Preference Models for Creative Artifacts and Systems

Debarun Bhattacharjya
Cognitive Computing Research
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598 USA
debarunb@us.ibm.com

Abstract

Although there is vigorous debate around definitions of creativity, there is general consensus that creativity i) has multiple facets, and ii) inherently involves a subjective value judgment by an evaluator. In this paper, we present evaluation of creative artifacts and computational creativity systems through a multiattribute preference modeling lens. Specifically, we introduce the use of multiattribute value functions for creativity evaluation and argue that there are significant benefits to explicitly representing creativity judgments as subjective preferences using formal mathematical models. Various implications are illustrated with the help of examples from and inspired by the creativity literature.

Introduction

Computational creativity (henceforth CC) is an interdisciplinary field that studies the role of computers in the creative process. Several success stories in domains such as visual art, music, literature, humor, science and mathematics have already been noted (Buchanan 2001; Cardoso, Veale, and Wiggins 2009; Colton and Wiggins 2012). One of the issues that has plagued the research community, analogous to the broader field of artificial intelligence, is around pinning down what it means to be creative. This is of course not surprising as creativity is often considered the pinnacle of human intelligence. Taylor (1988) summarizes several early definitions of creativity from the psychology literature.

Although there is significant debate around defining creativity, there appears to be general agreement about at least two aspects. First, creativity seems to involve 'novelty and more', i.e. there are multiple facets to creative artifacts and it is not enough to only be original to be considered creative. Second, creativity is a subjective judgment that is not meaningful without an evaluator of creative value.

In this paper, we frame evaluation in CC and creativity studies in general through a **multiattribute** preference modeling lens that embraces the subjectivity that is inherent in judging creative value, while effectively capturing the notion of creativity involving multiple facets. According to this framework, a (human or machine) evaluator's preferences are modeled using **preference functions**. There is a rich literature in formal mathematical models of preferences, mainly in fields that pertain to prescriptive decision making, and we believe there are significant benefits from applying these techniques to the field of CC.

The preference modeling framework is applied to evaluate artifacts as well as CC systems, and we explore the implications of various modeling assumptions through illustrative examples. For instance, creativity researchers and practitioners should be aware of the implications of taking weighted averages of scores along multiple criteria – we argue that such an approach need not always be appropriate in CC, or at the very least, the underlying assumptions should be appreciated. We begin by motivating our research effort.

Evaluating Creative Value

Evaluation, evaluation, evaluation! Evaluating creative artifacts and systems that produce them is to the field of CC what location is to real estate. We distinguish between evaluation of creative artifacts vs. CC systems, like in Pease and Colton (2011). In this section, we provide some background to motivate a multiattribute preference view of creativity.

Novelty and More

Attempts at defining creativity or towards identifying its properties can be seen in the early literature on artificial intelligence. Newell et al. (1958) opined that creative products needed to have novelty and value. Boden (1990) echoed this thought, suggesting that creativity involves generating ideas that are both novel and valuable. Mayer (1999) refers to these two facets as originality and usefulness, citing several alternatives for the latter term, including utility, adaptiveness, appropriateness and significance.

We will avoid the terms 'value' and 'utility' in how they have been used in the creativity literature because we reserve them for specific concepts from the preference modeling domain. (Our notion of creative value includes novelty.) Instead, we will use the term 'quality' to refer to non-novelty related aspects of creative artifacts (Ritchie 2001; Pease, Winterstein, and Colton 2001). Evaluation could in general involve several (> 2) attributes that sufficiently span novelty/originality as well as quality/usefulness.

Subjectivity in Evaluation

Various definitions of creativity explicitly acknowledge the relationship between the creator/creation and an observer (Wiggins 2006) and how a creative artifact must be

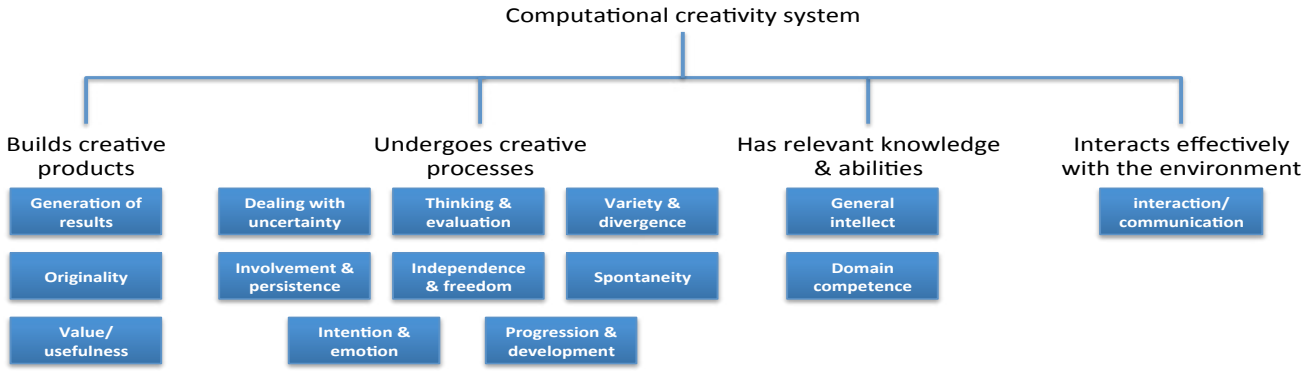


Figure 1: An example of objectives for a computational creativity system, with attributes from Jordanous (2012).

judged or deemed to be creative (Sawyer 2012). The line of research that deals with creativity assessment goes back at least around a hundred years (Cattell, Glascock, and Washburn 1918) and includes popular methods like the consensual assessment technique where artifacts are rated by two or more experts in the field (Amabile 1982).

