Computational Models of Surprise in Evaluating Creative Design

Mary Lou Maher[1], Katherine Brady[2], Douglas H. Fisher[2]
[1]Software Information Systems, University of North Carolina, Charlotte, NC
[2]Electrical Engineering & Computer Science, Vanderbilt University, Nashville, TN
m.maher@uncc.edu, katherine.a.brady@vanderbilt.edu, douglas.h.fisher@vanderbilt.edu

Abstract

In this paper we consider how to evaluate whether a design or other artifact is creative. Creativity and its evaluation have been studied as a social process, a creative arts practice, and as a design process with guidelines for people to judge creativity. However, there are few approaches that seek to evaluate creativity computationally. In prior work we presented *novelty*, *value*, and *surprise* as a set of necessary conditions when identifying creative designs. In this paper we focus on the least studied of these – surprise. Surprise occurs when expectations are violated, suggesting that there is a temporal component when evaluating how surprising an artifact is. This paper presents an approach to quantifying surprise by projecting into the future. We illustrate this approach on a database of automobile designs, and we point out several directions for future research in assessing surprising and creativity generally.

Evaluating Creativity and Surprise

As we develop partially and fully automated approaches to computational creativity, the boundary between human creativity and computer creativity blurs. We are interested in approaches to evaluating creativity that make no assumptions about whether the creative entity is a person, a computer, or a collective intelligence of human and computational entities. In short, we want a test for creativity that is not biased by the form of the entity that is doing the creating (Maher and Fisher 2012), but the test should be flexible enough to allow for many forms of creative output. Ultimately, such tests will imbue artificial agents with an ability to assess their own designs and will inform computational models of creative reasoning. Such tests will also inform the design of cognitive assistants that collaborate with humans in sophisticated, socially-intelligent systems.

Evaluating creativity by the characteristics of its results has a long history, including contributions from psychology, engineering, education, and design. Most descriptions of creative designs include *novelty* (sufficiently different from all other designs) and *value* (utilitarian and/or aesthetic) as essential characteristics of a creative artifact (Csikszenmihalyi & Wolfe, 2000; Amabile, 1996; Runco, 2007; Boden, 2003; Wiggins, 2006; Cropley & Cropley, 2005; Besemer & O'Quin, 1987; Horn & Salvendy, 2003; Goldenberg & Mazursky, 2002; Oman and Tumer, 2009; Shah, Smith, & Vargas-Hernandez, 2003).

Surprise is an aspect of creative design that is rarely given attention, even though we believe that it is distinct from novelty and value: a design can be both novel and valuable, but not be surprising. It may be tempting to think that surprise simply stems directly from its "novelty" or difference relative to the set of existing and known artifacts, but we believe that while surprise is related to novelty, it is distinct from novelty as that term is generally construed. In particular, surprise stems from a violation of expectations, and thus surprise can be regarded as "novelty" (or sufficient difference) in a space of projected or expected designs, rather than in a space of existing designs.

In earlier work, Maher and Fisher (2012) presented novelty, value, and surprise as essential and distinct characteristics of a creative design. They also forwarded computational models based on clustering algorithms, which were nascent steps towards automating the recognition of creative designs. This paper takes a closer look at surprise, adding an explicit temporal component to the identification of surprising designs. This temporal component enables a system to make projections about what designs will be expected in the future, so that a system can subsequently assess a new design's differences from expectations, and therefore judge whether a new design deviates sufficiently from expectations to be surprising.

AI Approaches for Assessing Surprise

There is little work on assessing surprise in computational circles; but there has been some, which we survey here.

Horvitz et al (2005) develop a computational model of surprise for traffic forecasting. In this model, they generate probabilistic dependencies among variables, for example linking weather to traffic status. They assume that when an event has less than 2% probability of occurring, it is marked as surprising. They temporally organize the data, grouping incidents into 15-minute intervals. Surprising events in the past are collected in a case library of surprises that is used to identify when a surprising event has occurred. Though related, the concept of rarity as an identifier of something surprising is not the same as difference ("novelty") as an interpretation of surprise – for example, perhaps the rare event differs on only one or two dimen-

sions from other events, and it is these slight differences that make the event rare, and thus surprising.

