Artificial Intelligence for Engineering Design, Analysis and Manufacturing (2015), *29*, 109–128.
© Cambridge University Press 2014 0890-0604/14 $25.00
doi:10.1017/S0890060414000699

Computing the creativeness of amusing advertisements: A Bayesian model of Burma-Shave's muse

KEVIN BURNS
MITRE Corporation, Bedford, Massachusetts, USA

(Received October 23, 2012; Accepted September 4, 2014)

Abstract

How do humans judge the creativeness of an artwork or other artifact? This article suggests that such judgments are based on the pleasures of an aesthetic experience, which can be modeled as a mathematical product of psychological arousal and appraisal. The arousal stems from surprise, and is computed as a marginal entropy using information theory. The appraisal assigns meaning, by which the surprise is resolved, and is computed as a posterior probability using Bayesian theory. This model is tested by obtaining human ratings of surprise, meaning, and creativeness for artifacts in a domain of advertising design. The empirical results show that humans do judge creativeness as a product of surprise and meaning, consistent with the computational model of arousal and appraisal. Implications of the model are discussed with respect to advancing artificial intelligence in the arts as well as improving the computational evaluation of creativity in engineering and design.

Keywords: Aesthetics; Bayesian; Creativity; Information; Semantics

1. INTRODUCTION

1.1. Computational creativity

Computational creativity is an area of artificial intelligence focused on two functions. One function is generation, in which creative artifacts are synthesized using combination, transformation, and exploration algorithms (Boden, 1991). The other function is evaluation, in which creative artifacts are analyzed using mathematical, psychological, and biological approaches (Galanter, 2012).

Many computer systems have been developed to generate creativity, with their developers serving as evaluators of the resulting artifacts. In that case, the feedback loop from evaluation to generation lies in the head and hand of the developer, so it is not clear how much of an artifact's creativeness can be attributed to the computer (Boden, 2009; Burns & Maybury, 2010; Jennings, 2010). It is also not clear how creativity is evaluated in the minds of developers or other critics (Gero, 2010; Brown, 2013). What is clear is that artificial intelligence will require a capability to evaluate creativity, such as that embodied in natural intelligence, if systems are to credibly generate artifacts that humans will agree are creative.

The development of evaluative methods can be informed by disciplines outside engineering, especially the fields of fine and applied art, where creativity is highly valued. In the arts as well as other areas of creative endeavor, humans do not usually consider their cogitations to be computations. However, computational models and measures of psychological processes are needed for artificial intelligence to create and critique its own artifacts along with those of natural intelligence. Thus, the present study is aimed at a computational understanding of how humans evaluate the creativeness of applied artworks, in a domain of advertising design.

Although advertising is not engineering, it involves creative design and its artifacts can be analyzed using computational methods of the kind that engineering employs. Moreover, advertising applies to all engineered artifacts that are promoted in communications aimed at selling, trading, or giving products to users. When creativity is important to consumers, then it is also important in advertising so that consumers can experience some of the creative spirit (before getting the product) as a basis for wanting the product (Walvis, 2008).

1.2. Engineering evaluations

In the realm of engineering design and analysis, researchers have recognized that evaluations of creativity are *judgments* relative to personal and cultural norms, and have explored the factors that underlie such judgments (for a recent review, see Brown, 2013). Although a number of different models

Reprint requests to: Kevin Burns, MITRE Corporation, 202 Burlington Road, Bedford, MA 01730-1420, USA. E-mail: kburns@mitre.org

have been proposed for evaluating creativity, most involve the same or similar factors as outlined below.

An early and influential example is the creative product analysis model (CPAM; Besemer & Treffinger, 1981), which assesses creativity using human judgments of three factors (each involving several subfactors). The first factor is *novelty*: the extent of newness in a product compared to existing products. The second factor is *resolution*: the degree to which a product meets the needs of a problematic situation, which other models more commonly call *value* or *utility*. The third factor is *style*: the degree to which a product combines unlike elements into a refined, developed, coherent whole. This notion of unity (a coherent whole) amid variety (unlike elements) is akin to what other authors refer to as *aesthetics* or *beauty* (Blinderman, 1962; Hofstadter & Kuhns, 1964; Berlyne, 1971).

CPAM relies on human judgments to quantify each factor, so it does not specify how any factor can be computed from features of an artifact itself. Similarly, the SAPPhIRE model (Srinivasan & Chakrabarti, 2010; Sarkar & Chakrabarti, 2011) solicits human judgments about functional, behavioral, and structural aspects of a product. These judgments are combined to obtain measures of *novelty* and *utility*, which are then combined to obtain a measure of creativity. However, a key difference compared to CPAM is that SAPPhIRE does not address the aesthetic style of an artifact.

Other authors (Ritchie, 2001, 2007; Maher, 2010; Maher & Fisher, 2012; Grace et al., 2014, in press) have proposed computational models in which factors that contribute to creativity can be quantified directly from features of artifacts, without human judgments. However, unlike the conceptual models cited above, these computational models have not been tested against human judgments of the individual factors or overall creativity itself. The computational models have also not addressed aesthetic style, although some of these models include *surprise* (Maher, 2010; Maher & Fisher, 2012; Grace et al., 2014, in press) as a third factor along with novelty and value (utility).

1.3. Addressing aesthetics

A computational model for evaluating creativity would ideally specify how contributing factors are combined in an overall measure of creativeness, including interdependencies. For example, novelty and surprise are not independent factors, because novel artifacts are usually surprising when one encounters them. Similarly, value or utility may depend on surprise and novelty, especially when the underlying value judgments include aesthetics.

Psychological theories and empirical testing suggest that aesthetic pleasure is a hedonic value (Berlyne, 1970) produced from arousal and appraisal (Berlyne, 1971; Parsons, 1987; Scherer, 1999; Oatley, 2003; Silvia, 2005), where arousal stems from surprise in sensing stimuli that exhibit novelty, complexity, incongruity, uncertainty, and so on (Berlyne, 1957, 1963, 1970; Silvia, 2006). Thus aesthetic *value* depends on *surprise*, which in turn depends on *novelty*. This aesthetic value, produced by a pleasing form (Postrel, 2003; Norman, 2004), is in addition to any pragmatic value or utility associated with an artifact's function.

In sum, there seems to be some consensus that novelty and utility are two factors that contribute to judgments of creativity, but other factors like aesthetic style and surprise are also important. Computational models have been developed to quantify some of these factors, but they have not addressed aesthetics, nor accounted for how individual factors may interact, nor been benchmarked against human judgments of creativity or contributing factors. Artificial intelligence is in need of computational models that can overcome these limitations, in engineering as well as the arts, because empirical studies have demonstrated the importance of an affective (aesthetic) factor in human judgments of creative product designs (Besemer & Treffinger, 1981; Besemer & O'Quin, 1986; O'Quin & Besemer, 1989; Besemer, 1998; Besemer & O'Quin, 1999; Horn & Salvendy, 2006).

Aesthetics apply to any artifact, but are especially important in fine art and applied arts such as architecture and advertising. Thus, the present paper develops and validates a computational model of aesthetics for evaluating creativity in a domain of advertising design. In addition to aesthetic style, the computational model also addresses novelty, utility, surprise, value, and other factors that can affect assessments of creativity, including personal and cultural differences between individuals.

1.4. Amusing advertisements

Within the realm of advertising design, the present study deals with amusing advertisements known as *Burma-Shave* jingles (Rowsome, 1965). This genre was chosen because it employs humor in compact verses that are conducive to analyses and experiments, which are required for rigor in modeling and measuring aesthetic experiences. As a practical matter, humor is often employed by natural intelligence for the purpose of persuasion in advertisements (Weinberger & Gulas, 1992; Sternthal & Craig, 1973), and computational humor remains a difficult challenge for artificial intelligence (Binstead, 2006).

Burma-Shave was a shaving cream sold in jars and tubes, containing essential oils from the Malay peninsula and Burma. When it was sold during the mid-1900s, the product was an innovative alternative to other shaving lathers that had to be worked up and applied with a wet shaving brush. Equally innovative was Burma-Shave's unique approach to advertising, known as *The Verse By the Side of the Road* (Rowsome, 1965).

Each verse was a short jingle written in five lines followed by a sixth line that said simply "Burma-Shave." Each line was painted on its own sign in plain text, and signs were spaced at uniform intervals along roadways such that it took about 3 s to proceed from one sign to the next. The short pauses built suspense as readers anticipated each next line, and the effort was usually rewarded with a surprising twist in the fifth line, which served as the punch line of the jingle. This advertising approach actively engaged a reader for about 15 s, which was far longer than most newspaper or magazine ads could accomplish. An example is the following jingle from 1939:

past
schoolhouses
take it slow
let the little
shavers grow

This verse is prototypical in eliciting two feelings that often arise in humor (Roeckelein, 2002). First, there is a feeling of surprise from incongruity, when the reader reaches the punch line. Second, there is a feeling of success and superiority, when the reader resolves the surprise. Together the arousal from surprise and appraisal of success produce an experience that people find *pleasurable*, and this aesthetic is pragmatic in making advertisements more *persuasive* (Meyers-Levy & Malaviya, 1999).

The remainder of this article presents analyses and experiments with a focus on such advertisements, as follows: Section 2 derives a computational model of aesthetic creativity, informed by psychological science and employing mathematical theories of Bayesian inference and Shannon information; Section 3 dissects a prototypical Burma-Shave jingle and quantifies an audience's aesthetic experience using the computational model; Section 4 details the results of an experiment to validate the model's formulation of aesthetic creativity, by comparing the model to human judgments (as well as to competing models); and Section 5 explains how the model can advance artificial intelligence in the arts and improve the evaluation of engineering artifacts.

2. FORMULATION

Aesthetic experiences involve *feelings* that arise as an audience assigns *meanings* to *signals*. In particular, it is important to distinguish signals, such as the lines of a Burma-Shave jingle, from the meaning that may be assigned to those signals. The signals are evidence, observed with certainty, whereas meaning is a hypothesis, subject to uncertainty. It is also important to distinguish a feeling from the meaning of a signal; the feeling is a *consequence* of the media experience, rather than a *hypothesis* (meaning) or *evidence* (signal). These distinctions are highlighted here because they are important in the derivation that follows.

2.1. Bayesian inference

Bayesian inference (Bayes, 1763; McGrayne, 2011) is the rational method for updating confidence in competing hypotheses as evidence is obtained. For example, assume that a reader initially has two hypotheses about the possible meanings, A and B, of a signal. Here A and B are *hypothesized meanings*, so the terms hypothesis and meaning will be used interchangeably below. Similarly, the terms *evidence* and *signal* are synonymous and will be used interchangeably below.

Before receiving any evidence a reader has prior confidence in each hypothesis, which can be measured by the probability that each meaning is true. Note that A and B are competing hypotheses, which means they are exclusive (only one hypothesis, A or B, can actually be true) and exhaustive (the probabilities of A and B sum to 1). With these assumptions, Bayes' rule specifies how the probability of each hypothesis should be updated in light of evidence, based on likelihoods of the evidence assuming the truth of each hypothesis.

Mathematically, Bayes' rule is expressed as follows:

$$P(H_i|e) = P(H_i) \times P(e|H_i)/P),$$

where $P(H_i|e)$ is the *posterior* probability of a hypothesis (H_i) after receiving some evidence (e); $P(H_i)$ is the *prior* probability of the hypothesis before receiving the evidence; $P(e|H_i)$ is the conditional *likelihood* of the evidence assuming the hypothesis is true; and $P(e)$ is the *marginal* probability of the evidence over all hypotheses in the set $\{H_i\}$ of mutually exclusive and exhaustive hypotheses. This marginal probability $P(e)$ is a normalizing factor computed as the following sum of numerators in Bayes' rule:

$$P(e) = \sum_i P(H_i) \times P(e|H_i).$$

Now consider the case of a Burma-Shave advertisement, or any other media experience in which an audience is attempting to understand the *meaning* of the *signals* that are received. The meaning is not known for certain, especially early in the message when few or no signals have been received as evidence. Therefore, various meanings are the hypotheses, M_i, and various signals are the evidence, s_j. Using Bayes' rule, we can compute confidence in hypothesized meanings given evidential signals, and we can do so incrementally as signals are received sequentially. That is, the posteriors after one signal become the priors before the next signal, and so forth.

Of course this requires more than just Bayes' rule, because priors and likelihoods are needed as input in order to compute posteriors as output. For human audiences with natural intelligence, these inputs will come from past experience and associated background knowledge about possible meanings M_i with prior probabilities $P(M_i)$, as well as knowledge about the likelihoods $P(s_j|M_i)$ of various signals s_j given the truth of each possible meaning M_i. Clearly this knowledge of semantic categories M_i and probabilistic quantities $P(M_i)$ and $P(s_j|M_i)$ is critical. Armed with these inputs (from natural intelligence if not from artificial intelligence), Bayes' rule provides a principled method for modeling how meanings are assigned to signals by an audience engaged in reading an advertisement.

