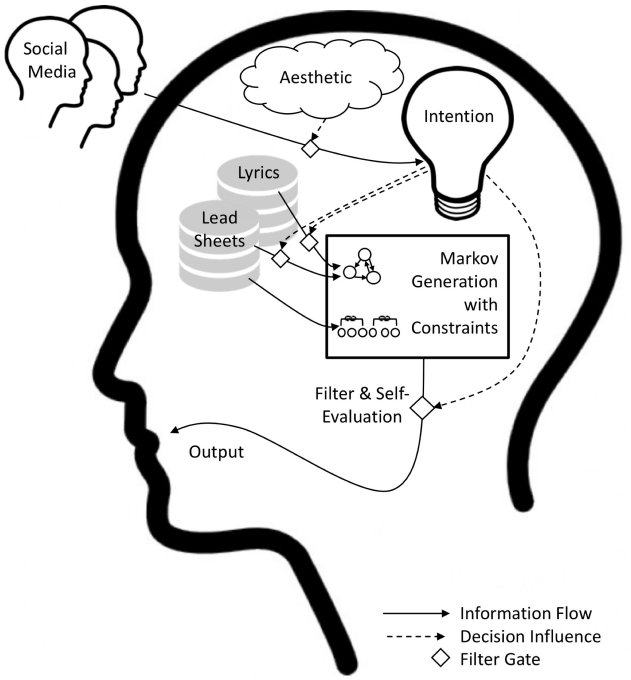

Figure 1: *Pop* process diagram.* The system originates an intention from social media posts that appeal to its aesthetic, which guides the training of generative models through a targeted selection of lyrics and lead sheets. Generated artifacts are output if they pass an intention-driven self-evaluation.

Characteristic #1: Generation

A generative system produces artifacts that can be argued to be more or less creative in the context of a broader, culturally defined class (Wiggins 2006). In addition to traditional conceptualizations of creative artifacts, the process of generation can also include more abstract elements including descriptions of production methods, process, and means of evaluation (Ritchie 2007; Colton 2008; Colton, Pease, and Charnley 2011). Generative systems can range from the purely stochastic to creation with the aid of perceptual faculties (Ventura 2016), raising questions as to whether knowledge of the system's process is needed to determine its creativity (Boden 2004; Kasof 1995; Ventura 2017).

To achieve local cohesive modeling without loss of global structure, Pop* uses *constrained* or *non-homogeneous Markov models* (NHMMs) (Bodily and Ventura 2022). These models—built from a Markov model and a set of constraints—were first designed to model small-scale transition patterns while allowing for constraints to be imposed at various positions. These constraints can be used to create the impression of global structure (Barbieri et al. 2012).

A NHMM $N = (n, M, C_1)$ models the probability of a sequence of random variables $X = X_1, \ldots, X_n$. Defining N requires defining a sequence length n, a Markov model M, and a set of unary constraints $C_1 = \{c_1, \ldots, c_n\}$. Each unary constraint $c_i \in C_1$ represents a function $f_i(x)$ that maps assignments of X_i (i.e., $x \in \Sigma$) to either 1 (c_i is satisfied) or 0 (c_i is not satisfied). A particular assignment $x = x_1, \ldots, x_n$ to $X = X_1, \ldots, X_n$ *satisfies* C_1 if $\forall i$, x_i satisfies c_i. $P_N(x) \propto P_M(x)$ if x satisfies C_1 and is 0 otherwise. For details see (Pachet, Roy, and Barbieri 2011).

Pop* builds a modified NHMM $N = (n, M, C_1, C_2)$ that allows the definition of a set C_2 of *binary* constraints. These binary constraints are essential to creating rhymes, motifs, and sectional form structure. Such a model is trained for each of the chord, pitch, rhythm, and lyric viewpoints, resulting in the four models: N_c, N_p, N_r, N_l. The details of this implementation are presented in (Bodily and Ventura 2022).

Characteristic #2: Knowledge Representation

A knowledge representation defines a structured model of problems, the cognitive process for solving these problems, and the artifacts themselves. There is often a distinction between internal and external representations, with internal representation in some cases being conflated with the choice of generative model (Ventura 2017). The internal representation determines the ability to generalize within or across domains (Lake, Salakhutdinov, and Tenenbaum 2015; Ventura 2016). Some forms of knowledge representation are manually crafted (e.g., rule-based systems or predefined grammars) whereas others are learned from observing existing domain artifacts (i.e., machine learning).