An important characteristic of the Horvitz et al model is that it makes time explicit, by grouping events into temporal intervals.

A possible limitation of considering rarity as an interpretation of surprise is that as rare events recur, as they are apt to do, many observers would regard them as less surprising. So conditioning surprise by prior precedent might be a very desirable addition to the model. Indeed, Rissland (2009) advances a case-based approach to reasoning about rare and transformative legal cases, where the first appearance of a rare case is surprising and transformative, but subsequent appearances of similar, but still rare events, are neither transformative, nor surprising.

While Rissland's research is not concerned with computational assessment of surprise per se, it recognizes that there are certain legal precedents that radically alter the legal landscape. Rissland calls such precedents 'black swans,' which are rare, perhaps only differing from past legal cases in "small" ways, but they are surprising nonetheless. Importantly, as cases that are similar to the black swan surface, these 'grey cygnets' (as she calls them) are covered by the earlier black swan precedent; a grey cygnet is not transformative and not surprising. The general lesson for approaches to assessing surprise is that rarity may not be enough, because over any sufficient time span the recurrence of rare events is quite likely! But of course, an observer's memory may be limited to a horizon, so that when time intervals are bounded by these horizons, rarity may in fact be a sufficient basis for assessing surprise.

Itti and Baldi (2004) describe a model of surprising features in image data using a priori and posterior probabilities. Given a user dependent model M of some data, there is a P(M) describing the probability distribution. P(M|D) is the probability distribution conditioned on data. Surprise is modeled as the distance d between the prior, P(M), and posterior P(M|D) probabilities. In this model, time is not an explicit attribute or dimension of the data. There are only two times: before and now.

Ranasinghe and Shen (2008) develop a model of surprise as integral to developmental robots. In this model, surprise is used to set goals for learning in an unknown environment. The world is modeled as a set of rules, where each rule has the form: Condition → Action → Predictions. A condition is modeled as: Feature → Operator → Value. For example, a condition can be feature1 > value1 where greater than is the operator. A prediction is modeled as: Feature → Operator. For example, a prediction can be feature1 > where it is expected that feature1 will increase after the action is performed. Comparisons can detect the presence or absence of a feature, and the change in the size of a feature (<, ≤, =, ≥, >). If an observed feature does not match its predicted value, then the system recognizes surprise. This model does not make any explicit reference to time and uses surprise as a flag to update the rule base.

Maher and Fisher (2012) have used clustering algorithms to compare a new design to existing designs, to identify when a design is novel, valuable, and surprising. The clustering model uses distance (e.g., Euclidean distance) to assess novelty and value of product designs (e.g., laptops) that are represented by vectors of attributes (e.g., display area, amount of memory, cpu speed). In this approach, a design is considered surprising when it is so different from existing designs that it forms its own new cluster. This typically happens when the new design makes explicit an attribute that was not previously explicit, because all previous designs had the same value for that attribute. Maher and Fisher use the example of the Bloom laptop, which has a detachable keyboard (i.e., detachable keyboard = TRUE), where all previous laptop designs had value FALSE along what was a previously implicit, unrecognized attribute. Thus, like one of Rissland's black swans, the Bloom transformed the design space.

In Maher and Fisher, the established clusters of design are effectively representing the expectation that the next new design will be associated with one of the clusters of existing designs, and when a new design forms its own cluster it is surprising and changes our expectations for the next generation of new designs.

Maher and Fisher (2012) focused on evaluation of creativity on the part of an observer, not an active designer. Brown (2012) investigates many aspects of surprise in creative design, such as who gets surprised: the designer or the person experiencing or evaluating the design. Brown (2012) also presents a framework for understanding surprise in creative design by characterizing different types of expectations, active, active knowledge, and not active knowledge, as alternative situations in which expectations can be violated in exploratory and transformative design.

To varying extents, many of the computational approaches above model surprise as a deviation from expectation, where the expectation is an expected value that is estimated from data distributions or a prediction made by simulating a rule-based model. In these, however, there is no explicit representation of time as a continuum, nor explicit concern with projecting into the future.