Assessing a creative artifact is necessarily a cognitive task that involves several complex processes (Varshney et al. 2013). In this paper, we attempt to model an evaluator's (subjective) preferences for a creative artifact – one approach to understanding these preferences is directly through an overall creativity rating; another is to break down the assessment along multiple attributes. We consider the second approach and adopt the view that the evaluator's preferences can be captured, at least approximately, through some abstract mathematical representation.

Multiattribute Preference Models: A Review

Decision analysis is an approach to prescriptive decision making that applies the norms of decision theory to practical decision situations, tracing its roots to stalwarts such as Bernoulli, Laplace and Pascal. The field explicitly models decision makers' subjective beliefs and preferences; the latter aspect has spawned the overlapping fields of multiattribute utility theory (MAUT) and multi-criteria decision making (MCDM). Here we provide a brief summary of relevant concepts, mainly using terminology from the seminal Keeney and Raiffa (1976), but the reader is encouraged to peruse related work (Belton and Stewart 2002; Wallenius et al. 2008).

Preliminaries

A formalization of preferences is useful in the spirit of breaking a larger unmanageable problem into smaller pieces. An **objective** indicates the 'direction' in which one strives to do better. It is often convenient to generate objectives by organizing them in a hierarchy, thereby meaningfully structuring them. **Attributes** (or criteria) are measures that adequately determine how well an objective has been met. Figure 1 shows an example objectives hierarchy (with only one level) for a CC system: four high-level objectives are spanned by fourteen attributes from Jordanous (2012).

There are several desirable properties of attributes: they should be *complete* (sufficiently capture degree to which objectives are met), *operational* (meaningful and understandable), *nonredundant* (double counting should be avoided) and *minimal* (with manageable problem dimension).

Making a judgment about how much to give up on an objective for another is the essence of a trade-off, and this is represented by a functional form over attributes. There are two types of preference functions: **utility functions** and **value functions**, also referred to as cardinal and ordinal utility functions: the distinction between them is whether uncertainty is involved or not, respectively. A utility function thus captures both a decision maker's strength of preference as well as their attitude towards risk. In this paper, we focus solely on value functions as all of the situations that are studied are those of certainty, but the concepts apply broadly. Also, we do not expound upon how preference functions are elicited/assessed here – instead, we refer the reader to the literature (Keeney and Raiffa 1976; von Winterfeldt and Edwards 1986).

Value functions

Let $X_1, \ldots, X_M$ be a set of M attributes where lower case x_i denotes the score/consequence along attribute X_i. We use the notation $\bar{X}_i$ to refer to the complement set of attributes to X_i, and x_i^* and x_i^0 for the best and worst scores of attribute X_i. A **preference structure** is defined over the domain of attributes if all points in the domain are comparable and no intransitivities exist. In that case, if $\mathbf{x} = \{x_1, \ldots, x_M\}$ and $\mathbf{y} = \{y_1, \ldots, y_M\}$ are two alternatives, then a (measurable) value function $v(\cdot)$ (Dyer and Sarin 1979) is one such that $v(\mathbf{x}) \geq v(\mathbf{y})$ if and only if $\mathbf{x} \succeq \mathbf{y}$, where the symbol $\succeq$ reads 'preferred or indifferent to'.

A preference structure over attributes is determined by **indifference curves**, i.e. complete sets of points in the domain of attributes where the decision maker is indifferent. It can be shown that monotonic transformations of value functions do not change the preference structure, and it is often convenient to normalize value functions between most and least preferred alternatives such that $v(\mathbf{x^0}) = 0$ and $v(\mathbf{x}^*) = 1$.

Additive value functions The simplest and most widely used value function is the **additive** function, of the form:

$$v(x_1, \ldots, x_M) = \sum_{i=1}^{M} \lambda_i v_i(x_i), \quad (1)$$

where $\lambda_i \geq 0 \ \forall i$ are the weights, $\sum_{i=1}^{M} \lambda_i = 1$, and $v_i(\cdot)$ are marginal (one-dimensional) value functions bounded between 0 and 1. The additive value function is thought to be fairly robust (Stewart 1996) and is extremely popular in practice, but unfortunately it is often misused. This is because its proponents often forget or are unaware that it applies if and only if there is **mutual preferential independence** among attributes (for $M > 2$). This condition holds if every $\mathbf{Y} \subset \{X_1, \ldots, X_M\}$ is preferentially independent of its complement, i.e. the preference structure over $\mathbf{Y}$ does not depend on the scores of the attributes in the complement set. The following example illustrates the implications.