The knowledge representation in Pop* uses a hierarchical Bayesian program learning (HBPL) model consisting of a hand-crafted hierarchy of human-level concept models each trained from data (Lake, Salakhutdinov, and Tenenbaum 2015). The composition of a lead sheet γ for a given aesthetic A is modeled as

$$P(\gamma|A) = P(\nu|A) \cdot P(\tau) \cdot P(\eta|\nu,\tau) \cdot P(\phi|\nu,\tau,\eta) \cdot P(\rho|\nu,\tau) \cdot P(\lambda|\nu,\tau,\rho),$$

where ν, τ, η, ϕ, ρ, and λ represent intentions, structure, chords, pitch, rhythm, and lyrics, respectively.

Pop* is able to individually train each submodel on a potentially different dataset: $P(\eta|\nu,\tau)$, $P(\phi|\nu,\tau,\eta)$, and $P(\rho|\nu,\tau)$ on a knowledge base of music and $P(\lambda|\nu,\tau,\rho)$ on a knowledge base of lyrics. We use NHMMs to implement the models for sampling η, ϕ, ρ, and λ. Pop* is able to generalize its learning in specific areas and generate descriptions of design decisions at varying levels of detail. In addition to sheet music and audio artifacts, Pop* outputs a short description that explains its intention for the composition, the source of its intention, and a comment evaluating how well the intention was accomplished.

Characteristic #3: Intentionality

An intentional system is a deliberative system whose artifacts are the result of a directed process towards a particular objective (Ventura 2016; Ackerman et al. 2017). Intentionality is reflected in a system's ability to choose its *own* objectives (Guckelsberger, Salge, and Colton 2017) and to evaluate its own success in accomplishing goals (Ackerman et al. 2017). Intentions can relate to content, style, external impact, or type of generative act (Colton, Pease, and Charnley 2011), or other facets of creativity. To effectively enhance the perception of creativity, intention must be specific

joy	optimism	fear	lust
surprise	disappointment	hate	shame
anger	positive_emotion	envy	disgust
sadness	negative_emotion	love	timidity

Table 1: *Emotion Topics*. Empath topics used to find emotionally charged tweets.

enough to pose a challenge without being so focused as to suggest determinism. (cf. the balance required for inducing flow states as described by (Csikszentmihalyi 1997).)

Pop* explicitly models intention as a vector of topics or emotions $V = ((v_1, w_1), \ldots, (v_n, w_n))$ where (v_i, w_i) represents a topic and its weight. The system's objective is to create music that communicates each topic v_i to an extent commensurate with its weight w_i.

Pop* computes an intention V using Stanford's Empath library (Fast, Chen, and Bernstein 2016). Given a text input ι, Empath creates $Em(\iota) = ((v_1, w_1), \ldots, (v_{200}, w_{200}))$ where v_i represents one of a set of 200 topics (defined by Empath) and $w_i \in [0, 1]$ represents the strength of the semantic relationship between ι and v_i. The weights are normalized such that $\sum_i w_i = 1$. To establish an intention, Pop* assigns $V = Em(\tilde{t})$ where $\tilde{t}$ is some text (e.g., a tweet). The weights of all but the two highest scoring topics in V are set to 0.0 to narrow and focus the system's intention on the most important topics (in the case of ties, scores for all topics with tying scores are kept).

Characteristic #4: Aesthetic

In contrast to intention, aesthetic plays a more persistent role in determining style and is represented by the ability to have opinions, attitudes, or beliefs about what is beautiful, good, and creative (Papadopoulos and Wiggins 1999; Koren 2010; Mothersill 2004; Wiggins 2006). Simple systems may rely on manual encoding of an aesthetic (Ventura 2017); however, more creative systems may autonomously initiate and make changes to their aesthetic standards (Colton, Pease, and Charnley 2011; Jennings 2010). A system may form aesthetic standards with respect to *skill*, *imagination*, and *appreciation* (Colton 2008); *value* and *surprise* (Boden 2004); complexity (Hofstadter 1980); order (Birkhoff 1933); and entropy (Shannon 2001). An ideal aesthetic should be *explainable* (Bodily and Ventura 2018).

For Pop*, an aesthetic $A = (\theta_0, \ldots, \theta_m)$ is a list of interests where θ_i represents any keyword or phrase in which Pop* is "interested". In this formulation, Pop* has a favorable opinion of any interest listed in A and no opinion about anything else. For the results below, $A =$ ("being in love", "feeling depressed", "new beginnings").

For each theme $\theta_i \in A$, Pop* searches Twitter for (up to) 500 tweets from the prior 24 hours using θ_i as the search key to obtain a set of tweets T. From T, Pop* filters out retweets and tweets of less than 100 characters. For each remaining tweet $t \in T$, the Empath vector $Em(t)$ is computed. An $emotion\_score(t)$ is computed using the Empath vector $Em(t) = ((v_1, w_1), \ldots, (v_{200}, w_{200}))$ as

$$emotion\_score(t) = \sum_i w_i \text{ s.t. } v_i \in E$$

where E is the subset of Empath topics representing emotions (see Table 1). Pop* retains the set T_{θ_i} of ten tweets with the highest $emotion\_score$ values for each interest θ_i. From $\bigcup_i T_{\theta_i}$ one tweet $\tilde{t}$ is selected at random as an originating tweet for the system's generative process.