2.2. Beyond inference

Although Bayes' rule is useful for modeling how humans assign *meanings* to *signals*, it says nothing about how *feelings* arise in the process. As such, Bayes' rule addresses only half of the problem, from advertisements (signals) to understanding (meaning), and another approach is needed to address the remaining half of the problem, from understanding (meaning) to aesthetics (feelings).

The second half of the problem is addressed by a theory dubbed EVE′ (Burns, 2006*a*), which postulates a psychological progression of *expectation* (E), *violation* (V), and *explanation* (E') as the essence of every aesthetic experience. This theory is grounded in the ecological assumption that the brain evolved to reward itself with pleasure for successes in expectations and explanations, because these successes would help humans survive and thrive in the natural world. The pleasure of explanation is especially important because it serves as a biological reward for experiential learning. That is, if an organism only experienced pleasure when expectations were satisfied, then the organism would have no affective motivation for seeking out and learning from unexpected stimuli.

EVE′ holds that humans are constantly forming expectations of signals that might possibly be received, as well as forming explanations of meaning for the signals that are actually received. The aesthetic feeling of pleasure by which an audience gauges beauty or creativity is thus computed as the sum of two terms: pleasure that arises when an expectation is satisfied, that is, when one receives a signal that is consistent with one's expectation (E); and pleasure-prime that arises after a violation of expectation, if and when the violation is resolved by an explanation (E').

Per EVE′, the psychological arousal produced by a violation of expectation is modeled mathematically by Shannon's measure of *surprisal* (Shannon & Weaver, 1949) and computed as a marginal entropy equal to $-\log P(s_j)$. This is a positive quantity that increases as $P(s_j)$ decreases. Here and hereafter, the log is assumed to be $\log_2$ such that entropy is measured in *bits* of information. Note that this Shannon entropy is a function of the signal probability $P(s_j)$, which is not the same as the Bayesian posterior probability $P(M_i|s_j)$ or prior probability $P(M_i)$ that other authors have used to model a notion of "Bayesian surprise" (Itti & Baldi, 2009). Instead, EVE′ uses the Bayesian probability as a measure of *meaning* (discussed below), and uses the Shannon entropy of a signal as a measure of *surprise*.

By this model, if a signal s_j is expected at probability $P(s_j)$, then when s_j is received, there is a violation measured as $V = -\log P(s_j)$. Conversely, expectation E is measured as the additive inverse of violation V, such that $E = \log P(s_j)$. Finally, an explanation E' of the violation V is modeled using a Bayesian measure of *meaning*, computed as the posterior probability $P(M_i|s_j)$ for the hypothesized meaning M_i with the highest posterior probability. This probability is a number between zero and one, so EVE′'s measure of explanation E' is given by the fraction of the violation V that is resolved,

$$E' = V \times P(M_i|s_j) = -\log P(s_j) \times P(M_i|s_j).$$

Using the above expressions, aesthetic pleasure (X) is computed as a weighted combination of E and E' as follows:

$$\begin{aligned} X &= G \times E + G' \times E' \\ &= G \times \log\ P(s_j) - G' \times \log\ P(s_j) \times P(M_i|s_j), \end{aligned}$$

where the first term in the sum models pleasure from successful expectation (E) scaled by a factor G, and the second term in the sum models pleasure-prime from successful explanation (E') scaled by a factor G'. These scaling factors, G and G', are used to model personal preferences for the two types of pleasure, which may vary between individuals.

The parameters G and G' are limitations of the EVE′ model, because numerical values must be provided as input in order to compute X as output. This limitation is similar to the priors and likelihoods (described above) that are also required as input to the model. However, all of these model limitations are useful in the sense that they specify what would be required as input (in the form of natural knowledge and personal factors) for artificial intelligence to feasibly and credibly compute aesthetics and judgments of creativeness.

Although the EVE′ model itself does not provide numerical values for G and G', the underlying theory of expectation–violation–explanation does suggest that values will generally satisfy the inequality $G' > G$ (assuming both are positive values). This is because a gain of pleasure-prime from explanation can only be obtained at the cost of pleasure from expectation, so the brain would need to have evolved some incentive for giving up a unit of pleasure at E (scaled by G) in order to gain a fractional unit of pleasure-prime at E' (scaled by G'). In EVE′, this incentive is captured by the ratio $G'/G > 1$, which represents a person's drive or desire for arousal, akin to the personality trait known as *sensation seeking*.

Sensation seeking, defined as "the seeking of varied, novel, complex, and intense sensations and experiences" (Zuckerman, 1994), is a personality trait that has been found to affect aesthetic judgments as well other behavioral preferences (Demaree et al., 2008). In particular, high sensation seekers tend to prefer more abstract art (Zuckerman, 1994) and nonsensical humor (Ruch, 1988), whereas low sensation seekers tend to prefer less abstract art and more conventional humor. These personal preferences are captured by EVE′'s ratio G'/G, where higher G'/G implies a preference for less predictable stimuli, and lower G'/G implies a preference for more predictable stimuli.

Besides magnitude, the valence (sign) of G'/G may also vary between individuals depending on the meaning M_i of the signal s_j. For instance, one person may be offended by the meaning of a signal that another person would find amusing, based on cultural background and/or personal values that differ between the two individuals. A specific example of this is provided later, in Section 3.4.

Now, returning to the previous equation and rearranging terms, EVE′'s equation for aesthetics (X) can be expressed as follows:

$$X = -\log\ P(s_j) \times [G' \times P(M_i|s_j) - G] = Y \times Z.$$

In this form, we can see that X is the product of two terms, $X = Y \times Z$, where $Y = -\log P(s_j)$ is a Shannon-entropy measure of *surprise*, and $Z = [G' \times P(M_i|s_j) - G]$ is a Bayesian-probability measure of *meaning*.

Figure 1 offers more insight into the equation, by plotting a family of curves for X versus $P(s_j)$, with each curve representing a different value of $P(M_i|s_j)$ in the range 0 to 1 (in 0.1 increments). On the left side of Figure 1, $P(s_j)$ is small so the surprise of receiving s_j is large. In that case, we find both the most positive and the most negative aesthetic X, depending on the meaning as measured by $P(M_i|s_j)$. When $P(M_i|s_j)$ is high, then X is very positive (pleasurable), and when $P(M_i|s_j)$ is low, then X is negative (displeasurable). Moving to the right of Figure 1, surprise Y decreases as $P(s_j)$ increases, so the product $X = Y \times Z$ is less affected by the meaning Z, although higher and lower values of $P(M_i|s_j)$ at a given $P(s_j)$ still cause higher and lower X values, respectively.

3. DEMONSTRATION

As a concrete example of how EVE′ applies to amusing advertisements, this section analyzes the earlier Burma-Shave jingle about *little shavers*. In performing the analysis, I will parse the jingle into three signals (evidence) as follows: s_1 = *past schoolhouses take it slow*, s_2 = *let the little*, and s_3 = *shavers grow*. In the analysis, I will consider four meanings (hypotheses), to be denoted A, B, C, and D.

The analysis will show how the most likely meaning progresses from A to A to C as the signals proceed from s_1 to s_2 to s_3. As such, the verse can be characterized as an AAC version of the AAB pattern that is common in humor and music (Rozin et al., 2006). A similar pattern applies to humor in haiku form, which has also been analyzed by EVE′ (Burns, 2012).

Based on the computed confidence in meanings at each stage of signal processing, the three signals can be characterized as: a setup (s_1), after which A is the most probable meaning; a buildup (s_2), after which A is even more probable; and a punch line (s_3), after which C is the most probable meaning. Although the punch line is rather obvious as the last line, the preceding four lines of the jingle might be grouped differently into a setup and buildup (or perhaps more or less stages). However, it is reasonable to assume the setup ends at the word *slow*, because this word rhymes with the last word of the punch line (*grow*); hence, the intervening line acts as a bridge (buildup). Moreover, a different grouping of lines into setup and buildup *signals* would not change the main point of the Bayesian analysis below, namely, that the aesthetic experience is driven by a reversal in *meaning* after the punch line is read.

Similarly, the hypothesized meanings A, B, C, and D are certainly debatable and probably impoverished compared to the set of possible meanings being considered by human readers. However, again those details would not change the main thrust of the analysis, which illustrates a reversal of the most likely meaning based on the perceived *intent* of the message.

3.1. The setup: s_1 = *past schoolhouses take it slow*

To begin, assume a reader has read the first signal, s_1 = *past schoolhouses take it slow*. The reader will wonder why he should drive slowly past schoolhouses, and an obvious reason is that children who attend school may be coming or going or playing nearby. Thus, the reader forms the hypothesis A = drive slowly past schoolhouses to avoid harming schoolchildren. At the same time, however, the reader realizes there are other possible meanings, captured in a hypothesis B = not A, which includes all other reasons why one should take it slow past schoolhouses.

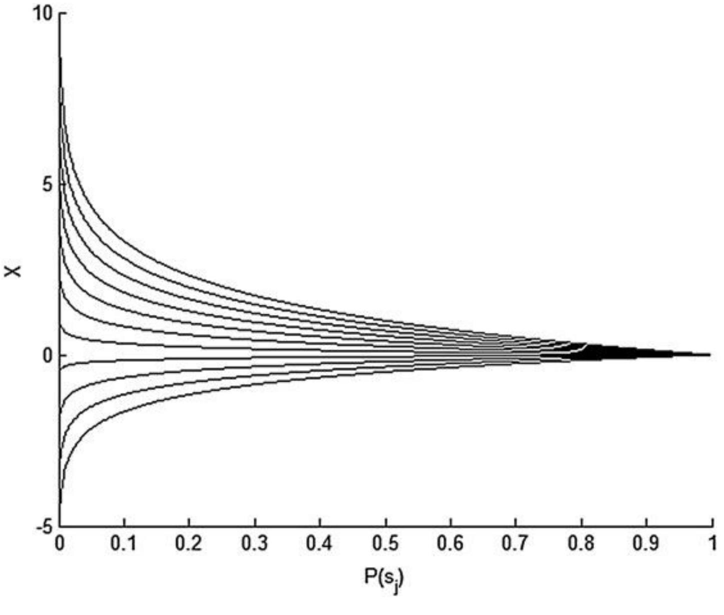

Fig. 1. Plot of EVE′'s equation $X = Y \times Z = -\log P(s_j) \times [G' \times P(M_i|s_j) - G]$, assuming $G' = 1.5$ and $G = 0.5$. The family of curves varies $P(M_i|s_j)$ from 0.0 (lowest curve) to 1.0 (highest curve) in increments of 0.1.

The signal s_1 gives no hint of reasons other than to avoid harming schoolchildren, so a reader will infer that A is relatively probable. In contrast, there are many other possible meanings included in B, and the reader knows that further signals (which may support different meanings) are yet to come. Thus, the reader may think $P(\mathrm{A}|s_1) > P(\mathrm{B}|s_1)$, but these probabilities would be moderate (i.e., close to 0.5) rather than extreme (i.e., close to 0.0 or 1.0). Here and hereafter, I will assume numbers for probabilities in order to make the analysis more concrete. In doing so, I am not suggesting that readers are explicitly thinking in terms of numbers, or that the assumed numbers are exactly the ones implicitly represented in cognitive processing. Instead, I assume numbers (based on introspection) as approximate and representative values in order to illustrate the computational model more clearly and specifically.

Here I assume $P(\mathrm{A}|s_1) = 0.60$ and $P(\mathrm{B}|s_1) = 0.40$, consistent with the logic outlined above, as initial probabilities that a reader assigns to meanings A and B after receiving the first signal, s_1. These values then become prior probabilities for a Bayesian update to be performed after the next signal, s_2, is received.

In addition to the meanings A and B, which serve as explanations of the signal s_1, a reader will also be forming expectations about the next signal, s_2, to be received as evidence. Per EVE′, the reader is a bounded-Bayesian whose working memory (Baddeley, 1992; Cowan, 2001) cannot possibly represent all possible signals, just as his working memory cannot possibly represent all possible meanings that may be recalled or constructed from long-term memory. Instead, a reader's working memory is assumed to represent *classes of signals* that are *consistent* with the *currently held meanings*, A and B. This assumption is an important aspect of the EVE′ model, discussed further in Section 3.5.