Example 1. [Evaluating cartoon captions]
Sternberg et al. (2006) performed creativity assessments on captions provided by students for cartoons from the New Yorker. These were adjudicated based on three attributes: cleverness, humor and originality, all scored on a 5 point scale, and the total creativity score was computed by summing up the three scores. From a preference modeling perspective, this procedure implicitly assumes that preferences for these cartoon captions, when viewed as creative artifacts, follow an additive value function. This in turn implies mutual preferential independence among attributes. We pose the question – is this condition appropriate?

Consider indifference curves for humor and originality, given a cleverness score. If the score on cleverness is high, then it seems reasonable that the evaluator may generally be willing to give up a large score of humor for some originality. The rationale behind this claim is that the evaluator may view cleverness and humor as two attributes for a higher level objective (quality), and may be willing to compensate one for the other. On the other hand, if the cleverness score is low, the evaluator may no longer be willing to give up as much humor for the same score increment on originality.

These (hypothetical) preferences are clearly inconsistent with mutual preferential independence, which in a three attribute problem enforces identical indifference curves for every pair of attributes, conditional on the score of the third attribute. This would make the additive value function inappropriate in this case. Our conjecture is that preferences of this sort are probably commonplace for creative artifacts. Understanding the preferential assumptions behind implicit functional forms could be broadly beneficial to the creativity community. □

Value copulas The value function $v(\cdot)$ can take any form, including one that does not need to subscribe to independence assumptions. A recent approach to modeling potentially complex preference functions is that of **copulas**, and although they were introduced to model utility functions (Abbas 2009), they can also be used for value functions. A copula $C_\lambda(z_1, \ldots, z_M)$ is a multivariate function that is a continuous mapping from the hypercube $[0,1]^M$ to the interval $[0,1]$, normalized such that $C(\mathbf{0}) = 0$ and $C(\mathbf{1}) = 1$ (Sklar 1959). It is non-decreasing in each of its arguments z_i, and for each argument, there exists some reference scores of the complement attributes for which it is an affine function. A value function can be constructed from a copula as follows:

$$v(x_1, \ldots, x_M) = C_\lambda \left(v_1(x_1|\bar{x}_1^{\lambda_1}), \ldots, v_M(x_M|\bar{x}_M^{\lambda_M}) \right), \quad (2)$$

where $C_\lambda(\cdot)$ is a copula and $v_i(x_i|\bar{x}_i^{\lambda_i})$ are normalized conditional value functions, defined as:

$$v_i(x_i|\bar{x}_i^{\lambda_i}) = \frac{v_i(x_i, \bar{x}_i^{\lambda_i}) - v_i(x_i^0, \bar{x}_i^{\lambda_i})}{v_i(x_i^*, \bar{x}_i^{\lambda_i}) - v_i(x_i^0, \bar{x}_i^{\lambda_i})}, \quad (3)$$

with $\bar{x}_i^{\lambda_i}$ denoting a particular reference score for the set of complement attributes to X_i. Assessing a conditional value function therefore entails determining the marginal rate of value for an attribute when all other attributes are set to some reference scores. It is typical to assess this function at the complementary maximum ($\bar{x}_i^*$) or minimum ($\bar{x}_i^0$) scores.

The model of equation (2) represents any value function that is continuous, bounded, non-decreasing in each argument, and strictly increasing with each argument for at least one reference score of the complement attributes. The power of the copula is that like the additive function, it models a high-dimensional function by aggregating one-dimensional functions, yet allows for a much wider class of functions.

An example of a copula where conditional value functions are assessed at the maximum scores of the complement of each attribute is the extended Archimedean copula:

$$E(z_1, \ldots, z_M) = a\psi^{-1}\left[\prod_{i=1}^{M} \psi(l_i + (1 - l_i)z_i)\right] + b, \quad (4)$$

where $l_i \in [0,1)$, $a = 1/\left(1 - \psi^{-1}\left[\prod_{i=1}^{M} \psi(l_i)\right]\right)$, $b = 1 - a$, and the **generating function** ψ has the same mathematical properties as a strictly increasing cumulative probability distribution function. A special case occurs when $l_i = 0 \ \forall i$ and the generating function is linear, $\psi(z_i) = z_i$, resulting in the **multiplicative** form:

$$v(x_1, \ldots, x_M) = \prod_{i=1}^{M} v_i(x_i|\bar{x}_i^*). \quad (5)$$

Both additive and copula value functions will be applied in subsequent sections.

A Two-Attribute Model for Artifacts

As we highlighted earlier, maximizing novelty (or originality) as well as quality (or usefulness) are reasonable high-level objectives for creative artifacts. The simplest preference model therefore involves two attributes: novelty X_N and quality X_Q. In this section, we first discuss some potential value functions over these two attributes and then consider a scenario that explores the implications of potentially mis-characterizing a user's value function.