Characteristic #5: Domain Knowledge

Domain knowledge includes an understanding of the structure of artifacts within the domain, requirements for inclusion within the domain, and principles governing the evaluation of artifacts within the domain. Domain knowledge is often based on an *inspiring set* of domain-representative artifacts, which serves as inspiration during the generative process, enabling the system to evaluate novelty and typicality as a function of similarity to existing items (Ritchie 2007); learn from examples that are devoid of the designer's (explicit) biases or "fingerprints" (Ventura 2016); and socially evolve through addition and subtraction of information, including adding its own (successful) artifacts to its knowledge base (Pérez y Pérez and Sharples 2004).

Pop* leverages domain knowledge through machine learning on two primary knowledge bases. The lyric knowledge base (LKB) comprises lyrics from 369,606 songs scraped from www.lyrics.com. Lyrics $\mathcal{L}$ for each song are annotated with an Empath vector $Em(\mathcal{L})$. Given an intention ν, Pop* selects a subset of the LKB for training determined by computing the Euclidean distance $\|\nu - Em(\mathcal{L})\|$ for each song $\mathcal{L} \in$ LKB. The closest k songs are selected, with k varying as a function of the length of the song to be generated (generally 3000-4000 songs). The music knowledge base (MKB) comprises 6,673 pop music lead sheets from the Wikifonia dataset. Each contains chords, melody, and lyrics in Music XML format. Lyrics $\mathcal{L}$ for each song are annotated with an Empath vector $Em(\mathcal{L})$. Pop* selects a subset of the MKB (generally 75-150 songs) for training using the same Euclidean distance method as LKB subselection. The MKB is used to train chord, pitch, and rhythm models. For details on the preparation, parsing, and training from these knowledge bases, see (Bodily and Ventura 2022).

Pop* also possesses perceptual faculties for detecting musical structure in symbolic music (i.e., Music XML files) modeled on the way humans detect structure when listening to music through a process of mentally aligning musical subsequences that share similar chord, pitch, rhythm, or lyric features. Structural alignments in a single viewpoint are called *motifs*. Structural alignments across multiple viewpoints create *sectional form* (e.g., a *chorus* results from overlapping repeats of lyrics, pitch, rhythm, and chords).

Alignment is performed using a multi-Smith-Waterman (mSW) self-alignment algorithm which aligns a song against itself for each viewpoint and finds all unique local alignments with scores above some threshold. The scoring function used to calculate alignments considers different features for each viewpoint. Weights for the scoring function (as well

as the threshold) are learned *a priori* using a genetic algorithm trained on a small, structurally annotated subset of the MKB. See (Bodily and Ventura 2021) for details.

For each new composition, Pop* selects at random a song in the MKB after which to structurally model the new composition. The song's length determines the length n of the new composition needed to instantiate NHMMs N_c, N_p, and N_r for approximating $P(\eta|\nu,\tau)$, $P(\phi|\nu,\tau,\eta)$, and $P(\rho|\nu,\tau)$. Pop* performs an mSW-alignment on this training song for each viewpoint. Each viewpoint-specific alignment defines a set of binary constraints C_2 that is used to define NHMMs N_c, N_p, N_r, N_l.

Characteristic #6: Autonomy

Autonomy is commonly cited as one of the most essential yet difficult characteristics to achieve in computational creative systems (Mumford and Ventura 2015). Autonomy is not a binary characteristic in creative systems but rather manifests itself along a spectrum, with the most creative systems exhibiting autonomy through self-evaluation, the injection of knowledge "without leaving the injector's fingerprints," and (at the acme) the ability to use perceptual abilities to self-improve the system (Ventura 2016). In short, the more ways in which a system can exert autonomy, the greater its ascribed creativity (Colton, Pease, and Charnley 2011). Jennings defines three criteria required for a system to be autonomous: autonomous evaluation, autonomous change, and non-randomness (Jennings 2010). The first two refer to an ability to independently decide how well a creation appeals to standards and to independently initiate and guide changes to these standards. The third does not mean a system cannot employ randomness, but rather that decisions should *generally* reflect that the system operates (independently) on a basis of persistent standards.