In light of the two hypothesized meanings, A and B, there are two diagnostic and therefore meaningful classes to which the next signal (s_2) might belong: signals in class a (consistent with A) that are likely if A is true but unlikely if B is true; and signals in class b (consistent with B) that are likely if B is true but unlikely if A is true. Using $P = 0.90$ to represent "likely" and $P = 0.10$ to represent "unlikely," the likelihoods for these two classes of signals can be expressed as follows:

$$P(a|\mathrm{A}, s_1) = 0.90, \quad P(a|\mathrm{B}, s_1) = 0.10,$$

$$P(b|\mathrm{A}, s_1) = 0.10, \quad P(b|\mathrm{B}, s_1) = 0.90.$$

Finally, the priors and likelihoods together can be used to compute the marginal probabilities of signals in classes a and b that might be received next, as follows:

$$\begin{aligned} P(a|s_1) &= P(a|\mathrm{A}, s_1) \times P(\mathrm{A}|s_1) + P(a|\mathrm{B}, s_1) \times P(\mathrm{B}|s_1) \\ &= 0.90 \times 0.60 + 0.10 \times 0.40 = 0.58, \end{aligned}$$

$$\begin{aligned} P(b|s_1) &= P(b|\mathrm{A}, s_1) \times P(\mathrm{A}|s_1) + P(b|\mathrm{B}, s_1) \times P(\mathrm{B}|s_1) \\ &= 0.10 \times 0.60 + 0.90 \times 0.40 = 0.42. \end{aligned}$$

In words, after receiving s_1 the next signal, s_2, is expected to be in class a with probability 0.58 and is expected to be in class b with probability 0.42.

3.2. The buildup: s_2 = *let the little*

When the second signal is received, the reader will recognize that s_2 = *let the little* is consistent with not harming children per meaning A, and does not suggest any other meaning B. That is, the signal s_2 is in class a rather than class b. Based on this new evidence, the reader will update his prior confidence in each meaning across the set of hypotheses $\{\mathrm{A}, \mathrm{B}\}$, using the likelihoods of evidence $s_2 = a$ noted above. These likelihoods are $P(s_2|\mathrm{A}, s_1) = 0.90$ and $P(s_2|\mathrm{B}, s_1) = 0.10$, and the priors are $P(\mathrm{A}|s_1) = 0.60$ and $P(\mathrm{B}|s_1) = 0.40$, so the Bayesian update produces posteriors as follows:

$$P(\mathrm{A}|s_2, s_1) = P(\mathrm{A}|s_1) \times P(s_2|\mathrm{A}, s_1)/P(s_2|s_1) = 0.93,$$

$$P(\mathrm{B}|s_2, s_1) = P(\mathrm{B}|s_1) \times P(s_2|\mathrm{B}, s_1)/P(s_2|s_1) = 0.07.$$

In other words, the reader's confidence in A increases from a prior 0.60 before s_2 to a posterior 0.93 after s_2. According to EVE′, the consequence of this Bayesian update is a feeling of pleasure, computed as the product of surprise (Y_2) and meaning (Z_2). Per the model, $Y_2 = -\log P(s_2|s_1)$, because s_2 was the new signal received, and $Z_2 = [G' \times P(\mathrm{A}|s_2, s_1) - G]$, because A is the most likely posterior meaning after s_2 is received. The marginal probability of s_2 (conditional on s_1), needed to compute surprise, is given by the earlier equation for $P(a|s_1)$, i.e., $P(s_2|s_1) = P(a|s_1) = 0.58$. Thus we obtain the following equation:

$$\begin{aligned} X_2 = Y_2 \times Z_2 &= -\log P(s_2|s_1) \times [G' \times P(A|s_2, s_1) - G] \\ &= -\log 0.58 \times [G' \times 0.93 - G]. \end{aligned}$$

To compute a final number, we can assume $G' = 1.5$ and $G = 0.5$, consistent with the ratio $G'/G = 3$ used in previous studies (Burns, 2006*a*, 2012; Burns & Dubnov, 2006) and illustrated in Figure 1. With these values, we obtain $X_2 = 0.70$. Therefore, after receiving and processing s_2, the reader has experienced 0.70 bits of pleasure as a product of arousal (surprise) and appraisal (meaning).

Here as before, the posteriors become priors for a Bayesian update to be performed after the next signal, s_3, is received. At this point as well, the reader is forming expectations about possible signals that might be received next as s_3. Because there are only two hypotheses, A and B, we can assume the reader is expecting a signal in class a or a signal in class b, with the same likelihoods as his earlier expectations before s_2. However, here after s_2 the reader has different priors than before s_2, so the probabilities of potential signals s_3 (in classes a and b) are computed as follows:

$$
\begin{aligned}
P(a|s_2, s_1) &= P(a|\mathrm{A}, s_2, s_1) \times P(\mathrm{A}|s_2, s_1) \\
&\quad + P(a|\mathrm{B}, s_2, s_1) \times P(\mathrm{B}|s_2, s_1) \\
&= 0.90 \times 0.93 + 0.10 \times 0.07 = 0.84,
\end{aligned}
$$

$$
\begin{aligned}
P(b|s_2, s_1) &= P(b|\mathrm{A}, s_2, s_1) \times P(\mathrm{A}|s_2, s_1) \\
&\quad + P(b|\mathrm{B}, s_2, s_1) \times P(\mathrm{B}|s_2, s_1) \\
&= 0.10 \times 0.93 + 0.90 \times 0.07 = 0.16.
\end{aligned}
$$

In words, after receiving s_2, the next signal, s_3, is expected to be in class a with probability 0.84 and is expected to be in class b with probability 0.16.

3.3. The punch line: s_3 = *shavers grow*

When the third signal, s_3, is received, the reader recognizes it as being in class b that is consistent with B and inconsistent with A, because schoolchildren do not shave and because shaving by little adults appears to have nothing to do with child safety near schoolhouses. The signal $s_3 = b$ was quite unexpected ($P(b|s_2, s_1) = 0.16 \ll 1$), so s_3 produces a large violation of expectation. Fueled by this arousal, a reader with a good sense of humor who readily makes creative associations will think of two things based on his background knowledge. First, he will recall that children are fondly referred to as *little shavers* by adults using slang language. Second, he will realize that children who grow up and shave are potential customers of Burma-Shave. Thus, he will see that the punch line *shavers grow* implies a *compound meaning* for the jingle, which is C = drive slowly past schoolhouses to avoid harming schoolchildren *and* to avoid harming future profits of Burma-Shave.

This recognition of the compound meaning C represents a creative leap by the reader, much like the creative leap of the writer who authored the jingle in the first place. As is typical in humor (Roeckelein, 2002), understanding the meaning of the jingle and appreciating the author's intent gives the reader a feeling of dominance or superiority, which is shared by others who also understand and appreciate the jingle. This makes humor, when used appropriately, an effective tool for building a brand in the field of advertising (Walvis, 2008). In contrast, humor must be used carefully because failed attempts can do more harm than good (Weinberger & Gulas, 1992; Sternthal & Craig, 1973), as illustrated by negative values of X in Figure 1 and as exemplified in Section 3.4.

A reader who recognizes the compound meaning is effectively *abducting* (Thagard, 2007) a hypothesis C that was previously not explicit in his frame of discernment $\{\mathrm{A}, \mathrm{B}\}$. The result is a revised set of hypotheses, $\{\mathrm{A}, \mathrm{C}, \mathrm{D}\}$, where A is the same as before and C is as described above. The third hypothesis, D = not A or C, replaces the previous B = not A.

Notice that C includes two different reasons for driving slowly past schoolhouses, whereas A includes only one reason. In addition, C captures specific reasons, whereas D (= not A or C) includes only unspecified reasons that are not included in A or C. These features of C compared to A and D affect the relative likelihoods of s_3, which are estimated as follows:

$$P(s_3|\mathrm{A}, s_2, s_1) = 0.01,$$

$$P(s_3|\mathrm{C}, s_2, s_1) = 0.98,$$

$$P(s_3|\mathrm{D}, s_2, s_1) = 0.01.$$

Here the likelihoods $P(s_3|\mathrm{A}, s_2, s_1) = P(s_3|\mathrm{D}, s_2, s_1) = 0.01$ are assumed to be "very unlikely" and hence an order of magnitude lower than the previous likelihood $P(b|\mathrm{A}, s_1) = 0.10$. This reflects the specificity of s_3 = *shavers grow* and its inconsistency with the single meaning of A or the other unspecified meanings captured by D. However, the signal s_3 = *shavers grow* is completely consistent with child safety *and* Burma-Shave's future profits, so s_3 is "very likely" if the compound meaning C is true.

Besides these likelihoods, the use of Bayes' rule to update beliefs requires a prior probability for each hypothesis. Because C and D together capture a scope of meaning that was previously covered by B, we can split the magnitude of $P(\mathrm{B}|s_2, s_1) = 0.07$ between C and D to obtain $P(\mathrm{C}|s_2, s_1) = P(\mathrm{D}|s_2, s_1) = 0.035$. The priors are then updated via Bayes' rule, using likelihoods for s_3 as noted above, to obtain the following posteriors:

$$P(\mathrm{A}|s_3, s_2, s_1) = P(\mathrm{A}|s_2, \mathrm{s}_1) \times P(s_3|\mathrm{A}, s_2, s_1)/P(s_3|s_2, s_1) = 0.21,$$

$$P(\mathrm{C}|s_3, s_2, s_1) = P(\mathrm{C}|s_2, s_1) \times P(s_3|\mathrm{C}, s_2, s_1)/P(s_3|s_2, s_1) = 0.78,$$

$$P(\mathrm{D}|s_3, s_2, s_1) = P(\mathrm{D}|s_2, s_1) \times P(s_3|\mathrm{D}, s_2, s_1)/P(s_3|s_2, s_1) = 0.01.$$

In words, the reader's abduction of new meaning C causes a reversal in his beliefs: from A being the most probable hypothesis $P(\mathrm{A}|s_2, s_1) = 0.93$ before s_3; to C being the most probable hypothesis $P(\mathrm{C}|s_3, s_2, s_1) = 0.78$ after s_3. The same sort of reversal and recognition is common in comedy, as articulated by Aristotle in his *Poetics* (Butcher, 1951) and as analyzed in more modern studies of *Mathematics and Humor* (Paulos, 1980).

To complete the analysis, recall from above that the marginal probability of a signal $s_3 = b$ (conditional on s_2 and s_1), before receiving the signal s_3, was $P(s_3|s_2, s_1) = P(b|s_2, s_1) = 0.16$. Thus we obtain the following aesthetic measure:

$$X_3 = Y_3 \times Z_3 = -\log 0.16 \times [G' \times 0.78 - G].$$

Assuming the same values of $G' = 1.5$ and $G = 0.5$ used earlier, we obtain $X_3 = 1.8$. In words, after s_3 the reader experiences 1.8 bits of pleasure, on top of the 0.7 bits of pleasure already experienced after s_2. Taken together, the reader has experienced a total of $X = X_2 + X_3 = 2.5$ bits of pleasure from reading the jingle, with most of this pleasure coming after the punch line.

3.4. When humor fails

Now returning to the punch line, $s_3 = $ *shavers grow*, consider the aesthetic experience of a reader who does not know that *little shavers* is a slang term for children. Although this reader might realize that shavers are customers of Burma-Shave, he would not understand why little adults who shave would be near schoolhouses, or what that has to do with careful driving or child safety. In effect, the reader is left with confusion, rather than an explanation that resolves the violation. For example, the confused reader might be considering five hypotheses and assign a roughly equal probability to each hypothesis, that is, $P(M_i|s_3, s_2, s_1) \approx 0.2$, because none of the hypotheses stands out as being particularly probable.

In that case, the factor Z would be negative, $Z_3 = [1.5 \times 0.2 - 0.5] = -0.2$, such that any amount of surprise Y would cause the resulting $X = Y \times Z$ to be negative. In other words, the reader would feel displeasure rather than pleasure after reading the punch line, and displeasure would increase as the amount of surprise increases. This is consistent with the annoyance one feels after reading a punch line that is surprising but cannot be resolved in a meaningful manner.

Here I assume that surprise for a reader who did not understand the punch line is the same as for a reader who did understand the punch line, because surprise is a function of the signal probability $P(s_3|s_2,\ s_1)$ before the punch line, s_3, is received. Thus, we obtain $X_3 = Y_3 \times Z_3 = -\log 0.16 \times [1.5 \times 0.2 - 0.5] = -0.53$ bits of pleasure, which represents 0.53 bits of *displeasure*.

A similar but potentially much larger displeasure arises when one understands the punch line and does not appreciate the author's intent. EVE′ captures this effect via the valence of G', which scales the affective consequence of a person's explanation. For example, a reader who personally knows of a child hurt by reckless driving may find it offensive that Burma-Shave is joking about child safety in order to promote corporate profits. In that case, the reader's Bayesian inference would be the same as other readers who understand the jingle, but the G' factor would be negative. Assuming $G' = -1.5$ rather than the previous $G' = 1.5$, EVE′'s aesthetic measure would be computed as $X_3 = Y_3 \times Z_3 = -\log 0.16 \times [-1.5 \times 0.78 - 0.5] = -4.4$ bits, which is 4.4 bits of *displeasure*.