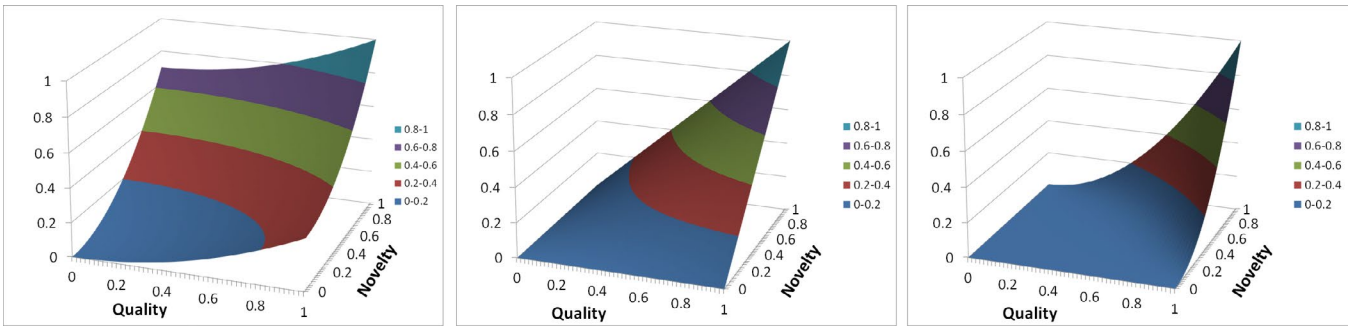


Figure 2: Some example value functions over novelty and quality: Left) additive with $\lambda = 0.7$ and power function marginals, $\beta_N = \beta_Q = 2$; Middle) multiplicative with linear conditionals at maximum reference points; Right) copula with exponential generating function, $\delta = -5$, and power function conditionals at maximum reference points, $\beta_N = \beta_Q = 2$.

Value functions for novelty and quality

An evaluator's preferences for an artifact with novelty x_N and quality x_Q can be represented by $v(x_N, x_Q)$. If the attributes are bounded, then they can be normalized to lie in $[0, 1]$. Assuming that $v(\cdot)$ is bounded, and since value functions are unique up to monotonic transformations, we set $v(0, 0) = 0$ and $v(1, 1) = 1$.

Applying the additive value function from equation (1):

$$v(x_N, x_Q) = \lambda v_N(x_N) + (1 - \lambda)\, v_Q(x_Q), \qquad (6)$$

where $\lambda \in [0, 1]$ is the weight for novelty and $v_N(\cdot)$ and $v_Q(\cdot)$ are marginal value functions bounded between 0 and 1. If more of an attribute is preferred to less, its marginal value function must be increasing. An example is the power function, where $v_j(x_j) = x_j^{\beta_j}$ for attribute j. $\beta_j > 1$ implies marginally increasing value – this seems like a reasonable form for creative artifacts, for instance, the user may deem that increasing novelty from say 0.8 to 0.9 is more valuable than from 0.2 to 0.3. $\beta_j = 1$ represents a linear marginal value function, i.e. $v_j(x_j) = x_j$ for attribute j.

Figure 2 (left) plots an additive value function over the entire domain, with weight on novelty $\lambda = 0.7$ and power marginal value functions with $\beta_N = \beta_Q = 2$. Four indifference curves are highlighted, equally spaced between the worst (0) and best (1) value. Note that when the quality score is 0 and novelty score is 1, the value is as high as 0.7, because the weight on novelty is 0.7.

The evaluator may however deem that there is no creative value (i.e. it equals 0) if either novelty or quality are at their lowest scores. Additive value functions do not support such a condition. Another aspect that additive functions fail to capture sufficiently is that of the **confluence effect**: much like how creativity in people involves more than a simple sum of their level on separate skills/abilities (Sternberg and Lubart 1991), we hypothesize that the value in a creative artifact, as deemed by an evaluator, often arises from the confluence of scores along attributes. Extended Archimedean copulas from equation (4) are examples of copulas that could potentially be used to model both these effects.

Figure 2 (middle) depicts the multiplicative form from equation (5) with linear conditional value functions at maximum reference points. The function is grounded, i.e. equals 0 when either novelty or quality is 0, and increases only when both attributes score high together. The confluence effect is heightened even further in Figure 2 (right), depicting a value copula with an exponential generating function:

$$\psi(z_i) = \frac{1 - e^{-\delta z_i}}{1 - e^{-\delta}}. \qquad (7)$$

The parameter δ models value dependence among attributes; $\delta = -5$ was chosen here. The figure indicates a low value for a significant region of the domain, and the value increases only for significantly high scores on both novelty and quality. A value function model should of course reflect the evaluator's preferences as much as possible.

A CC recommender scenario

A CC system should ideally cater to the user's preferences for artifacts/items – but what is the impact of a potential mischaracterization of the user's value function? Let us study this question using an illustrative scenario where the CC system either recommends one artifact or a list of artifacts.

Suppose there are N items produced with novelty x_N^i and quality x_Q^i of the i^{th} item. The standard approach in many creativity studies is to average out the scores to rate artifacts; the implicit value function in such a situation is additive with $\lambda = 0.5$ and marginal value functions that are linear.