Pop* implements autonomy in three principal ways: in choosing an intention, in choosing a relevant knowledge (sub)base, and in self-evaluation. Pop* derives an intention ν from a tweet $\tilde{t}$ where the probability of choosing $\tilde{t}$ varies according to its relation to the system's aesthetic A. Besides having dictated A, humans play no role in the selection of $\tilde{t}$. Pop* self-selects a training set from its knowledge base that relates to its intention ν. In the self-evaluation process, the decision of whether or not to keep a song is determined by how well the composition semantically achieves its intention ν. Note that of the three, the latter two rely on the system's intention, underscoring the important role that intention plays generally in enabling autonomy.

Characteristic #7: Self-Evaluation

Self-evaluation constitutes more than the ability to reflect on the novelty, typicality, and aesthetic value of its output; it also includes evaluation of how well the output achieves the system's goals (Ventura 2016). Self-evaluation occurs without consulting outside opinions and thus presupposes autonomy (Jennings 2010). (Seeking outside opinions is an important guide by which to initiate and change evaluative standards, but a system should maintain and apply a standard in self-evaluation that is distinct from those of other creative agents (Ackerman et al. 2017).) Self-evaluation can occur *post hoc* (e.g., in systems that filter) or via a "baked-in" approach where the system's notion of what is good is inherent in generative subprocesses. This can go so far as allowing subprocesses to be aware of and influence one another during generation (Linkola et al. 2017). Over time, self-evaluation can fine-tune the generative process either through direct modification of the generative model or through the addition of generated artifacts to the system's knowledge base (Pérez y Pérez and Sharples 2004).

For a given intention vector ν Pop* is given 6 hours to compose up to 10 candidate compositions using the same training sets (i.e., Markov models M) but each with a potentially unique structure (i.e., lengths n and binary constraint sets C_2). Pop* applies a scoring function $S(\gamma)$ to a composition $\gamma = (\nu, c, p, r, l)$ with intention ν, chords c, pitches p, rhythms r and lyrics l:

$$S(\gamma) = \delta + \sigma + \mathcal{E} + R$$

where $\delta = \|\nu - Em(l)\|$; $\sigma = |uniq(l)|/|l|$ (i.e., the ratio of the unique word count to the total word count); $\mathcal{E} = \sum_1^{|p|-31} |uniq((p_i, \ldots, p_{i+31}))|$ (i.e., the average number of unique pitch values in p in a 4-measure sliding window); and $R = |\{p_i | p_i \in \text{MIDI } [60, 76]\}|/|p|$ (i.e., the fraction of $p_i \in p$ for which the MIDI value of p_i is in the range [60,76]). δ assesses how well the lyrics l reflect the system's intention ν. σ protects against overly repetitive lyrics. As a way of measuring "catchiness" or managing boredom, $\mathcal{E}$ represents a collective density value (Eigenfeldt et al. 2017)). R represents the sing-ability of the melodic pitch sequence. The candidate γ with the highest self-evaluation score $S(\gamma)$ is output.

External Evaluation of Creative Characteristics

Ritchie suggests that a system's creativity may (and possibly should) be assessed through the external artifact—without any knowledge of the system's process (Ritchie 2007). In contrast, others suggest that consideration of the system's process is essential (Kasof 1995; Boden 2004; Colton 2008; Ventura 2016; Colton, Pease, and Charnley 2011). We evaluate Pop* in both modalities, using Jordanous' SPECS framework (Jordanous 2012) which requires stating one's definition of what it means to be creative (as done above) and then creating and implementing assessments to measure to what extent that creativity has been achieved.

We conducted an evaluation in the form of an online Qualtrics survey of 125 people. In each survey, the system presents itself: "Hi, my name is Pop*! I am a computer system that composes pop, rock, and show tune music. I read a lot on Twitter. When I find a tweet that I like, then I compose music." The survey then proceeds in two stages: an evaluation based solely on external artifacts and an evaluation based on an informed understanding of Pop*'s process.

Evaluation Based on Artifacts The survey first invites the participant to listen to and evaluate two original Pop* compositions. Compositions for evaluation were chosen completely at random from Pop*'s output. Twelve unique compositions were included in the evaluation. These songs

include significant variation in chords, melody, lyrics, length, structure, tempo, modality, key, and intention.

For each composition, the system's generated description of the piece is presented along with an audio recording of the song. After reading the description and listening to the audio, the participant is asked seven Likert scale questions:

1. How would you **rate** this composition overall?
2. How would you rate the **lyric composition** in this piece?
3. How would you rate the **music composition** in this piece?
4. How would you rate the global **structure** (i.e., form, layout) in this piece?
5. How **typical** is this song of pop/rock/show tunes music?
6. How **novel** is this song compared to other pop/rock/show tunes music?
7. How well did Pop* communicate its **intention** (X and Y) through the music?