Recognizing Surprising Designs

Our approach to projecting designs into the future assumes that each product design is represented by a vector of ordinal attributes (aka variables). For each attribute, a mathematical function of time can be fit to the attribute values of existing (past) designs, showing how the attribute's values have varied with time in the past. This best fitting function, obtained through a process of regression, can be used to predict how the attribute's values will change in the future as well. Our approach to projecting into the future is inspired by earlier work by Frey and Fisher (1999) that was concerned with projecting machine learning performance curves into the future (thereby allowing cost benefit analyses of collecting more data for purposes of improving prediction accuracy), and it was not concerned with creativity and surprise assessment per se. While Frey and Fish-

er used a variety of functional forms, most notably power functions, as well as linear, logarithmic, and exponential, we have thus far only used linear functions (i.e., univariate linear regression) for projecting designs into the future for purposes of surprise assessment.

In this paper we focus on regression models for recognizing a surprising design: a regression analysis of the attributes of existing designs against a temporal dimension is used to predict the ”next” value of the attributes. The distance from the observed value to the predicted value identifies a surprising attribute-value pair.

We illustrate our use of regression models for identifying surprising designs in an automobile design dataset, which is composed of 572 cars that were produced between 1878 and 2009 (Dowlen, 2012). Each car is described by manufacturer, model, type, year, and nine numerically-valued attributes related to the mechanical design of the car. In this dataset only 190 entries contain values for all nine attributes. These complete entries all occur after 1934 and are concentrated between 1966 and 1994. A summary of the number of designs and the number of attributes in our dataset is shown in Table 1.

Table 1: List of the mechanical design attributes and the number of automobile design records with an entry for each of the nine attributes in our dataset.

Attribute	Number of Designs
Engine Displacement	438
Bore Diameter	407
Stroke Length	407
Torque Force	236
Torque Displacement	235
Weight	356
Frontal Area	337
Maximum Speed	345
Acceleration	290

A variety of linear regression models are considered. The first model uses linear regression over the entire time period of the design data and fits a line to each attribute as a function of time. The results for one attribute, maximum speed, are shown in Figure 1. This analysis identifies the outliers, and therefore potentially surprising designs. For example, the Ferrari 250LM had a surprising maximum speed in 1964, and the Bugati Type 41 Royale has a surprising engine size (another attribute, and another regression analysis) in 1995.

This first model works well for identifying outliers across a time period but does not identify trendsetters (or ‘black swans’ as Rissland might call them) since data points that occurred later in the timeline were included in the regression analysis when evaluating the surprise of a design. A trendsetter is a surprising design that changes the expectations for designs in the future, and is not simply an outlier for all time. In other words, using the entire time line to identify surprising automobile designs does not help us identify those designs that influenced future designers. A design that is an outlier in its own time, but inspires future generations of designers to do something similar can only be found if we don’t use designs which came out after the model being measured in the training data.

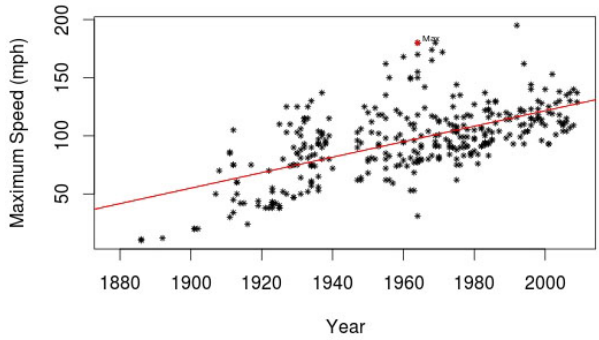

Figure 1. Regression analysis for maximum speed over the entire time period of car design data.

Thus, we considered a second strategy that performs a linear regression only on previously created designs and measures surprise of a new design as the distance from that design’s attribute value to the projection of the line at the year of the design in question. This second regression strategy, where the time period used to fit the line for a single attribute was limited to the time before each design was released (see Figure 2), found roughly the same surprising designs as the first model (over the entire time period) for most attributes, but there were two exceptions: torque displacement and maximum speed. In these exceptions, outliers earlier in time were sufficiently extreme so as to significantly move the entire regression line from before the early outliers to after, whereas in other cases the rough form of the regression lines created over time did not change much.