As described above, the EVE′ model predicts three modes in which humor can fail, with each mode corresponding to a specific model parameter. First, humor fails when there is little or no *surprise* (modeled by EVE′'s factor Y), regardless of how much meaning can be extracted from the signal. Section 4.2.5 provides a specific example of this failure mode. Second, humor fails when there is little or no *meaning* (modeled by EVE′'s factor Z), regardless of how surprising the signal may be. An example was given above, and Section 4.2.5 provides another example of this failure mode. Third, humor fails when the *intent* of the author, per the *meaning* inferred by the audience, is contrary to the value structure of the audience (modeled by EVE′'s parameter G'). A specific example of this failure mode was provided above.

The third failure mode, involving *intent*, appears to be especially challenging for artificial intelligence to address (Dennett, 1987). The reason is that humans have mental models of how they and others would act, and what they and others would want, in various situations, as shaped by cultural norms and personal goals. These mental models include rich representations of norms and goals, which machine systems would need to model as well. The same problem poses challenges to human artists who write advertisements (Sternthal & Craig, 1973; Weinberger & Gulas, 1992), because audience responses can be sensitive to gender, age group, cultural background, financial status, political party, religious beliefs, and so on. However, given the long-term success of Burma-Shave's advertising campaign, it is reasonable to assume that the humor in its ads was generally effective across individuals. Therefore, EVE′'s analysis in the above demonstration and in the below validation (Section 4) is focused on the other two modes in which humor can fail. That is, humor fails when there is little *surprise* Y or little *meaning* Z, because for humor to succeed there must be both high surprise and high meaning, $X = Y \times Z$.

3.5. EVE′'s advantages

The EVE′ model is not alone in assuming that *expectations* and *violations* are essential aspects of aesthetic experiences. Similar ideas have been advanced by other authors, primarily in the domains of music (Meyer, 1956; Narmour, 1992; Huron, 2006) and humor (Rozin et al., 2006). In addition, in the domain of music, Bayesian methods have been used by others (Temperley, 2007) to analyze how *explanations* of deeper structure (akin to EVE′'s meaning) can be inferred from surface features (akin to EVE′'s signals). However, EVE′ has been applied across domains, to graphics (Burns, 2006*a*, 2014), humor (Burns, 2012), and music (Burns & Dubnov, 2006), with unique advantages as discussed further below.

3.5.1. EVE′'s probabilities

EVE′ offers an overall equation $X = Y \times Z$ for how factors of surprise (Y) and meaning (Z) are combined to compute the magnitude of aesthetic experience (X), that is, as a measure of beauty or creativity. EVE′ also offers equations $Y = -\log P(s_j)$ and $Z = [G' \times P(M_i|s_j) - G]$ that specify how each factor, Y and Z, is computed. In particular, Z is computed from the most probable posterior probability, $P(M_i|s_j)$, along with scaling factors G' and G, that depend on an individual's personal preferences; whereas Y is computed as the marginal entropy of the signal s_j on which the meaning M_i is conditioned.

These aspects of EVE′ can be contrasted to a notion of *Bayesian surprise*, which has been used by others (Itti & Baldi, 2009) in modeling low-level visual attention and has been suggested as a factor for assessing creativity (Maher & Fisher, 2012). More specifically, Bayesian surprise is computed (Itti & Baldi, 2009) as the relative entropy (Kullback & Leibler, 1951) between a posterior distribution (over a set of hypotheses) and a prior distribution (over the same set of hypotheses). EVE′ differs in computing *Shannon surprise* as the marginal entropy of a signal, $Y = -\log P(s_j)$, because expectations (E) and violations (V) refer to *signals*, not to posterior and/or prior *meanings* of signals. EVE′ then models *Bayesian meaning* as a posterior probability, which measures how completely the surprise of a violation (V) is resolved by an explanation (E').

In addition, with respect to surprise, and different from other models, EVE′'s expectations and violations are governed by the probabilities of *signal classes* in the context of *causal meanings*, rather than by the probabilities of individual signals in isolation. This feature of the model is discussed further in Section 3.5.2 below, but first it is useful to highlight why signal classes and causal meanings are important from a psychological perspective. As a simple example, assume a fair coin is tossed six times, and consider the probability of getting an all-heads sequence "hhhhhh" as opposed to the mixed sequence "hhthtt." Mathematically, these two sequences are equally probable, because any single sequence of six outcomes has a probability of $2^{-6} = 0.016$. However, psychologically most people think the mixed sequence is much more probable, based on a heuristic known as *representativeness* (Tversky & Kahneman, 1974; Kahneman et al., 1982; Kahneman, 2011) that reflects *classes* of outcomes and *causes* for outcomes.

More specifically, consider a set of two possible causes {A, B} for the sequence of coin tosses, where A indicates tosses are random and B indicates tosses are rigged (i.e., by some nonrandom property such as a weighted coin). Based on our knowledge of the outcomes that are likely to be produced by each cause, we expect A would produce sequences in class *a* that do not exhibit the regularity of all-heads (i.e., sequences like "hhthtt" that are *representative* of A), whereas we expect B would produce sequences in class *b* that do exhibit the regularity of all-heads (i.e., sequences like "hhhhhh" that are *representative* of B). Thus, using the numbers for "likely" and "unlikely" that were assumed in the earlier analysis of a Burma-Shave jingle, our subjective likelihoods for signals in classes *a* and *b* conditional on each of the causes A and B are as follows:

$$P(a|\mathrm{A}) = 0.90, \quad P(a|\mathrm{B}) = 0.10,$$

$$P(b|\mathrm{A}) = 0.10, \quad P(b|\mathrm{B}) = 0.90.$$

Now assuming the coin tosses are fair, which implies that A is true (and B is not true), the above likelihoods lead us to believe that an all-heads (class *b*) sequence is much less probable than a mixed (class *a*) sequence. That is, $P(b|\mathrm{A}) = 0.10 \ll 0.90 = P(a|\mathrm{A})$. Therefore, psychologically, we would be quite surprised to see "hhhhhh" (class *b*), and not surprised to see "hhthtt" (class *a*), even though mathematically the two sequences are equally probable.

Now returning to the case of a Burma-Shave jingle (or any other media experience), human expectations and violations of expectations are governed by the same sorts of subjective probabilities (i.e., likelihoods of *signal classes* in the context of *causal meanings*). The only difference is that, unlike coin tosses, the causes are more complex in the case of artist–audience communications that involve meaning (intent). That is, an artist has some *meaning* that he intends to convey, but his artwork provides only a *signal* of that meaning. The audience, in turn, must infer the most likely meaning, and in so doing, the audience is inferring the most likely cause (intent) of the effect (artwork).

3.5.2. EVE′'s computability

As described above, EVE′'s modeling of surprise can be distinguished from other models that do not address *signal classes* and *causal meanings*. A relevant example comes from a study (Coulson & Kutas, 2001) to test the two-stage theory of jokes characterized as "surprise followed by reestablishment of coherence" (Suls, 1972). The study attempted to control for surprise in one-line jokes versus nonjokes by replacing the last word *b* (which was the punch word of the one-liner) with another word *a* that was equally surprising, for example, as follows: "When I asked the bartender for something cold and full of rum, he recommended his . . . [*a* or *b*]," where *b* = "wife" for the joke and *a* = "daiquiri" for the nonjoke.

To obtain these two endings, the study asked a group of people to read the one-liner without the last word, and then to report the first word that came to mind as an ending. Apparently, *wife* and *daiquiri* were each reported at approximately the same low frequency (≈3%), so researchers assumed that these two endings had equal probabilities and hence were equally surprising. Clearly the two endings are *not equally surprising*. That is, after we read the setup that implies A = bartender is thinking about a cold rum cocktail, the last word *wife* (in class *b*, consistent with meaning B = not A) produces a large violation of expectation whereas the last word *daiquiri* (in class *a*) is consistent with our expectation.

In short, surprise is not a *signal* that can be measured and modeled in a task of *generation*. Rather, surprise is a *feeling* that must be measured and modeled in a task of *evaluation*. That is why EVE′ models surprise Y as a violation of expectation, computed as a marginal entropy $-\log P(s_j)$, where $P(s_j)$ is a marginal probability computed from likelihoods of *signal classes* in the context of *causal meanings*. EVE′ also models the inferred meaning by which a surprise is resolved, that is, as a posterior probability $P(M_i|s_j)$. Finally, EVE′ models the aesthetic measure of a media experience (humorous or otherwise) as the product of surprise and meaning, in a formulation unlike others in that EVE′'s individual factors and overall equation $X = Y \times Z$ are actually computable.

With respect to being computable, the frame-shifting theory of humor (Coulson et al., 2006) is said to involve "semantic and pragmatic reanalysis in which elements of the existing message-level representation are mapped into a new frame retrieved from long-term memory." However, this theory does not specify what variables are represented by a frame retrieved from long-term memory, or how those variables relate to the elements of message-level representations, or how any of these variables can be computed. Similarly, others have proposed functional requirements for a computational system to experience humor (Hurley et al., 2011, p. 121), which are outlined as involving "time-pressured, involuntary heuristic search for valid expectations, which generates mental spaces in which elements are constantly being tested." This theory also includes a notion of epistemic commitment, but once again there is no formal specification of what is computed or how it is computed.

EVE′'s advantage lies in mathematical modeling of variables and operations needed to compute an aesthetic experience, including hypothetical meanings, evidential signals, and associated conditional and marginal probabilities, all in a bounded-Bayesian framework of signal classes and causal meanings. This makes EVE′'s model computable, at least in principle, as demonstrated by the earlier analysis of a prototypical Burma-Shave jingle.

4. VALIDATION

A comprehensive test of EVE′ would begin by computing Y, Z, and X, much like Section 3, but for a number of Burma-Shave jingles. The model-predicted values Y_m, Z_m, and X_m for each jingle would then be compared to human-reported judgments Y_h, Z_h, and X_h for that jingle. However, this approach is not currently feasible because of the inputs required to compute Y_m and Z_m. As explained in Section 2.1, the inputs include categorical knowledge of possible meanings (M_i), as well as numerical values of prior probabilities $P(M_i)$ and likelihoods $P(s_j|M_i)$, for all signals (s_j) in the sample of jingles to be tested. These input requirements are beyond what today's artificial intelligence systems can provide, and that is why the probabilities needed for model demonstration in Section 3 were developed from natural intelligence via introspection. However, as one test of EVE′, it is feasible to compute model-predicted values X_m given human-reported values Y_h and Z_h as inputs. The resulting predictions $X_m = Y_h \times Z_h$ can then be tested against human ratings of creativeness X_h across a sample of Burma-Shave jingles.

Clearly this is only a partial test of the model, because it addresses the overall equation $X = Y \times Z$ without addressing the underlying computation of factors Y and Z. However, the test is not trivial, and it is valuable for several reasons. First, meaning (Z) has not previously been featured as a factor in other computational models of creativity, so the test will help support or refute a role for this factor. Second, EVE′ assumes that novelty is captured by surprise (Y), and that meaning (Z) is an informatic utility by which surprise is resolved. Thus, the test will establish whether EVE′ can predict human ratings of creativeness without additional measures of novelty and utility like those of other computational models. Third, a test that measures surprise (Y_h) and meaning (Z_h) directly would provide useful data for future research in validating computational models of the factors Y_m and Z_m.

Another reason for testing in the manner described above is that EVE′'s equation differs markedly from an early theory of aesthetics that remains influential even today (Rigau et al., 2007; Phillips et al., 2010; Galanter 2012). According to this theory, proposed by Birkhoff (1933), aesthetic measure denoted M is equal to O/C, where O is a measure of order (i.e., symmetry or harmony) and C is a measure of complexity. Birkhoff's formula $M = O/C$ has been tested in numerous psychological experiments, with mixed results in predicting the magnitude of aesthetic pleasure experienced by human participants (Eysenck, 1941, 1942, 1957; Kreitler & Kreitler, 1972). Unfortunately, none of these studies tested the equation $M = O/C$ directly, because the factors O and C were computed by experimenters using ad hoc assumptions (some similar to and some different from Birkhoff's own assumptions), rather than obtaining ratings of O and C from the same humans who are providing ratings of M. In effect, the tests have confounded how experimenters compute measures of O and C with how participants combine judgments of O and C into an overall judgment of M.

Conceptually, Birkhoff's variables can be mapped to EVE′'s variables as follows: M corresponds to EVE′'s X as a measure of aesthetic pleasure (i.e., beauty or creativity); O corresponds to EVE′'s Z as a measure of meaning (i.e., symmetry or harmony); and C corresponds to EVE′'s Y as a measure of surprise (i.e., entropy), which has often been used as a measure of complexity by others who adopted or adapted Birkhoff's model (Bense, 1965; Moles, 1966; Rigau et al., 2007). With this mapping, an experiment that collects human judgments X_h, Y_h, and Z_h can be used to test EVE′'s model $X = Y \times Z$ against Birkhoff's model $X = Z/Y$ and an inverse model $X = Y/Z$, as well as simpler models $X = Y$ and $X = Z$.