Each question is rated on a scale from 1 to 5, the first four as star ratings with no labels (half stars allowed) and the remaining three as Likert scales which label only the lowest and highest ratings in defining the spectrum (Likert 1932).

Evaluation Based on Process Upon completion of the artifact-based evaluation, the participant is then introduced to the process of Pop*. This introduction consists of a brief description accompanied by a simplified version of Figure 1. The simplified version uses the labels "Tweets", "Interests", "Intention", "Lyrics", "Sheet Music", "Generation", "Evaluation", and "Output" in place of the pertinent labels in Figure 1 to avoid overtly biasing answers. The participant is then given 5 Likert scale questions:

8. How convinced are you that Pop* internally represents **knowledge** of music?
9. How convinced are you that Pop* has an **opinion**, belief, or attitude about what makes music "good"?
10. How much **autonomy** would you say Pop* has to make decisions on its own?
11. How good is Pop* at **self-evaluation** (i.e., judging its own success)?
12. How would you rate the **creativity** of Pop*

We also invited the participant to explain via free response their answer to question 12, to add other comments, and to provide some demographic information to control for biases. Of 125 respondents, 68 (54.4%) considered themselves musicians; 77 (61.6%) reported knowing the system designers personally; 53 (42.4%) believed "Absolutely" that computers are capable of creativity, 6 (4.8%) believe they "Never" will be, and 66 (52.8%) were "Not sure".

Results

Results of the survey are found in Tables 2-4. The "Best Song" reported in Table 2 refers to the song with the highest average ratings for question Q1 (overall rating) across all participants[1]. All scores are out of 5 with 1 representing the minimum score possible.

[1]*And I think I* took 10th place in the 2022 AI Song Contest

Familiarity Bias Fundamental to the notion of achieving creativity in computational systems is that they are deemed to be creative by *unbiased* observers. Although there is often bias against computers being creative, *familiarity* may introduce a bias *towards* ascribing creativity.

We find that there is a noticeable drop in scores between the group of survey participants who responded "Yes" and those that responded "No" to knowing one of the system designers personally, both in the overall average scores for each question, but also in the highest song-specific averages (see Table 3). There was not a noticeable skew in beliefs about the ability of computer creativity among the group not familiar with the system designers.

Even given the familiarity bias, we see the same general trends in which aspects of creativity participants are most willing to ascribe to Pop*. Novelty scores (Q6) have the highest average and max scores for both groups. Likewise both groups were least impressed by the system's lyric abilities (Q2). Significant to our work on incorporating structure is the fact that both groups have elevated ratings for the structure in Pop* compositions (Q4).

Generation Clearly the system generates artifacts. The question that remains is whether these artifacts are novel (Q6). The average and max novelty scores for both musicians and non-musicians were the highest scores for any of the characteristics of creativity as evaluated solely based on artifacts. One respondent (not familiar with the system designers) responded that "Creativity is the ability to create new things. Pop* can clearly create new things."

There were also some participants that did *not* feel Pop* expressed novelty. One respondent wrote that the songs (s)he heard "did not strike me as very original and creative. Sorry." This criticism may be in part a reflection of the two songs that this participant was randomly assigned. Pop*'s perceived novelty might be improved by constraining a distance measure between output songs.

Knowledge Representation Most groups felt that Pop* had some internal representation of knowledge (Q8). Even musicians—who by definition likely possess such a representation themselves—rated Pop* (slightly) above average. Those unfamiliar with the system designers rated knowledge representation as slightly below average.

Related to the notion of process, one respondent wrote that "combining ideas and generating from them something completely new is the epitome of creativity, regardless of whether or not the process was automated. Thus, even though the machine is a machine it does mimic, in some way at least, some form of creativity."

Intentionality Intentionality (Q7) received relatively low scores, particularly from the group of respondents unfamiliar with the system designers, who rated the intentionality of Pop* well below average .

Several comments were made to explain these low intentionality ratings. One respondent wrote: "Pop* [sic] seemed to always start with something in mind but delivered something different." We assume that this comment is referring to the way that Pop* explicitly compares the topics of its

	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Max Score	*3.20*/3.25	*2.63*/3.00	*3.50*/3.54	*3.38*/3.33	*3.14*/3.50	*3.77*/3.76	*3.40*/3.67
Avg Score	*2.66*/2.81	*2.09*/2.29	*2.98*/3.04	*2.75*/2.91	*2.86*/2.90	*3.16*/3.28	*2.88*/3.02
Min Score	*2.20*/2.48	*1.50*/1.83	*2.54*/2.67	*2.33*/2.50	*2.63*/2.50	*2.60*/2.95	*2.47*/2.48
Best Song	*2.71*/3.25	*2.29*/3.00	*2.71*/3.13	*3.00*/3.13	*3.14*/3.50	*3.57*/3.31	*2.86*/3.13

Table 2: *Evaluation Based on Artifact*. Average ratings for each song: average musicians' rating in italics followed by the average of all ratings. "Best Song" is the song with the highest overall rating score (Q1) across all participants.