If the system provides the user with the top candidate from the N items as determined by the mean rating, and if the user has value function $v(\cdot)$, then the loss in value is:

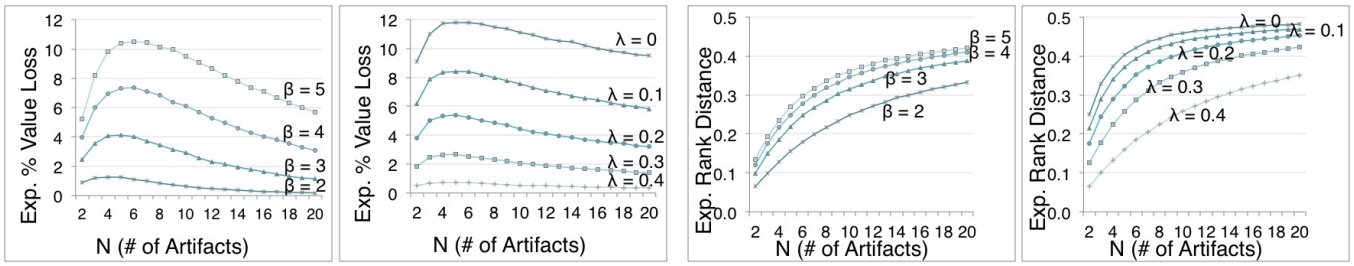
$$\text{Loss} = \max_i \left[v(x_N^i, x_Q^i)\right] - v(x_N^{i*}, x_Q^{i*}), \qquad (8)$$

where $i*$ is the item index with the highest mean rating, or formally, $i* = \text{argmax}_i \frac{x_N^i + x_Q^i}{2}$.

If the user is instead presented with an ordered ranking of items, then the discrepancy between the optimal rank ordering and the one suggested by the system is:

$$\text{Rank dist.} = D\left[r\left(i : v\left(x_N^i, x_Q^i\right)\right), r\left(i : \tfrac{x_N^i + x_Q^i}{2}\right)\right], \qquad (9)$$

where $r\left(i : C^i\right)$ denotes the rank ordering over items i based on condition C^i, and D is some distance metric. The



a) Metric: Exp. % value loss

b) Metric: Exp. rank distance

Figure 3: The two metrics in Example 2 for $N = \{2, \cdots, 20\}$ as a function of parameters β and λ. Left (both a and b): Sensitivity to β for $\lambda = 0.5$; Right (both a and b): Sensitivity to λ for $\beta = 1$.

following numerical example provides some insights into the implications of a potential discrepancy.

Example 2. [Sensitivity to additive function parameters]

Suppose a CC system draws items independently and uniformly from the unit cube: X_N^i and $X_Q^i \sim U(0,1)\ \forall i$. Furthermore, suppose that the user's value function is additive with weight on novelty λ and where marginal value functions are power functions. For simplicity, consider the case where parameters for the marginal functions are identical, i.e. $\beta_N = \beta_Q = \beta$. Simple probabilistic analysis reveals that the mean ratings of items follow a triangular distribution from 0 to 1 with mode at 0.5. One can compute the expected loss in value and expected rank distance using Monte Carlo simulations over the metrics in equations (8) and (9).

Figure 3(a) plots the % expected loss against the number of artifacts N for various parameter values of β and λ. This metric first increases with N but then decreases, after the mean rating approach has more items from which to select one that is closer in value to the optimal as per $v(\cdot)$. The figure indicates that when a parameter is significantly mischaracterized by the mean rating approach (after fixing the other parameter at its reference value), the potential % loss in value could be around $10 - 12\%$. The loss in value could be even higher if β and λ are jointly mis-specified. Due to the model assumptions, the results are symmetric around $\lambda = 0.5$ but not around $\beta = 1$.

Figure 3(b) repeats the exercise using a discrepancy in rankings where D is the normalized Kendall tau distance. This distance metric normalizes the number of swaps required to convert one rank order to another, such that identical orders result in 0 distance whereas an order and its reverse have distance 1. The rank distance increases with N and can reach distances of around $0.4 - 0.5$, implying that the system could potentially provide a user with a rank order of artifacts significantly different from the optimal order.

In this example, the user's value function was assumed to be additive. Even in this case, where the functional form is the same as the mean rating approach, not fully appreciating the user's strength of preferences over attributes could result in CC systems providing lower value artifacts to users. □

A Three-Attribute Model for Sets of Artifacts

Ritchie (2001; 2007) introduced an evaluation framework for a CC system that assesses the set(s) of artifacts it produces, where each artifact is associated with two measures: typicality $T \in [0,1]$ and quality $Q \in [0,1]$. The distinction between these two measures, for example, is how typical a joke is vs. how funny it is. Here we consider a preference model with three attributes based on this framework.

Consider a CC system that produces a large set of artifacts where the T and Q measures of each artifact are generated from a joint probability density function (pdf) $f_{T,Q}(t,q)$. Ritchie proposed several criteria based on these measures. We formulate a model with three attributes of a set of artifacts: novelty X_N and conformance X_C, both properties of the typicalities of the artifacts in the set – the idea is that some artifacts should conform to item type whereas others should be 'atypical' and therefore deemed novel – along with the quality X_Q of the set. We define these attributes based on the fraction of artifacts that are less or greater than specified thresholds. For a large enough set, these can be approximated as: $X_N \approx \int_0^{\alpha_N} f_T(t)dt$ (fraction such that $T \leq \alpha_N$), $X_C \approx \int_{\alpha_C}^1 f_T(t)dt$ (fraction such that $T > \alpha_C$), $X_Q \approx \int_{\alpha_Q}^1 f_Q(q)dq$ (fraction such that $Q > \alpha_Q$), where $f_T(t)$ and $f_Q(q)$ are marginal pdfs for each artifact's T and Q measures. Clearly, only the marginal pdfs of the generating distribution of typicality and quality matter here.