Thus, my approach to empirical testing is as follows: A number of Burma-Shave jingles are chosen as stimuli (see Table 1) for use in a survey of human judgments. Then, human subjects are asked to provide ratings of surprise (Y_h), meaning (Z_h), and creativeness (X_h) for each jingle. Next, for each jingle, the average human ratings for surprise (Y_h) and meaning (Z_h) are used to compute model-predicted values of creativeness (X_m) for five models, including EVE′ (see Table 2). Finally, across jingles, model-predicted X_m are compared to human-reported X_h as a test of how well each model matches the data for judgments of creativeness.

4.1. Methods

A sample of 20 jingles (Table 1) was taken from a collection of 600 published verses (Rowsome, 1965) based on three criteria. The first criterion relates to the general topics of jingles. A review of all published verses found that each could be fit into one of the following four categories: good things that happen to people who use Burma-Shave; bad things that happen to people who do not use Burma-Shave; distinguishing characteristics of Burma-Shave compared to its competitors; or public service messages that Burma-Shave provides to promote safe driving practices, such as the example jingle about *little shavers*. Thus, the sample of 20 jingles was chosen to include 5 jingles of each type.

Table 1. *Burma-Shave jingles used in the survey*

Jingle	Type	Date	Pun
1. *the happy golfer / finds with glee // the shave / that suits him / to a tee*	1	1935	Y
2. *he's the boy / the gals forgot // his line / was smooth / his chin was not*	2	1940	N
3. *put your brush / back on the shelf // the darn thing / needs a / shave itself*	3	1940	N
4. *thirty days / hath September // April / June and the / speed offender*	4	1960	Y
5. *Burma-Shave / was such a boom // they passed / the bride / and kissed the groom*	1	1950	N
6. *many a wolf / is never let in // because of the hair / on his /chinny-chin-chin*	2	1945	Y
7. *stores are full / of shaving aids // but all you need / is this / and blades*	3	1942	N
8. *if our road signs / catch your eye // smile / but don't forget / to buy*	4	1963	N
9. *this cooling shave / will never fail // to stamp / its user / first class male*	1	1959	Y
10. *substitutes / can let you down // quicker / than a / strapless gown*	2	1955	Y
11. *his brush is gone / so what'll we do // said / Mike Robe I / to Mike Robe II*	3	1950	Y
12. *slow down Pa / sakes alive // Ma missed signs / four / and five*	4	1955	N
13. *one shave lasts / all day through // face feels / cool and / smoother too*	1	1955	N
14. *bristles scratched / his cookie's map // that's what / made poor / Ginger snap*	2	1960	Y
15. *shaving brushes / you'll soon see'em // on the shelf / in some / museum*	3	1943	N
16. *her chariot / raced 80 per // they hauled away / what had / Ben Hur*	4	1950	Y
17. *tempted to try it / follow your hunch // be top banana / not one / of the bunch*	1	1960	Y
18. *dear lover boy / your photo came // but your doggone beard / won't fit / the frame*	2	1960	N
19. *thrifty jars for / stay at homes // handy tubes / for him / who roams*	3	1947	N
20. *drinking drivers / nothing worse // they put / the quart / before the hearse*	4	1959	Y

Note: A double slash (//) is used to delineate the two rhyming phrases, and a single slash (/) is used to delineate the two or three lines within each of those phrases.

The second criterion was to include jingles that appeared during all four decades of the long-running ad campaign, including the 1930s to the 1960s. The order in which jingles are presented in the survey is roughly chronological within each topic type, except for an additional constraint on puns. That is, 10 of the jingles involve puns, and 5 of these jingles are placed in the first half of the survey and the remaining half are placed in the last half of the survey.

A third criterion deals with the line-by-line rhythm of the verses. At the highest level, all 600 jingles are organized in two phrases that rhyme. At a lower level, some jingles have three lines in the first phrase (and two lines in the second phrase), whereas others have two lines in the first phrase (and three lines in the second phrase). Roughly half of the 600 verses are of each type, and the jingle about *little shavers* analyzed in Section 3 is an example of the former type with three lines in the first phrase. For the survey, all 20 jingles were selected to be of the latter type with two lines in the first phrase. This helps to control for differences in ratings that may arise from variations in rhythmic structure between jingles. It also equalizes the number of lines in the setup (two lines) and the buildup (two lines), under the assumption that the setup ends in a word that rhymes with the last word of the punch line.

Table 2. *Regression coefficients for EVE′'s model ($X = Y \times Z$) and other models*

Model	β	α	R^2
$X = Y \times Z$	0.14	1.5	0.70
$X = Y$	0.61	1.4	0.58
$X = Z$	0.17	2.5	0.06
$X = Y/Z$	0.30	2.2	0.25
$X = Z/Y$	–0.45	3.9	0.17

Each page of the survey presented one jingle in five lines, double-spaced, using all uppercase lettering to be consistent with how the jingles were originally painted on road signs. Beneath each jingle were three questions that subjects were asked to answer, after reading the jingle, on a 5-point scale as follows: *very slightly*, *slightly*, *moderately*, *strongly*, or *very strongly*. The three questions were the following: *Did you understand the verse? Did the ending surprise you?* and *Is the jingle creative?* Responses to these questions correspond to the modeled variables *Z*, *Y*, and *X*, respectively.

The 20 jingles were presented one at a time to each participant in an online survey, which was prefaced by background on Burma-Shave's product and ad campaign. This background included a description of the four hallmark topics and the example jingle about *little shavers* (which fits the fourth topic). All 20 jingles were presented in the same order to all participants. No counterbalancing was performed because the jingles were purposely arranged roughly chronologically (within each topic type) and because the object of the survey was to obtain ratings while holding the context for each rating as constant as possible across all human judges.

Along with a balanced testing of four topics over four decades of the ad campaign, the survey was also designed to explore the effectiveness of puns (Pollack, 2011) in the realm of creative advertising (van Mulken et al., 2005). Although a comedian's use of puns may often receive groans from the audience, many clever and funny jokes use puns with much success. This suggests that puns may be a particularly sensitive test of creativity, leading to some of the best and worst results depending on how *surprising* and how *meaningful* an audience finds a pun. Puns and other wordplay frequently (by some estimates up to 40%; see Leigh, 1994) appear in advertisements, so insight into what makes a good pun or a bad pun from an evaluative perspective would have practical implications for writing better ads from a generative perspective.

4.2. Results

4.2.1. Validating the EVE′ model

The survey of 20 Burma-Shave jingles was taken by 81 adult volunteers, including 52 females and 29 males, aged 18 to 87 years with a median age of 48. The analysis here is focused on average ratings across genders and ages, computed by converting the qualitative ratings to quantitative values as follows: *very slightly* = 1, *slightly* = 2, *moderately* = 3, *strongly* = 4, and *very strongly* = 5.

Figure 2 shows the mean ratings of creativeness (X_h) for all 20 jingles, with error bars representing standard errors of the means. The ratings varied across participants, with individual ratings ranging from 1 to 5 for each jingle. However, one-way analysis of variance for each variable (X_h, Y_h, Z_h) showed a significant difference in mean ratings (each $p <$ 1E-06) across the 20 jingles. Therefore, mean ratings for creativeness (X_h), surprise (Y_h), and meaning (Z_h) are the human data used to test various models (X_m) listed in Table 2.

Model comparisons require a consistent approach to control for the arbitrary scale of 1–5 used in converting qualitative ratings to quantitative values. Therefore, for the set of 20 jingles, human data X_h are regressed onto modeled values X_m in the same manner for each model. The regression results include two coefficients, β and α, along with a goodness-of-fit (coefficient of determination) statistic R^2. For example, the regression equation for EVE′'s model is $X_\mathrm{m} = \beta \times (Y_\mathrm{h} \times Z_\mathrm{h}) + \alpha$, and the associated R^2 measures how much of the variability in the data (X_h) across jingles can be accounted for by the model $X_\mathrm{m} = Y_\mathrm{h} \times Z_\mathrm{h}$. The regression results for this model, and other models, are presented in Table 2.

As seen in Figure 3, EVE′'s model of creativeness (X_m) accounts for $R^2 = 70\%$ of the variability in mean ratings (X_h) reported by human judges. As shown in Table 2, this is better or much better than the R^2 for other models tested in the same manner. Further comparisons to the other models are made below.

4.2.2. Examining simpler models

The model $X = Y$ implies that creativeness is solely a function of surprise. This model scores $R^2 = 58\%$, which is less than EVE′ but still high and better than all other models in Table 2. These results suggest that surprise is by far the most important factor in judgments of creativeness, but that conclusion must be tempered by two further observations. First, most of the Burma-Shave jingles received high ratings for meaning (mean $Z = 4.18$), as one would expect because the jingles were actual advertisements. All of the jingles were selected over the years by Burma-Shave's officers and directors, based on thousands of entries submitted each year in response to an annual jingle-writing contest. Presumably, the executives were primarily interested in jingles that would be coherent to consumers, and secondarily interested in the surprise factor. In that case, EVE′'s model $X = Y \times Z$ would be approximated by $X = Y$. Second, the surprise of reading a punch line is made possible by a strong meaning that is set up and built up before

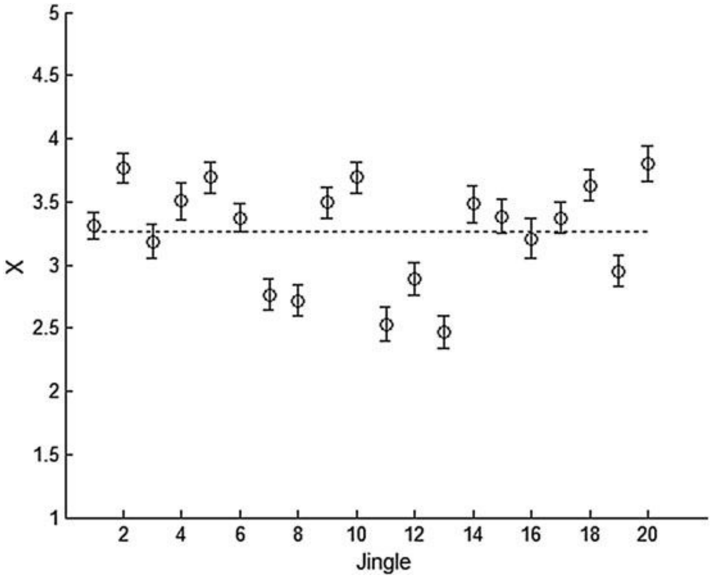

Fig. 2. Mean human ratings for creativeness (X_h) of 20 Burma-Shave jingles, across 81 human participants. Error bars represent standard errors of the means. Dotted line is mean rating across all 20 jingles.

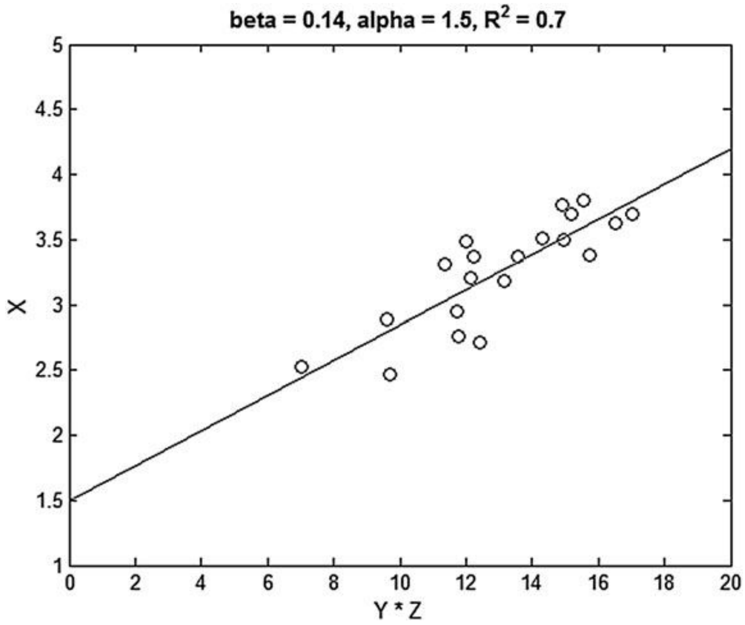

Fig. 3. Results of regressing human data for creativeness (X_h) onto EVE′'s model ($X_m = Y_h \times Z_h$).

being reversed in the punch line. Therefore, although *surprise* is clearly a very important factor, *meaning* is an equally important factor that cannot be ignored in the assessment of creativeness (Martindale et al., 1990).

Now with a focus on meaning, Table 2 shows that the model $X = Z$ performs much worse than the model $X = Y$, scoring only $R^2 = 6\%$. One reason for this is that increased creativeness is strongly predicted by increased surprise (discussed above), and in general it is more difficult for an audience to extract meaning as surprise is increased. Figure 4 shows this inverse relation between meaning and surprise, by plotting the mean values of Z_h versus Y_h for all 20 jingles. The regression of Z onto Y exhibits a negative slope, showing that meaning (Z) tends to decrease as surprise (Y) increases.