		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Knows us	Survey Avg	2.93	2.36	2.93	3.00	2.64	4.00	3.00	3.53	2.87	3.40	2.88	3.56
	Song Max	3.40	3.11	3.73	3.53	3.33	4.00	4.08					
Doesn't know us	Survey Avg	2.00	2.00	2.20	2.40	2.80	2.80	2.00	2.90	2.60	3.04	2.48	3.08
	Song Max	3.29	2.88	3.22	3.33	3.71	3.78	3.20					

Table 3: *Familiarity Bias*. Scores by participants' familiarity with study designers. Averages are over all songs. Max is the highest-rated song. The system was typically rated more creative by those familiar with the system designers.

intention with the topics it perceives in its generated song and how these often tend to be different topics. Based on this assumption, this comment is valuable because it may suggest one way that at least some audiences perceive intentionality: assessing the success of the system in meeting its goals is based, at least partially, on whether or not the *system* believes it has met its goals. This is surely not universally the case; many observers likely autonomously assess the system's success in meeting its goals. However, finding ways to emphasize aspects in which the system believes it has achieved its intention may in some cases improve the perception of intentionality.

Some observers felt that the system had some intentionality: "good at finding ideas and creating sounds to match those ideas and feelings." This may result from the fact that some songs exhibit intentionality better than others.

Aesthetic The presence of aesthetic (Q9) is a quality Pop* currently struggles to exhibit. This attribute scored third lowest across the general survey population as well as across the musical subpopulation. Among participants not biased by familiarity, the question of whether or not Pop* has an opinion or belief was also answered with ratings below average (though not far below those of the general survey group).

One of the challenges in creating a convincing model of aesthetic is that unlike novelty, or even intention, aesthetic gets at the heart of what many people see as the fundamental difference between computers and humans. As one respondent put it: "It's a machine, it has no autonomy or beliefs." Conversely, others were willing to give Pop* some credit: "there were some sentiments communicated" and "Pop*'s music sounds . . . different enough to have its own style."

Domain Knowledge Typicality (Q5) was the only characteristic for which the group unbiased by familiarity rated Pop* higher than the group biased by familiarity. This could be explained by the fact that this word can have a negative connotation outside of CC, and thus the bias for this question might be expected toward lower scores. Either way, we consider elevated scores for this attribute to be a good sign, especially given the parallel trends in the novelty score.

Also related to domain knowledge were questions Q2, Q3, and Q4, which rated Pop*'s lyric model, music model, and structure model respectively. Ratings for lyrics were universally the lowest ratings for any question, with musicians being particularly critical. Although Pop* has focused relatively little on semantic cohesiveness of lyrics, it has focused a great deal on how lyrics should be effectively combined with music, e.g., to emphasize appropriate stresses and to effectuate rhyme. This subtlety was observed by one commenter that wrote: "To put together lyrics to match the music . . . takes a lot of creativity."

Overall *music* scores (Q3) were, on the other hand, among the highest for any of the questions, suggesting that the chord, pitch, and rhythm models are learning and applying knowledge effectively. Many made comments such as "the generated chord structure of the two songs impressed me."

Structure scores (Q4) were generally below average, although there were a few songs that earned average ratings above 3.0. This is an area that would benefit from further assessment about what kinds of impact the structure may or may not be having (e.g., does structure make the song easier to remember or to get stuck in one's head?).

One respondent remarked on the importance of leveraging domain knowledge, explaining that their attribution of creativity to Pop* was based on its perceived ability "to create songs based off of lots of different information and still make it sound appealing and applicable."

Autonomy Autonomy (Q10) scored above average in nearly every group, including those not biased by familiarity with the system's designers (the one exception was those with the belief that computer creativity is impossible). Autonomy was rarely the highest scoring characteristic, and several of the comments against the creativity of the system were aimed at the absence of autonomy. For example, "Pop*' [sic] follows the formulas really well, but the ideas are not innovative, they are formulaic", or, "Pop [sic] follows a sequence of preprogrammed algorithms. I would therefore

		Musicianship		Belief about computer creativity		
	All Groups	Musicians	Non-Musicians	Believers	Skeptics	Unbelievers
Q8	3.29	3.07	3.54	3.38	3.29	2.50
Q9	2.77	2.59	2.98	2.94	2.67	2.33
Q10	3.26	3.21	3.33	3.43	3.20	2.5
Q11	2.73	2.56	2.93	2.79	2.70	2.5
Q12	3.38	3.25	3.53	3.36	3.40	3.33

Table 4: *Evaluation Based on Process*. Average scores for survey questions Q8–Q12, broken down by demographic.

describe the output of Pop* as being more representative of the creativity of its designer as well as the thoughts/moods of the people giving input tweets."