If the system builder can adjust the parameters $\boldsymbol{\theta}$ of the generating distribution $f(\cdot)$, and if the user's value function over the three attributes of the set of artifacts is $v(\cdot)$, then the optimal parameters are:

$$\boldsymbol{\theta}* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[v\left(x_N(\boldsymbol{\theta}), x_C(\boldsymbol{\theta}), x_Q(\boldsymbol{\theta}) \right) \right]. \qquad (10)$$

The reader should note that according to this three-attribute formulation, a large fraction of highly typical artifacts results in low novelty in the set – but other interpretations of Ritchie's model are possible; for instance, an artifact may be considered typical in its form but novel in its content. Consider the following illustrative numerical example.

Example 3. [System design: Typicality vs. quality]

Suppose the typicality and quality of each artifact of a CC system are generated from independent truncated Gaus-

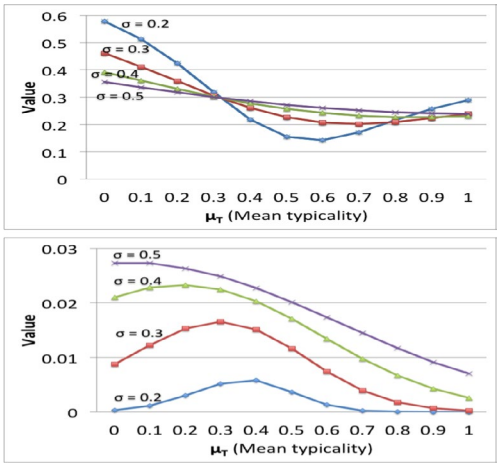


Figure 4: Value of a set of artifacts from Example 3 as a function of parameters μ_T and σ. Top: Mean rating value function; Bottom: Multiplicative value copula.

sian distributions with parameters μ_T, σ_T and μ_Q, σ_Q respectively. To examine a specific simple case, suppose $\sigma_T = \sigma_Q = \sigma$, and thresholds $\alpha_N = 1 - \alpha$, $\alpha_C = \alpha_Q = \alpha$ for $\alpha = 0.7$. Furthermore, suppose the system builder(s) can choose the mean typicality but they would need to sacrifice it for mean quality – formally, they can determine μ_T under the constraint $\mu_T + \mu_Q = 1$. How should they choose μ_T?

Figure 4 (top) helps analyze this problem for a user with a *mean rating value function*, which is the additive function from equation (1) with equal weights and linear marginal value functions: $v(x_N, x_C, x_Q) = \frac{x_N + x_C + x_Q}{3}$. Clearly, the optimal $\mu_T* = 0$ for all the displayed σ curves. Although a low mean typicality results in poor conformance, it yields both high novelty and quality, and the additive function willingly sacrifices conformance for the other two attributes.

Figure 4 (bottom) repeats the analysis for a user with a *multiplicative value copula* from equation (5) with linear conditional value functions: $v(x_N, x_C, x_Q) = x_N * x_C * x_Q$. The solution is no longer straightforward because poor performance on any individual attribute needs to be avoided. As σ decreases, a more intermediate μ_T that effectively balances the three attributes should be chosen. This example highlights how the user's value function can (and should) impact a CC system builders' decisions.

The bottom figure also reveals that a higher standard deviation improves value to the user, because it increases the fraction of the set of artifacts above or below specified thresholds. In other words, more randomness in the system is preferable; some may view this result as antithetic to the notion of CC. Several cases have been made against focusing solely on the properties of products generated by a CC system, without making other considerations – for instance, there is an argument that creative artifacts will eventually be produced by a random generation system that is "nearly equivalent to the proverbial room full of monkeys pounding on typewriters" (Ventura 2008). □

Multiple Attributes for CC Systems

The preference modeling techniques discussed for artifacts and sets of artifacts also apply more generally to CC systems. Naturally, the context in which a CC system operates should have an impact on the system builders' objectives and therefore decisions. For instance, the objectives of a CC system that provides a user with a joke every morning would be substantially different from a culinary CC system attempting a 'moonshot' recipe like the next Oreo cookie. In the previous section, we formulated a model that evaluated CC systems based on the set(s) of artifacts produced, disregarding the process by which they were created. Significant research has been pursued on frameworks that also incorporate the processes involved (Pease, Winterstein, and Colton 2001; Colton 2008; Colton, Charnly, and Pease 2011).

It is not our intention here to identify the appropriate attributes for CC systems; there is vibrant discussion in the community on such matters. Instead, we merely highlight that the preference modeling view is entirely consistent with calls for making the criteria of judging CC systems explicit (Jordanous 2012). According to this view, the system builder(s) should first deliberate over their objectives, building a hierarchy as necessary, and then identify a desirable set of attributes that ideally satisfy the properties mentioned earlier. The system builders' preferences can be represented by a value function over the attributes – this is where the proposed approach goes above and beyond current guidelines for evaluating CC systems. The following numerical example explores the use of value functions to evaluate CC systems. It is intended mainly for illustrative purposes.