Thus, meaning has both a direct effect and an indirect effect on judgments of creativeness. The direct effect (independent of surprise) is that creativeness increases as meaning *increases*. The indirect effect (dependent on surprise) is that creativeness increases as meaning *decreases*. This is because creativeness increases as surprise increases (see $\beta = 0.61 > 0$

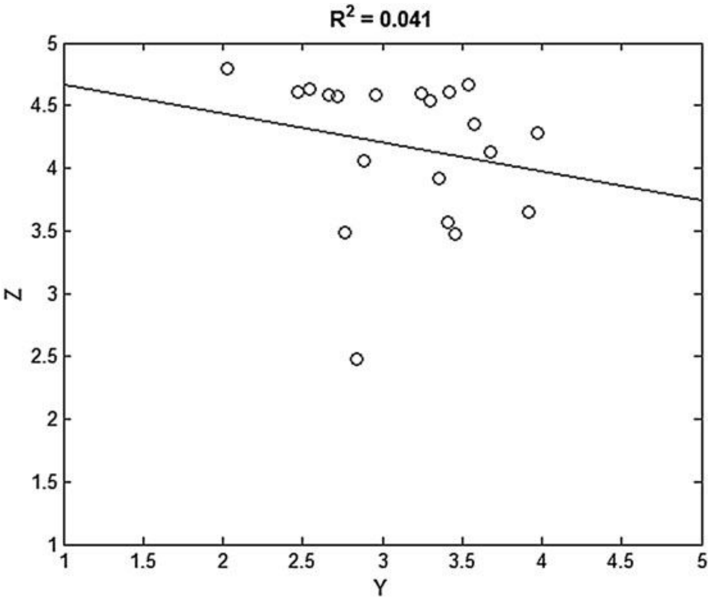

Fig. 4. Results of regressing human ratings of meaning (Z) onto human ratings of surprise (Y).

for $X = Y$ in Table 2) and meaning decreases as surprise increases (see Fig. 4). Although the direct effect is stronger (see $\beta = 0.17 > 0$ for $X = Z$ in Table 2), the two effects act in opposition, which helps to explain why Z alone accounts for little of the variability in X (see $R^2 = 6\%$ in Table 2). Nevertheless, the regression of Z onto Y in Figure 4 shows an R^2 of only 4%, so surprise (Y) accounts for little of the variability in meaning (Z). As such, surprise and meaning can be treated as relatively independent factors in the equation $X = Y \times Z$. This independence is necessary, if some Burma-Shave jingles are to be judged as highly creative, because a strong inverse relation between Y and Z would imply that only one factor (Y or Z) could be high for any given jingle.

4.2.3. *Critiquing Birkhoff's model*

Table 2 presents the results of two additional models. One is Birkhoff's model $X = Z/Y$, and the other is an inverse model $X = Y/Z$. Table 2 shows that the inverse model $X = Y/Z$ scores $R^2 = 25\%$, which is much worse than the EVE′ model and even worse than the simple model $X = Y$ that ignores Z.

Table 2 also shows that Birkhoff's model $X = Z/Y$ scores $R^2 = 17\%$, but with a highly negative coefficient β, which follows from the positive coefficient β of the inverse model $X = Y/Z$. This means that Birkhoff's model is negatively correlated with the human data, such that model-predicted creativity $X_m = Z_h/Y_h$ increases when the human-reported creativity X_h decreases, and vice versa. Therefore, of all the above models, Birkhoff's model is by far the worst, even worse than a null model with $\beta = 0$. This poor performance is driven by the fact that complexity (modeled by surprise Y) appears in the denominator of Birkhoff's equation, whereas the human data clearly show a highly direct (not inverse) relation between surprise and creativeness. Upon reflection, it seems clear that complexity cannot possibly appear in the denominator of the equation for aesthetic measure, because that would imply the very simplest (i.e., extremely predictable) stimuli would be judged as having the greatest beauty and creativity.

Birkhoff proposed $M = O/C$ as a measure of aesthetic beauty or pleasure, rather than a measure of creativity per se. However, it is hard to imagine cases where aesthetic beauty and creativity would be seen as opposites rather than correlates, and yet Birkhoff's model is implying exactly such an opposition. Moreover, even for judgments of beauty rather than creativity, Eysenck (1941, 1942 1957) reviewed how well Birkhoff's model performed in a wide range of human experiments across diverse media (graphics, sculpture, music, and poetry). Based on his review, Eysenck suggested that $M = O \times C$ was a better model than $M = O/C$. Note that $M = O \times C$ is analogous to EVE′'s model $X = Y \times Z$, per the above-noted mapping of Y to C and Z to O.

4.2.4. *Creating fun for readers*

An underlying assumption of EVE′ in this study is that aesthetic pleasure provides a gauge by which humans judge the creativeness of advertisements. As a test of this assumption, the survey included a final question after all jingles were rated: *Would you say the most creative jingles were also the most amusing?* The answer was reported on the same 5-point scale as that used to rate jingles themselves (i.e., 1 = *very slightly*, 5 = *very strongly*). The mean response was 4.42, so clearly participants felt quite strongly that their judgments of creativeness were correlated with the amusement they experienced. Of course, this does not prove that feelings of pleasure *caused* judgments of creativeness to vary as they did from jingle to jingle. However, that conjecture seems more plausible than the opposite, which would imply that judgments of creativeness (made in some unspecified fashion) somehow preceded and affected feelings of amusement.

In short, EVE′ predicts that arousal from surprise and appraisal of meaning together give rise to aesthetic pleasure, which in turn forms the basis for judgments of beauty and creativity. Although the present study cannot prove the postulated causality, empirical findings strongly support an association between feelings of amusement and judgments of creativeness.

4.2.5. *Analyzing the case of puns*

A final analysis of the human data was performed to test for differences between verses with and without puns. As noted earlier, the survey of 20 jingles was designed to include 10 jingles with puns and 10 jingles without puns. The mean rating of creativeness was found to be higher for jingles with puns ($X = 3.38$) than for jingles without puns ($X = 3.14$), although a Student t test showed no significant difference ($p = 0.21$) between the two means.

This result suggests that puns and wordplay are not inherently more or less creative than other forms of wit and humor (Pollack, 2011). Instead, feelings of amusement and judgments of creativeness are predicted by the product of *surprise* and *meaning*, regardless of the linguistic mechanism (Ritchie, 2004) employed by the author to provoke surprise and evoke meaning. Two insightful examples are jingle #13 (without a pun) and jingle #11 (with a pun), which were the two jingles that received the lowest ratings for creativity. Results for these two jingles are compared below, to show how each fails for a different reason: one because it is not surprising and the other because it is not meaningful.

Jingle #13 reads as follows: *one shave lasts / all day through // face feels / cool and / smoother too*. The ending of this jingle was judged as having the least surprise of any jingle ($Y = 2.02$, mean = 3.14). Thus, consistent with the EVE′ model, even if jingle #13 is perfectly understood, it should be judged as having very little creativity. This jingle received the highest rating for meaning ($Z = 4.79$, mean = 4.18) and the lowest rating for creativity ($X = 2.47$, mean = 3.26). In addition, in answer to a question at the end of the survey, none of the 81 subjects chose jingle #13 as being one of the three most creative jingles.

These results for jingle #13 (without a pun) can be contrasted to those for jingle #11 (with a pun), which reads as follows: *his brush is gone / so what'll we do // said / Mike Robe I / to Mike Robe II*. This jingle received the second-lowest rating of all jingles for creativity ($X = 2.53$), which was only slightly higher than the rating of jingle #13 ($X = 2.47$). However, unlike jingle #13, which received the highest rating for meaning ($Z = 4.79$), jingle #11 received the lowest rating for meaning ($Z = 2.48$). This Z rating was far lower than for any other jingle (see the lowest point in Fig. 4). Based on posttest interviews with selected participants, it was clear that most readers simply did not understand the pun. That is, they did not recognize that "Mike Robe" could be a "microbe" living in a damp shaving brush.

However, some of the 81 subjects did understand the pun of "Mike Robe," and their ratings provide a further and more focal test of the EVE′ model. According to EVE′, subjects who do decipher the obscure pun should feel especially pleased at their explanatory success, and therefore find jingle #11 to be one of the most creative jingles. Consistent with that conjecture, 6 subjects chose jingle #11 as being one of the three most creative jingles in answer to a question at the end of the survey. For this jingle, the mean ratings of $Z = 3.67$ and $X = 3.67$ by these 6 subjects were much higher than the average ratings (across all 81 subjects) of $Z = 2.48$ and $X = 2.53$, respectively.

To recap, it is the product of surprise and meaning that governs feelings of pleasure and judgments of creativeness. These two factors pose a design trade-off to artists, which makes creativity challenging, because in general it is more difficult for an audience to extract large amounts of meaning from signals that present large amounts of surprise. The trade-off is illustrated in Figure 1, where we see that signals with very large surprise are especially risky but potentially most rewarding. On the left of Figure 1, the large surprise brings an audience high pleasure if they can find much meaning in it, but low pleasure (or even displeasure) if they cannot find much meaning in it. For jingle #11 about "Mike Robe," only a few subjects extracted much meaning from the pun and hence made it to the upper left of Figure 1. Instead, most subjects were stuck on the lower left of Figure 1, so the mean rating of creativeness across all subjects was very low. On the right of Figure 1, small surprise holds little potential for aesthetic pleasure even if the meaning is completely understood. This occurred for jingle #13, so the mean rating of creativeness was very low, but for a different reason (i.e., not surprising) than for jingle #11.

5. DISCUSSION

5.1. Contributions

This paper makes two contributions to the computational study of creativity. The first is a computational model of how humans evaluate the creativeness of aesthetic artifacts. This model, dubbed EVE′, combines concepts of Shannon entropy and Bayesian probability into an integrated theory of aesthetic beauty and artistic creativity. The second contribution is empirical testing of how humans judge the creativeness of amusing advertisements. The results show that creativity is computed as a product of surprise and meaning, which substantiates the model's main equation $X = Y \times Z$ and differentiates it from other models of aesthetics and creativeness.

These contributions have implications for advancing computational creativity in artificial intelligence and engineering evaluations, as discussed in Section 5.3.

5.2. Objections

One objection to my approach might be that humans do not carry probability estimates around in their heads. However, humans do carry such estimates, as the strengths of associative memories between causes and effects. That is, *signals* are evidential effects and *meanings* are hypothetical causes of those effects, so the vast store of cause-and-effect knowledge that humans possess is representing exactly the sorts of likelihoods needed as input to a Bayesian model, that is, the probability $P(s_j|M_i)$ that a meaning (M_i) will cause a signal (s_j).

Another objection is that a Bayesian approach computes posterior probabilities by combining prior probabilities $P(M_i)$ with conditional likelihoods $P(s_j|M_i)$, such that temporal and contextual effects will arise. Actually that is a strength of the model, rather than a weakness. The reason is that all judgments are made in the context of background knowledge and current beliefs, so context-invariant approaches are not adequate to model how humans evaluate creativity (Martindale, 1990).

A related objection might be that, even if the model is theoretically plausible, it would be impossible to obtain estimates for the likelihoods required as input. However, Section 3 demonstrated how these inputs can be obtained from natural intelligence, via introspection, based on intuitive notions like the consistency of signal classes to what would be expected if various causal meanings were true.

Finally, one might object to a Bayesian framework for cognitive modeling because humans are not perfectly rational. However, my approach does not imply that humans are perfect Bayesians, and EVE′ includes important aspects of bounded rationality that are well established from previous research on heuristics and biases (Tversky & Kahneman, 1974; Kahneman et al., 1982; Kahneman, 2011). The analysis of Section 3 demonstrated, in a bounded-Bayesian fashion, that *incomplete* representations of probabilities in working memory are actually *necessary* for an audience to achieve the aesthetic experience. Thus, there is nothing inconsistent about using a bounded-Bayesian formulation to model human-heuristic cogitations. In addition, as discussed in Section 3.5, the bounded-Bayesian framework of EVE′ offers important advantages over informal theories that do not specify how cognitive processes (such as expectation and explanation) and affective responses (such as surprise and pleasure) can be modeled computationally.

5.3. Applications

The subsections below offer specific suggestions as to how EVE′'s computational model and empirical findings can be applied in artificial intelligence and engineering evaluations.

5.3.1. Artificial intelligence

EVE′ is a computational model of aesthetics, which may be implemented in machine systems for evaluating (and generating) creativity in domains of fine and applied arts. The present study did not implement such a system, because the purpose was to motivate the model's formulation, demonstrate the model's computations, and validate the model's predictions (at least to some extent) against human judgments of creativeness. Based on these contributions, EVE′ illustrates what would be needed for a machine system to compute aesthetics, and thereby evaluate the creativity of artworks generated by man or machine.