Mumford and Ventura reported similar responses in their work to assess the autonomy of CC systems (Mumford and Ventura 2015). However, in their work, what respondents thought was a computer being creative was actually a *human* being creative. There exists a preconceived notion that regardless of what occurs, "It's a machine, it has no autonomy." To this point Mumford and Ventura suggest that "even creative humans could be argued to be following a complex set of chemical and psychological instructions" (Mumford and Ventura 2015). An interesting future study might examine how changing people's perception about their *own* autonomy would impact their perception of computer autonomy.

Self-Evaluation The collective survey group rated the self-evaluative abilities of Pop* (based on an understanding of its process) lower than any other property of the system except its lyric generation (Q11). This may result from the general perception that Pop* consistently composes below-average music (per responses to Q1). It may be hard to ascribe self-evaluative abilities to a system that can't generate cohesive English. It may also stem from the way that "Pop* seemed to always start with something in mind but delivered something different".

Q11 was included as an assessment based on Pop*'s *process* under the assumption that a knowledge of the *process* is critical to assessing whether the system uses a reasonable self-evaluation method. However, it may be that the question of efficacy in self-evaluation is better suited as an *artifact*-based evaluation, if attributing success in self-evaluation depends on the reviewer *agreeing* with the individual evaluations that the system makes rather than agreeing with the process it uses to make those evaluations.

Like creativity, each characteristic attribute occurs on a spectrum. The results of the evaluation point out opportunities for improvement; however, the results also suggest that, at least to some extent, Pop* possesses each of the characteristics necessary for creativity.

Creativity Assessing any one of the above characteristics of creativity may prove to be just as difficult as directly asking the question of interest: "Is the system creative?" In asking this question, we were careful to explicitly differentiate the creativity of the system and that of its designers. Responses to this question (Q12) in the survey reported the highest scores across all groups: those familiar with the designers, those not familiar, musicians, non-musicians, believers, skeptics, and unbelievers (novelty score for those familiar with the designers, which was higher than that of the group's creativity score, is the only exception). Scores across all groups were above average.

The high ratings for the direct question of Pop*'s "creativity" is quite remarkable. It may suggest that, despite a relative lack of the other characteristics examined, creativity *can* exist. Perhaps it might suggest the wrong criteria have been chosen, and that if we had chosen the right criteria then the scores for those criteria would have been a better reflection of the scores for creativity. However, there seems a more probable explanation, which is that creativity (Q12) is greater than the sum of its individual characteristics. In other words, the perception of creativity in computational systems can exceed the perception of individual creative characteristics *when these characteristics are found together*.

Discussion

Still in its early stages, CC has largely been defined by systems that exhibit one or a few creative characteristics. This is understandable considering that endowing or augmenting a system with any one of these characteristics is a non-trivial task and that there are cases where it might be desirable to focus exploration on a single characteristic. It is important, however, that we embrace the challenge and keep an eye on the goal of holistic computational creativity—the idea that creativity most effectively emerges from the confluence of the set of creative attributes. A system that possesses any one of the characteristics of creativity is just that: a system that possesses one of the characteristics of creativity. For such systems, it is easy for critics to focus on how the system is *not* creative. Conversely, by simply possessing *some* level of many creative attributes a system may reasonably be seen as creative, particularly among non-specialists.

We have identified seven common, fundamental characteristics that we believe to be necessary (though perhaps not sufficient) for creativity. We have demonstrated the joint application of these characteristics in an applied example of a music composition system. We have externally evaluated this system for these characteristics through the use of a combination of artifact-based and process-based assessments. Our findings from these assessments suggest that the system does possess the characteristics of creativity to varying extents, but that more importantly, the system overall is perceived to be creative. This suggests that holistic CC represents a promising direction for future work in our efforts to expand the "final frontier" of artificial intelligence.

References

Ackerman, M.; Goel, A.; Johnson, C. G.; Jordanous, A.; León, C.; Pérez Pérez, R.; Toivonen, H.; and Ventura, D. 2017. Teaching Computational Creativity. In *Proceedings of the Eighth International Conference on Computational Creativity*, 9–16.

Barbieri, G.; Pachet, F.; Roy, P.; and Esposti, M. D. 2012. Markov constraints for generating lyrics with style. In *Proceedings of the Twentieth European Conference on Artificial Intelligence*, 115–120. IOS Press.