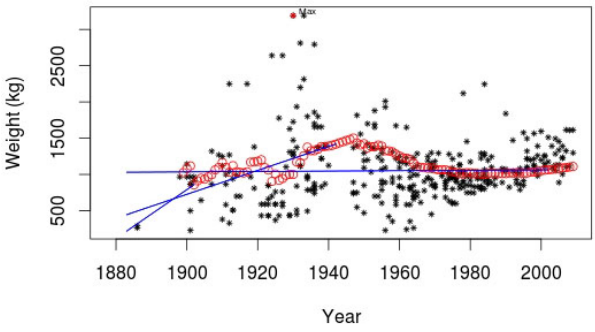

Figure 2: Using strategy 2, linear models are constructed using all previous-year designs. The circles show the predicted (or projected) values for EACH year from the individual regression lines; the dots show actual values. We show three sample regression lines, each ending at the year (circle) it is intended to predict, but there is actually one regression line for each year.

When training this second model, designs from every previous year were weighted equally for predicting future designs. Thus, outliers in the beginning of the dataset perpetually shifted the model and skewed the surprise measurements for all subsequent designs. And why shouldn't they – these early designs correspond roughly to what Rissland called black swans, which understandably diminish the surprise value of subsequent 'grey cygnets'. However, it is also the case that when using model 2, taking into account all past history, that a large mass of 'bland' designs earlier can exaggerate the perceived surprise of a design, even when that design is in the midst of a spurt of like designs.

These observations inspired a third linear regression strategy that makes predictions (or sets expectations) by only including designs within a specified time range before the designs being measured. We use a sliding window, rather than disjoint bins. In either case though, limited time intervals can mimic perceptions of surprise when the observer has a limited memory, only remembering up to a myopic horizon into the past.

The window (aka interval) size used for the cars dataset was ten years. This number was chosen because histograms of the data revealed that all ten-year periods after 1934 contained at least one design with all nine attributes while smaller periods were very sparsely populated in the 1950s. Larger window sizes converged to the second regression model as window size increased.

In general, the size of windows has a large influence on the results. Though we won't delve into the results of this final strategy here, its sensitivity has appeal. In fact, relative to our longer-term goal of modeling human surprise, this sensitivity to window size may map nicely on to different perceptions by people with different experiences. An older adult may have a very different surprise reaction than a young person, depending on past experience. In general, the selection of an appropriate range of years for the third regression model can be correlated with typical periods of time over which a person can remember. That is, if we want to compare our computational model of surprise with human expectations, we should use time intervals that are meaningful to people rather than based on the distribution of data. People will be surprised when expectations based on a time period relevant to their personal knowledge and experience of a series of designs is not met, rather than on the entire time period for all designs.

Directions for Further Research

This paper presents an approach to evaluating whether a design is surprising, and therefore creative, by including a temporal analysis of the conceptual space of existing designs and using regression analysis projected into the future to identify surprising designs. There are a number of directions we plan to follow.

1. We want to further develop the regression models, and in particular move beyond linear regression, to include other functional forms such as polynomial, power, and logarithmic. After all, a design might be regarded as surprising if we used linear regression to project into the future, but not at all surprising if we used a higher-order polynomial regression into the future! Identifying means of distinguishing when one functional form over another is most appropriate for regression will be a key challenge.

2. We want to move beyond our current univariate assessments of surprise through univariate regression, to holistic, multivariate model assessments of surprise through multivariate regression. We can apply multivariate regression methods to designs as a function of time, or combine our earlier work on clustering approaches (Maher and Fisher, 2012) with our regression approaches, perhaps by performing multivariate regression over multivariate summaries of design clusters (e.g., centroids).

3. We have thus far been investigating novelty and value (Maher and Fisher, 2012) and surprise as decoupled characteristics of creativity, but an important next step is to consider how measures of these three characteristics can be integrated into a single holistic measure of creativity, probably parameterized to account for individual differences among observers.

4. Assessments of creativity are conditioned on individual experiences; such individual differences in measures of surprise, novelty, and value are critical – surprise to one person is hardly so to another. We made a barest beginning of this study in Maher and Fisher (2012), where we viewed clustering as the means by which an agent organized its knowledge base, and against which creativity would be judged. The methods for regression that we have presented in this paper will allow us to build in an "imagining" capacity to an agent, adding expectations for designs that do not yet exist to the knowledge base of agents responsible for assessing creativity.