According to EVE′, a system for evaluating artistic creativity would require rich semantic knowledge, including conceivable intentions of the artist (Dennett, 1987; Cohen, 2010), in the form of hypothetical *meanings* M_i for evidential *signals* s_j that an audience might extract from artworks. A system would also require probabilistic knowledge representing prior (before signal) confidences in meanings, $P(M_i)$, and likelihoods of signal classes assuming the truth of each causal meaning, $P(s_j|M_i)$, in order to compute the marginal probabilities of signals: $P(s_j) = \sum_i P(s_j|M_i) \times P(M_i)$. The system could then use EVE′'s equations to compute marginal entropies of signals ($Y = \text{fn}(P(s_j))$), posterior probabilities of meanings ($Z = \text{fn}(P(M_i|s_j))$), and aesthetic beauty ($X = Y \times Z$) as a measure of artistic creativity. Examples of all these semantic and probabilistic inputs and outputs of EVE′ were provided in Section 3.

By explicating these knowledge requirements, EVE′ is useful for understanding the inherent difficulties of evaluating creativity, and for identifying domains of design in which they might feasibly be addressed by artificial intelligence. For example, elsewhere EVE′ has been implemented (and further validated) in a system to evaluate and generate abstract artworks composed of Mondrian-like grid-line patterns (Burns, 2014). The grid-line constraint on these artworks greatly reduced the semantic and probabilistic knowledge required as input to EVE′, and thereby enabled machine implementation of the model.

Music is another area where practical applications appear promising. Much like the "lines" of an abstract visual design, the "notes" of a musical composition are *signals* that do not refer to objects or other aspects of a world outside the artwork itself. This can be contrasted to figurative art and narrative verse, where the *meanings* of signals do pertain to a world outside the artwork, which exacerbates the semantic and aesthetic problems of understanding and appreciating the artwork (Parsons, 1987; Danto, 2013).

In music, it is currently possible to construct probabilistic models from signal statistics computed within and across compositions (Dubnov, 2010). Using these knowledge models, surprise can be computed as a marginal entropy for each signal s_j as it is received in the temporal sequence of a musical score. Using the same knowledge models, contextual sequences of adjacent signals S_i can be used to compute conditional probabilities that measure how completely the surprise of each signal s_j is resolved. Music compositions have been evaluated in this manner, using EVE′'s equations and an equivalent formulation whereby an "information rate" models the magnitude of an audience's emotion (Burns & Dubnov, 2006; Dubnov et al., 2006; Dubnov, 2010).

5.3.2. Engineering evaluations

The present study provided computational modeling and empirical testing of how humans assess aesthetic creativeness. The testing was limited to an artistic domain of advertising design, so the empirical results cannot establish the relative importance of aesthetic form and pragmatic function(s) for engineered artifacts in which both form and function contribute to creativity. However, EVE′'s computational model does offer unifying insights into how humans assess the creativeness of any artifact as discussed below.

One insight is that aesthetics can be pragmatic (i.e., when a consumer's emotional response to an artifact accomplishes a practical function intended by the creator of the artifact). For example, the pleasure arising from an amusing advertisement can persuade consumers to buy a product, which satisfies the main function of advertising. Similarly, in the entertainment industry, aesthetic form and pragmatic function are one in the same because the purpose of entertainment is to engage audiences in pleasurable experiences.

Another insight is that the evaluation of creativity is a *judgment* (Brown, 2013) and hence governed by psychological processes. According to EVE′, the processes of *arousal* and *appraisal* combine to produce a feeling of pleasure, which is how a person gauges the creativeness of an artifact. This model is consistent with how humans make judgments of other quantities involving probabilities (surprise) and utilities (value), based on feelings or *affect* (Finucane et al., 2000; Gilovich et al., 2002; Kahneman, 2011).

Thus, EVE′ suggests that the creativeness of engineered artifacts and artworks are evaluated using the same psychological processes. The main difference lies in the dimension of utility, which for some artifacts will include pragmatic functions as well as aesthetic form. However, the same basic product of *arousal* × *appraisal* applies to each aspect of utility, regardless of whether it involves a pragmatic function or an aesthetic form.

This insight points to the importance of modeling and measuring psychological phenomena. Existing computational models for evaluating creativity (Ritchie, 2001, 2007; Maher, 2010; Maher & Fisher, 2012; Grace et al., 2014, in press) have not modeled cognitive–affective processes, and they have not been tested against human judgments of creativeness or contributing factors. The present study differs in modeling psychological processes and performing empirical testing. The resulting model suggests how factors of novelty, utility, surprise, and value can be combined to compute judgments of creativeness as outlined below.

Some computational models (Grace et al., 2014, in press) assume that novelty measures the difference(s) between an artifact and other artifacts, whereas surprise measures the difference(s) between an artifact and *expectations* formed from previously experienced artifacts. However, it is also important to acknowledge that surprise always occurs in the context of an *experience*. Specifically, if an artifact exhibits novelty and is experienced by an evaluator, then the perceived novelty is what produces *surprise*. Similarly, human judgments of utility are always made relative to personal and cultural wants and needs, so utility is a *value* judgment made by the one who is experiencing the artifact and evaluating creativity.

In that light, I would suggest the terms *surprise* and *value* can serve as synonyms for *subjective novelty* and *subjective utility*, respectively. I would also suggest that judgments of novelty and utility are always subjective, because they are judgments and hence dependent on the knowledge and experiences of the one(s) doing the judging. Finally, for aesthetics, EVE′'s model suggests that the utility of concern is an informatic value, which represents the semantic *meaning* by which *surprise* is resolved. With that terminology, EVE′'s product of *arousal* × *appraisal* = *surprise* × *value* can be extended to an equation for evaluating the creativity of any artifact, as follows:

$$\text{Creativity} = \sum_{k} \text{Applicability}_k \times \text{Novelty}_k \times \text{Utility}_k,$$

where the sum is taken over all *k* aspects of an artifact that are judged to be applicable to that artifact, including various pragmatic functions as well as aesthetic form. Each term in this equation is a *subjective* assessment (i.e., a judgment) and hence sensitive to the knowledge and experiences of the one(s) evaluating creativity.

Referring to the three factors in this equation, applicability is a subjective assessment of *weighting* (i.e., fraction of a total 1.0) or importance for aspect *k* relative to other aspects of the artifact. For example, aesthetic form might have a weight of 1.0 for an artwork but lower weight for an engineered artifact in which pragmatic functions are important. Novelty is a subjective assessment of *surprise* that quantifies arousal with respect to aspect *k* (and may be different for each aspect *k*). Utility is a subjective assessment of *value* that quantifies appraisal with respect to aspect *k*. The product of *arousal* × *appraisal* = *surprise* × *value*, with respect to some aspect *k*, is a measure of the pleasure or other positive affect that is used to gauge the creativeness of the artifact with respect to that aspect *k*. The weighted sum, over all *k* applicable aspects, represents an overall measure of creativity for the artifact.

For pragmatic functions, the applicable aspects of utility may include efficacy, efficiency, usability, affordability, reliability, and other properties that are valued by consumers. For aesthetic form, the applicable utility is an informatic value, which represents semantic meaning and is of ultimate importance in the appreciation of artworks (Parson, 1987; Danto, 2013). Actually, the same sort of informatic utility applies to many engineered artifacts with pragmatic functionality, such that aesthetics may be pragmatic in engineering much like in advertising and entertainment. A case in point is the engineering of information systems, where the artifacts of interest include the outputs computed by systems and consumed by users. These outputs provide users with information that is relevant to task demands, hence *meaningful*; and not readily available from other sources or systems, hence *surprising*. Therefore, in some sense the information systems are analogous to synthetic artists who create aesthetic messages that combine surprise and meaning.

Of course, it is arguable to what extent such systems are aesthetic or creative in their computations. However, the analogy is useful here because it suggests how research on aesthetic creativity can be applied to the engineering of systems designed for informatic functionality. A relevant example comes from the field of intelligence analysis (George & Bruce, 2008), where informatic creativity is of critical importance. Intelligence analysts often speak of "anticipating surprise" (Grabo, 2004) as they seek to create knowledge in the form of indications and warnings for use by strategic policy makers and tactical commanders. This requires the identification and analysis of available information in order to explain the causes of world events that have occurred (or are occurring) and predict the courses of future events. Information systems are designed to support analysts who create such actionable intelligence in a process known as "sensemaking" (IARPA, 2010; Klein et al., 2006*a*; 2006*b*).

The underlying psychology of sensemaking has been previously characterized by a "data-frame" theory (Klein et al., 2007), much like the notion of frame-shifting that others have used to explain humor (see Section 3.5.2). However, like frame-shifting theories of humor, the data-frame theory does not specify how sensemaking can be computed, as needed for validation in human experiments and implementation in artificial intelligence. A computational model of sensemaking has been developed more recently (Burns, 2005, 2011, in press-*a*) using the Bayesian framework employed by EVE′. This model has been used to engineer software systems for administering and analyzing human experiments, on problems that pose cognitive challenges of intelligence analysis (Burns, 2011, in press-*b*). The model has also been used to identify facets of sensemaking that might feasibly be accomplished by artificial intelligence, and to guide the design of advanced information systems that can support the analytical efforts of natural intelligence (Burns, 2006*b*, in press-*a*). A key function of these systems is to help humans extract all the certainty that is available from uncertain sources of information (Edwards et al., 1968; Edwards, 1982), and thereby obtain the informatic utility needed for anticipating surprise.

6. CONCLUSION

This article addressed the computational evaluation of creativity, focusing on the aesthetics of amusing advertisements. A formal model was developed to analyze these applied artworks, and the model was extended beyond aesthetics to include pragmatic aspects of engineered artifacts. In this model, judgments of creativeness are computed as a product of subjective novelty (surprise) and subjective utility (value), over all applicable aspects of an artifact, including form and function(s).

For aesthetic form, the model has been tested against human judgments in the present study of amusing advertisements and in another study of abstract artworks (Burns, 2014). For pragmatic functions, other authors have proposed a similar product of novelty and utility to evaluate the creativeness of engineered artifacts (Srinivasan & Chakrabarti, 2010; Sarkar & Chakrabarti, 2011). Thus, one direction for future research is to test the present model on artifacts for which pragmatic function(s) and aesthetic form both contribute to judgments of creativity (i.e., as a sum of *novelty* × *utility* over all applicable aspects of form and function).

Another direction for future research is to specify and validate how subjective novelty (surprise) and subjective utility (value) can be computed for various pragmatic functions, as well as for various aesthetic forms. The present study was limited to one form of applied artworks, and the testing required that human judgments of contributing factors (i.e., surprise and meaning) be input to the model. Future research might attempt to construct the knowledge and value structures needed to compute these inputs directly, using artificial intelligence without support from natural intelligence.

Like the present research, future research in either direction would require human judgments of creativeness and contributing factors, in order to establish that the artificial intelligence can compute judgments consistent with those of natural intelligence.

REFERENCES

Baddeley, A. (1992). Working memory. *Science 255*, 556–569.

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions 53*, 370–418.

Bense, M. (1965). *Aesthetica: Einfürung in die Neue Aesthetik*. Baden-Baden: Agis-Verlag.

Berlyne, D. (1957). Uncertainty and conflict: a point of contact between information-theory and behavior-theory concepts. *Psychological Review 64(6)*, 329–339.

Berlyne, D. (1963). Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology 17*, 274–290.

Berlyne, D. (1970). Novelty, complexity, and hedonic value. *Perception and Psychophysics 8(5)*, 279–286.

Berlyne, D. (1971). *Aesthetics and Psychobiology*. New York: Appleton Century Crofts.

Besemer, S. (1998). Creative product analysis matrix: testing the model structure and a comparison among products—three novel chairs. *Creativity Research Journal 11(4)*, 333–346.

Besemer, S., & O'Quin, K. (1986). Analyzing creative products: refinement and test of a judging instrument. *Journal of Creative Behavior 20(2)*, 115–126.

Besemer, S., & O'Quin, K. (1999). Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal 12(4)*, 287–296.

Besemer, S., & Treffinger, D. (1981). Analysis of creative products: review and synthesis. *Journal of Creative Behavior 15(3)*, 158–178.

Binstead, K. (2006). Computational humor. *IEEE Intelligent Systems 21*, 22–32.

Birkhoff, G. (1933). *Aesthetic Measure*. Cambridge, MA: Harvard University Press.

Blinderman, C. (1962). T. H. Huxley's theory of aesthetics: unity in diversity. *Journal of Aesthetics and Art Criticism 21(1)*, 49–55.

Boden, M. (1991). *The Creative Mind: Myths and Mechanisms*. New York: Basic Books.

Boden, M. (2009). Computer models of creativity. *AI Magazine* Fall, 23–34.

Brown, D. (2013). Developing computational design creativity systems. *International Journal of Design Creativity and Innovation 1(1)*, 43–55.

Burns, K. (2005). Mental models and normal errors. In *How Professional Make Decisions* (Montgomery, H., Lipshitz, H., & Brehmer, B., Eds.), pp. 15–28. Mahwah, NJ: Erlbaum.