Birkhoff, G. D. 1933. *Aesthetic Measure*. Harvard University Press Cambridge.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms: Second Edition*. Routledge.

Bodily, P. M.; and Ventura, D. 2018. Explainability: An Aesthetic for Aesthetics in Computational Creative Systems. In *Proceedings of the Ninth International Conference on Computational Creativity*, 153–160.

Bodily, P. M.; and Ventura, D. 2021. Inferring Structural Constraints in Musical Sequences via Multiple Self-Alignment. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, volume 43, 1112–1118.

Bodily, P. M.; and Ventura, D. 2022. Steerable Music Generation which Satisfies Long-Range Dependency Constraints. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 5(1).

Colton, S. 2008. Creativity Versus the Perception of Creativity in Computational Systems. In *Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium: Creative Intelligent Systems*, volume 8.

Colton, S.; Pease, A.; and Charnley, J. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the Second International Conference on Computational Creativity*, 90–95.

Colton, S.; Pease, A.; Corneli, J.; Cook, M.; Hepworth, R.; and Ventura, D. 2015. Stakeholder groups in computational creativity research and practice. In *Computational Creativity Research: Towards Creative Machines*, 3–36. Springer.

Colton, S.; and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Proceedings of the Twentieth European Conference on Artificial Intelligence*, 21–26. IOS Press.

Csikszentmihalyi, M. 1997. Flow and the Psychology of Discovery and Invention. *HarperPerennial, New York*, 39.

Eigenfeldt, A.; Bown, O.; Brown, A. R.; and Gifford, T. 2017. Distributed musical decision-making in an ensemble of musebots: Dramatic changes and endings. In *Proceedings of the Eighth International Conference on Computational Creativity. Association for Computational Creativity*, 88–95.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 4647–4657. ACM.

Glines, P. W. 2022. *Imposing Structure on Generated Sequences: Constrained Hidden Markov Processes*. Ph.D. thesis, Idaho State University.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the "Why?" in Computational Creativity: A Non-Anthropocentric, Minimal Model of Intentional Creative Agency. In *Proceedings of the Eighth International Conference on Computational Creativity*, 128–135.

Hadjeres, G.; and Nielsen, F. 2018. Anticipation-RNN: Enforcing unary constraints in sequence generation, with application to interactive music generation. *Neural Computing and Applications*, 1–11.

Hofstadter, D. R. 1980. *Godel, Escher, Bach: An Eternal Golden Braid*. New York: Vintage Books.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines*, 20(4): 489–501.

Jordanous, A. 2012. A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative. *Cognitive Computation*, 4(3): 246–279.

Jordanous, A. 2014. Stepping Back to Progress Forwards: Setting Standards for Meta-Evaluation of Computational Creativity. In *Proceedings of the Fifth International Conference on Computational Creativity*, 129–136.

Kasof, J. 1995. Explaining creativity: The attributional perspective. *Creativity Research Journal*, 8(4): 311–366.

Koren, L. 2010. *Which "Aesthetics" Do You Mean? : Ten Definitions*. Point Reyes, California: Imperfect Publishing.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140: 54–55.

Linkola, S.; Kantosalo, A.; Männistö, T.; and Toivonen, H. 2017. Aspects of Self-awareness: An Anatomy of Metacreative Systems. In *Proceedings of the Eighth International Conference on Computational Creativity*, 189–196.

Mothersill, M. 2004. The Blackwell Guide to Aesthetics. In Kivy, P., ed., *Beauty and the Critic's Judgment: Remapping Aesthetics*, 152–166. Blackwell Publishing Ltd.

Mumford, M.; and Ventura, D. 2015. The man behind the curtain: Overcoming skepticism about creative computing. In *Proceedings of the Sixth International Conference on Computational Creativity*, 1–6.

Pachet, F.; Roy, P.; and Barbieri, G. 2011. Finite-length Markov processes with constraints. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 635–642.

Papadopoulos, G.; and Wiggins, G. A. 1999. AI Methods for Algorithmic Composition : A Survey, a Critical View and Future Prospects. *Artificial Intelligence and Simulation of Behaviour Symposium on Musical Creativity*, 110–117.

Pérez y Pérez, R.; and Sharples, M. 2004. Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge-based systems*, 17(1): 15–29.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1): 67–99.

Shannon, C. E. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1): 3–55.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; and Lanctot, M. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587): 484–489.

Ventura, D. 2016. Mere Generation: Essential Barometer or Dated Concept? In *Proceedings of the Seventh International Conference on Computational Creativity*, 17–24.

Ventura, D. 2017. How to Build a CC System. In *Proceedings of the Eighth International Conference on Computational Creativity*, 253–260.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems*, 19(7): 449–458.