Example 4. [Comparing jazz improvization systems] Jordanous (2012; 2013) compared three CC systems for jazz improvization using scores from three judges on a scale of $0 - 10$ across fourteen attributes. The attributes are organized by four high-level objectives, inspired by the four Ps (product, process, person, press) model (Rhodes 1961), as shown in Figure 1. A weighted average method was used to compare the systems, but the attribute weights that were determined for the analysis did not really reflect parameters of a user's additive value function. We will take the liberty of making additional assumptions to craft the data further into a hypothetical example, so as to illustrate some implications of taking a preference modeling approach.

First, we assume that the three CC systems' scores for each attribute are the mean values of the three judges' scores, normalized to between 0 and 1. Next, we assume that each high-level objective can be measured by a proxy attribute that aggregates the corresponding attribute scores from the lower level. Specifically, we assume that the value functions for each of the four high-level objectives are additive and equally weighted over the corresponding low-level attributes. The underlying assumption is that there is mutual preferential independence for each of the four high-level objectives. This effectively transforms the original fourteen attribute problem into a four attribute problem. From a modeling perspective, it is often helpful to construct preferences in a hierarchical fashion, but such a dimensional reduction is also useful for simplifying preference assessments.

Attribute	GAmprovising	GenJam	Voyager
'Product'	0.41	0.73	0.48
'Process'	0.34	0.70	0.38
'Person'	0.36	0.72	0.45
'Press'	0.40	0.55	0.57

Table 1: Scores for three jazz improvization systems.

The data that is generated through this transformation is displayed in Table 1; for the original data, see Jordanous (2013), Ch. 6, Table 6.3. If the goal of the exercise is to determine the best CC system then there is no need to go further – system GenJam dominates system GAmprovising by scoring higher on all four attributes, and almost dominates system Voyager. GenJam is almost surely the most preferred system, but it may be of interest to gauge how valuable it is when compared with others, in which case one needs to assess a value function over the four attributes.

We model a hypothetical value function over the four attributes using a value copula (equation (2)) so as to capture the confluence effect described earlier, where a CC system exhibits higher value only when multiple effects 'kick in' together. Specifically, we consider an exponential generating function (equation (7)) for an Archimedean copula (equation (4)) with $l_i = 0\ \forall i$. The limiting case of $\delta = 0$ makes the generating function linear, resulting in the multiplicative form (equation (5)). Conditional value functions $v(x_i|\bar{x}_i^*)$ are assumed to be power functions, and for simplicity we assume they are identical, i.e. with the same parameter β.

Figure 5 compares the values of the three systems when parameters β and δ of the value function are varied. The reference case is where conditional value functions are linear ($\beta = 1$) and where the copula is multiplicative ($\delta = 0$). Here, GenJam has value 0.2 but still beats the other systems by a significant margin. Increasing β makes the conditional value functions at the margins more convex and therefore decreases the value; there is not much difference between the three systems for $\beta > 2$. Making the dependence parameter more negative strengthens the confluence effect and decreases the value (compare the middle and right panels of Figure 2 to observe this effect). A positive parameter makes the function concave and increases value. In all cases, GenJam dominates the other systems, and even though this was evident without the need for formulating a value function, the function (and the chosen model) clearly has an impact on the value of the systems. □

Conclusions

It can be challenging to deliberate over the most pertinent objectives and attributes in many real-world decision situations, and perhaps it is even harder to identify attributes that sufficiently characterize ones preferences for creative artifacts. As Boden (1998) muses: "It is ... difficult to express (verbally or computationally) just what it is that we like about a Bach fugue, or an impressionist painting, ... And to say what it is that we like (or even dislike) about a new, or previously unfamiliar, form of music or painting

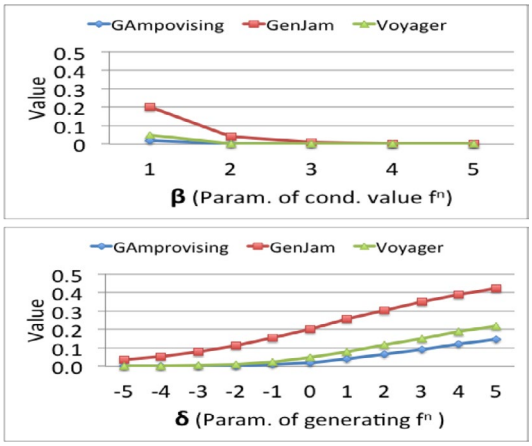


Figure 5: Value of three jazz improvization systems from Example 4 as a function of parameters β and δ. Top: Sensitivity to β for $\delta = 0$; Bottom: Sensitivity to δ for $\beta = 1$.

is even more challenging." However, if creativity is inherently subjective and involves a user's preference judgment, then understanding these preferences is a crucial aspect of CC, regardless of how challenging the task may be and how it is conducted, i.e. whether they are assessed through surveys or estimated through machine learning and related techniques (Fürnkranz and Hüllermeier 2010).