Burns, K. (2006*a*). Atoms of EVE′: a Bayesian basis for esthetic analysis of style in sketching. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing 20(3)*, 185–199.

Burns, K. (2006*b*). Bayesian inference in disputed authorship: a case study of cognitive errors and a new system for decision support. *Information Sciences 176(11)*, 1570–1589.

Burns, K. (2011). The challenge of iSPIED: intelligence sensemaking to prognosticate IEDs. *International C2 Journal 5(1)*, 1–36.

Burns, K. (2012). EVE′'s energy in aesthetic experience: a Bayesian basis for haiku humor. *Journal of Mathematics and the Arts 6*, 77–87.

Burns, K. (2014). *Entropy and optimality in abstract art: an empirical test of visual aesthetics*. Manuscript submitted for publication.

Burns, K. (in press-*a*). *A Computational Basis for ICArUS Challenge Problem Design* (MITRE Technical Report MTR140415). McLean, VA: MITRE.

Burns, K. (in press-*b*). *ICArUS Phase 2 Challenge Problem Design* (MITRE Technical Report MTR140412). McLean, VA: MITRE.

Burns, K., & Dubnov, S. (2006). Memex music and gambling games: EVE′'s take on lucky number 13. *Proc. AAAI Workshop on Computational Aesthetics: Artificial Intelligence Approaches to Beauty and Happiness, WS-06-04*, pp. 30–36. Menlo Park, CA: AAAI Press.

Burns, K., & Maybury, M. (2010). The future of style. In *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning* (Argamon, S., Burns, K., & Dubnov, S., Eds.), pp. 317–332. Berlin: Spinger–Verlag.

Butcher, S., Trans. (1951). *Aristotle Poetics*. New York: Dover.

Cohen, H. (2010). Style as emergence (from what?). In *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning* (Argamon, S., Burns, K., & Dubnov, S., Eds.), pp. 3–20. Berlin: Springer–Verlag.

Coulson, S., & Kutas, M. (2001). Getting it: human event-related brain responses to jokes in good and poor comprehenders. *Neuroscience Letters 316(2)*, 71–74.

Coulson, S., Urbach, T., & Kutas, M. (2006). Looking back: joke comprehension and the space structuring model. *Humor 19(3)*, 229–250.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences 24(1)*, 87–114.

Danto, A. (2013). *What Art Is*. New Haven, CT: Yale University Press.

Demaree, H., DeDonno, M., Burns, K., & Everhart, D. (2008). You bet: how personality differences affect risk-taking preferences. *Personality and Individual Differences 44(7)*, 1484–1494.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dubnov, S. (2010). Information dynamics and aspects of musical perception. In *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning* (Argamon, S., Burns, K., & Dubnov, S., Eds.), pp. 127–157. Berlin: Spinger–Verlag.

Dubnov, S., McAdams, S., & Reynolds, R. (2006). Structural and affective aspects of music form from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology 57(11)*, 1526–1536.

Edwards, W. (1982). Conservatism in human information processing. In *Judgment Under Uncertainty: Heuristics and Biases* (Kahneman, D., Slovic, P., & Tversky, A., Eds.), pp. 359–369. Cambridge: Cambridge University Press.

Edwards, W., Phillips, L., Hayes, W., & Goodman, B. (1968). Probabilistic information processing systems: design and evaluation. *IEEE Transactions on Systems, Man, and Cybernetics 4(3)*, 248–265.

Eysenck, H. (1941). The empirical determination of an aesthetic formula. *Psychological Review 48(1)*, 83–92.

Eysenck, H. (1942). The experimental study of "good gestalt." *Psychological Review 49(4)*, 344–364.

Eysenck, H. (1957). *Sense and Nonsense in Psychology*. Harmondsworth: Penguin.

Finucane, M., Alhakami, A., Slovic, P., & Johnson, S. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making 13(1)*, 1–17.

Galanter, P. (2012). Computational aesthetic evaluation: past and future. In *Computers and Creativity* (McCormack, J., & d'Inverno, M., Eds.), pp. 255–293. Berlin: Springer–Verlag.

George, R., & Bruce, J. (2008). *Analyzing Intelligence: Origins, Obstacles, and Innovations*. Washington, DC: Georgetown University Press.

Gero, J. (2010). Future directions for design creativity research. In *Design Creativity 2010* (Taura, T., & Nagai, Y., Eds.), pp. 15–22. Berlin: Springer–Verlag.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Grabo, C. (2004). *Anticipating Surprise: Analysis for Strategic Warning*. Lanham, MD: University Press of America.

Grace, K., Maher, M., Fisher, D., & Brady, K. (2014). Modeling expectation for evaluating surprise in design creativity. In *Design Computing and Cognition* (Gero, J., Ed.), pp. 201–220. Berlin: Springer–Verlag.

Grace, K., Maher, M., Fisher, D., & Brady, K. (in press). A data-intensive approach to predicting creative designs based on novelty, value, and surprise. *International Journal of Design, Creativity, and Innovation*.

Hofstadter, A., & Kuhns, R. (1964). *Philosophies of Art and Beauty: Selected Readings in Aesthetics from Plato to Heidegger*. Chicago: University of Chicago Press.

Horn, D., & Salvendy, G. (2006). Product creativity: conceptual model, measurements, and characteristics. *Theoretical Issues in Ergonomics Science 7(4)*, 395–412.

Hurley, M., Dennett, D., & Adams, R. (2011). *Inside Jokes: Using Humor to Reverse-Engineer the Mind*. Cambridge, MA: MIT Press.

Huron, D. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press.

IARPA. (2010). *Integrated Cognitive-Neuroscience Architectures for Understanding Sensemaking* (Broad Agency Announcement, Intelligence Advanced Research Projects Activity, IARPA-BAA-10-04). Washington, DC: Author.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research 49(10)*, 1295–1306.

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

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Strauss & Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Klein, G., Moon, B., & Hoffman, R. (2006*a*). Making sense of sensemaking 1: alternative perspectives. *IEEE Intelligent Systems 21(4)*, 70–73.

Klein, G., Moon, B., & Hoffman, R. (2006*b*). Making sense of sensemaking 2: a macrocognitive model. *IEEE Intelligent Systems 21(5)*, 88–92.

Klein, G., Phillips, J., Rall, E., & Peluso, D. (2007). A data-frame theory of sensemaking. In *Expertise Out of Context* (Hoffman, R., Ed.), pp. 113–155. New York: Erlbaum.

Kreitler, H., & Kreitler, S. (1972). *Psychology of the Arts*. Durham, NC: Duke University Press.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics 22(1)*, 79–86.

Leigh, J. (1994). The use of figures of speech in print ad headlines. *Journal of Advertising 23(2)*, 18–33.

Maher, M. (2010). Evaluating creativity in humans, computers, and collectively intelligent systems. *Proc. DESIRE'10: Creativity in Innovation and Design*. Aarhus, Denmark.

Maher, M., & Fisher, D. (2012). Using AI to evaluate creative designs. *Proc. 2nd. Int. Conf. Design Creativity*, pp. 45–54.

Martindale, C. (1990). *The Clockwork Muse: The Predictability of Artistic Change*. New York: Basic Books.

Martindale, C., Moore, K., & Borkum, J. (1990). Aesthetic preference: anomalous findings for Berlyne's psychobiological theory. *American Journal of Psychology 103(1)*, 53–80.

McGrayne, S. (2011). *The Theory That Would Not Die: How Bayes Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy*. New Haven, CT: Yale University Press.

Meyer, L. (1956). *Emotion and Meaning in Music*. Chicago: University of Chicago Press.

Meyers-Levy, J., & Malaviya, P. (1999). Consumers' processing of persuasive advertisements: an integrative framework of persuasion theories. *Journal of Marketing 63*, 45–60.

Moles, A. (1966). *Information Theory and Esthetic Perception*. Urbana, IL: University of Illinois Press.

Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity: The Implication–Realization Model*. Chicago: University of Chicago Press.

Norman, D. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*. New York: Basic Books.

Oatley, K. (2003). Creative expression and communication of emotions in the visual and narrative arts. In *Handbook of Affective Sciences* (Davidson, R., Scherer, K., & Goldsmith, H., Eds.), pp. 481–502. Oxford: Oxford University Press.

O'Quin, K., & Besemer, S. (1989). The development, reliability, and validity of the revised creative product semantic scale. *Creativity Research Journal 2(4)*, 267–278.

Parsons, M. (1987). *How We Understand Art: A Cognitive Developmental Account of Aesthetic Experience*. Cambridge: Cambridge University Press.

Paulos, J. (1980). *Mathematics and Humor*. Chicago: University of Chicago Press.

Phillips, F., Norman, J., & Beers, A. (2010). Fechner's aesthetics revisited. *Seeing and Perceiving 23(3)*, 263–271.

Pollack, J. (2011). *The Pun Also Rises: How the Humble Pun Revolutionized Language, Changed History, and Made Wordplay More Than Some Antics*. New York: Gotham.

Postrel, V. (2003). *The Substance of Style: How the Rise of Aesthetic Value Is Remaking Commerce, Culture, and Consciousness*. New York: Harper–Collins.

Rigau, J., Feixas, M., & Sbert, M. (2007). Conceptualizing Birkhoff's aesthetic measure using Shannon entropy and Kolmogorov complexity. In *Computational Aesthetics in Graphics, Visualization, & Imaging* (Cunningham, D., Meyer, G., & Neumann, M., Eds.), pp. 105–112. Goslar, Germany: Eurographics Association.

Ritchie, G. (2001). Assessing creativity. *Proc. AISB Symp. Artificial Intelligence and Creativity in Art and Science*, York.

Ritchie, G. (2004). *The Linguistic Analysis of Jokes*. London: Routledge.

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

Roeckelein, J. (2002). *The Psychology of Humor: A Reference Guide and Annotated Bibliography*. Westport, CT: Greenwood Press.

Rowsome, F. (1965). *The Verse by the Side of the Road: The Story of Burma-Shave Signs and Jingles with All 600 of the Roadside Rhymes*. New York: Plume.

Rozin, P., Rozin, A., Appel, B., & Wachtel, C. (2006). Documenting and explaining the common AAB pattern in music and humor: establishing and breaking expectations. *Emotion 6(3)*, 349–355.

Ruch, W. (1988). Sensation seeking and the enjoyment of structure and content of humor: stability of findings across four samples. *Personality and Individual Differences 9(5)*, 861–871.

Sarkar, P., & Chakrabarti, A. (2011). Assessing design creativity. *Design Studies 32(4)*, 348–383.

Scherer, K. (1999). Appraisal theory. In *Handbook of Cognition and Emotion* (Dalgleish, T., & Power, M., Eds.), pp. 637–663. New York: Wiley.

Shannon, C., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Silvia, P. (2005). Emotional responses to art: from collation and arousal to cognition and emotion. *Review of General Psychology 9(4)*, 342–357.

Silvia, P. (2006). *Exploring the Psychology of Interest*. New York: Oxford University Press.

Srinivasan, V., & Chakrabarti, A. (2010). Investigating novelty-outcome relationships in engineering design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing 24(2)*, 161–178.

Sternthal, B., & Craig, C. (1973). Humor in advertising. *Journal of Marketing 37*, 12–18.

Suls, J. (1972). A two-stage model for the appreciation of jokes and cartoons: an information processing analysis. In *The Psychology of Humor: Theoretical Perspectives and Empirical Issues* (Goldstein, J., & McGhee, P., Eds.), pp. 81–100. New York: Academic Press.

Temperley, D. (2007). *Music and Probability*. Cambridge, MA: MIT Press.

Thagard, P. (2007). Abductive inference: from philosophical analysis to neural mechanisms. In *Inductive Reasoning: Experimental, Developmental, and Computational Approaches* (Feeney, A., & Heit, E., Eds.), pp. 226–247. Cambridge: Cambridge University Press.

Tversky, A., & Kahneman, K. (1974). Judgment under uncertainty: heuristics and biases. *Science 185*, 1124–1131.

van Mulken, M., van, Enschot-van Dijk, R. & Hoecken, H. (2005). Puns, relevance, and appreciation in advertisements. *Journal of Pragmatics 37(5)*, 707–721.

Walvis, T. (2008). Three laws of branding: neuroscientific foundations of effective brand building. *Journal of Brand Management 16*, 176–194.

Weinberger, M., & Gulas, C. (1992). The impact of humor in advertising: a review. *Journal of Advertising 21*, 35–59.

Zuckerman, M. (1994). *Behavioral Expressions and Biosocial Bases of Sensation Seeking*. New York: Cambridge University Press.

Kevin Burns is a Principal Scientist at the MITRE Corporation and holds engineering degrees from the Massachusetts Institute of Technology. His interests are in cognitive computing as applied to hazardous operations, intelligence analysis, interactive entertainment, and artistic endeavors.