We introduced a preference modeling perspective for evaluating creative artifacts as well as systems – specifically, we formulated various multiattribute value function models. We focused primarily on additive and copula functions, stressing on the importance of the latter family of models and highlighting their potential advantages for creativity evaluation with the help of several illustrative examples. We argue for the explicit study of attributes, including the modeling of preference functions, over ad-hoc analyses that neglects to consider the implications of various assumptions.

There are various benefits to formulating preference models for CC. At an operational level, models that accurately reflect users' preferences can help in the generation of ideas and artifacts, for instance, they could improve search techniques in the conceptual space. A better understanding of preferences would also result in more effective optimization methods for CC. Furthermore, we have demonstrated that a careful consideration of the objectives of a CC system could help system builders make better strategic decisions. A CC system that could generate new attributes for meeting higher level objectives would be particularly powerful.

There are also potential limitations to using preference models in CC. Although they allow flexibility, more complex models require more parameters, and it can be far from trivial to accurately assess a complicated value function. It remains to be seen how easy or difficult it is for people to respond to preference elicitation schemes that assess multiattribute preference functions for creative artifacts. However, there is little doubt that it is essential for the best empirical research methods to effectively understand how creativity is evaluated in products, processes and ideas.

Acknowledgments

I am grateful to Lav Varshney, Anna Jordanous and four anonymous reviewers for their helpful suggestions.

References

Abbas, A. E. 2009. Multiattribute utility copulas. *Operations Research* 57(6):1367–1383.

Amabile, T. M. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology* 43:997–1013.

Belton, V., and Stewart, T. 2002. *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer.

Boden, M. 1990. *The Creative Mind: Myths and Mechanisms*. London, UK: Weidenfield and Nicolson Ltd.

Boden, M. 1998. Creativity and artificial intelligence. *Artificial Intelligence* 103:347–356.

Buchanan, B. 2001. Creativity at the metalevel. *AI Magazine* 22(3):13–28.

Cardoso, A.; Veale, T.; and Wiggins, G. A. 2009. Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI Magazine* 30(3):15–22.

Cattell, J.; Glascock, J.; and Washburn, M. F. 1918. Experiments on a possible test of aesthetic judgment of pictures. *American Journal of Psychology* 29:333–336.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, 21–26.

Colton, S.; Charnly, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity (ICCC)*, 90–95.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of AAAI Symposium on Creative Systems*, 14–20.

Dyer, J. S., and Sarin, R. K. 1979. Measurable multiattribute value functions. *Operations Research* 27(4):810–822.

Fürnkranz, J., and Hüllermeier, E. 2010. *Preference Learning*. Berlin Heidelberg: Springer-Verlag.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computing* 4:246–279.

Jordanous, A. 2013. *Evaluating computational creativity: A standardised procedure for evaluating creative systems and its application*. Ph.D. Dissertation, University of Sussex.

Keeney, R. L., and Raiffa, H. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. J. Wiley, New York.

Mayer, R. E. 1999. Fifty years of creativity research. In Sternberg, R. J., ed., *Handbook of Creativity*. Cambridge, UK: Cambridge University Press. 449–460.

Newell, A.; Shaw, J. C.; and Simon, H. 1958. The processes of creative thinking. Technical Report Report P-1320, The RAND Corp., Santa Monica, California.

Pease, A., and Colton, S. 2011. On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. In *Proceedings of the AISB Convention*, 1–8.

Pease, A.; Winterstein, D.; and Colton, S. 2001. Evaluating machine creativity. In *Proceedings of Workshop Program of ICCBR-Creative Systems: Approaches to Creativity in AI and Cognitive Science*, 129–137.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Ritchie, G. 2001. Assessing creativity. In *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in Arts and Science*, 3–11.

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

Sawyer, R. K. 2012. *Explaining Creativity: The Science of Human Innovation*. USA: Oxford University Press.

Sklar, A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* 8:229–231.

Sternberg, R. J., and Lubart, T. I. 1991. An investment theory of creativity and its development. *Human Development* 34(1):1–31.

Sternberg, R. J., and The Rainbow Project Collaborators. 2006. The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence* 34:321–350.

Stewart, T. J. 1996. Robustness of additive value function methods in MCDM. *Journal of Multi-Criteria Decision Analysis* 5(4):301–309.

Taylor, C. W. 1988. Approaches to and definitions of creativity. In Sternberg, R. J., ed., *The Nature of Creativity: Contemporary Psychological Perspectives*. Cambridge, UK: Cambridge University Press. 99–121.

Varshney, L. R.; Pinel, F.; Varshney, K. R.; Schorgendorfer, A.; and Chee, Y.-M. 2013. Cognition as a part of computational creativity. In *Proceedings of the 12th IEEE International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC)*, 36–43.

Ventura, D. 2008. A reductio ad absurdum experiment in sufficiency for evaluating (computational) creative systems. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*, 11–19.

von Winterfeldt, D., and Edwards, W. 1986. *Decision Analysis and Behavorial Research*. Cambridge, U.K.: Cambridge University Press.

Wallenius, J.; Dyer, J. S.; Fishburn, P. C.; Steuer, R. E.; Zionts, S.; and Deb, K. 2008. Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Management Science* 54(7):1336—1349.

Wiggins, G. A. 2006. Searching for computational creativity. *New Generation Computing* 24(3):209—222.