5. In all the variants that we plan to explore, we want to match the results of our models in identifying surprising designs to human judgments of surprise, and of course to assessments of creativity (novelty, value, surprise) of the designs, generally.

6. Finally, our work to date assumes that designs are represented as attribute-value vectors; these propositional representations are clustered in Maher and Fisher (2012), or time-based regression is used in this paper. We want to move to relational models, however, perhaps first-order representations and richer representations still. Relational representations would likely be required in Rissland's legal domain, if in fact that domain were formalized.

A domain that we find very attractive for exploring relational representations is the domain of computer programs, which follow a formal representation and for which a number of well established tools exist for evaluating novelty, value, and surprise. For example, consider that tools for identifying plagiarism in computer programs measure "deep" similarity between programs, and can be adapted as novelty detectors), and for assessing surprise as well.

An ability to measure creativity of "generic" computer programs will allow us to move into virtually any (computable) domain that we want. For example, consider mathematical reasoning in students. In an elementary

course, we can imagine seeing a large number of programs that are designed to compute the variance of data values, as composed of two sequential loops – the first to compute the mean of the data, and the subsequent loop to compute the variance given the mean. These programs will be very similar at a deep level. Imagine then seeing a program that computes the variance (and mean) with ONE loop, relying on a mathematical "simplification." These are the kinds of assessments of creativity that we can expect in more sophisticated relational domains, all enabled by capabilities to assess computer programs.

Acknowledgements: We thank our anonymous reviewers for helpful comments, which guided our revision.

References

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

Amabile, T. 1996. Creativity in Context: Update to "The Social Psychology of Creativity". Boulder, CO: Westview Press.

Besemer, S., and O'Quin, K. 1987. Creative product analysis: Testing a model by developing a judging instrument. Frontiers of creativity research: Beyond the basics 367–389.

Besemer, S. P., and O'Quin, K. 1999. Confirming the three-factor creative product analysis matrix model in an American sample. Creativity Research Journal 12:287–296.

Boden, M. 2003. The Creative Mind: Myths and Mechanisms, 2nd edition. Routledge.

Brown, D. C. 2012. Creativity, surprise and design: An introduction and investigation. In The 2nd International Conference on Design Creativity (ICDC2012), 75–84.

Cropley, D. H., and Cropley, A. J. 2005. Engineering creativity: A systems concept of functional creativity. In Creativity Across Domains: Faces of the muse, 169–185. Hillsdale, NJ: Lawrence Erlbaum.

Cropley, D. H.; Kaufman, J. C.; and Cropley, A. J. 1991. The assessment of creative products in programs for gifted and talented students. Gifted Child Quarterly 35:128–134.

Cropley, D. H.; Kaufman, J. C.; and Cropley, A. J. 2011. Measuring creativity for innovation management. Journal Of Technology Management & Innovation

Csikszentmihalyi, M., and Wolfe, R. 2000. New conceptions and research approaches to creativity: Implications of a systems perspective for creativity in education. International handbook of giftedness and talent 2:81–91.

Dowlen, C. 2012. Creativity in Car Design – The Behavior At The Edges. A. Duffy, Y. Nagai, T. Taura (eds) Proceedings of the 2nd International Conference on Design Creativity (ICDC2012), 253-262.

Forster, E., and Dunbar, K. 2009. Creativity evaluation through latent semantic analysis. In Proceedings of the Annual Conference of the Cognitive Science Society, 602–607.

Frey, L., and Fisher, D. 1999. Modeling decision tree performance with the power law. In Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics, 59–65. Ft. Lauderdale, FL: Morgan Kaufmann.

Goldenberg, J., and Mazursky, D. 2002. Creativity In Product Innovation. Cambridge University Press.

Horn, D., and Salvendy, G. 2003. Consumer-based assessment of product creativity: A review and reappraisal. Human Factors and Ergonomics in Manufacturing & Service Industries 16:155–175.

Horvitz, E.; Apacible, J.; Sarin, R.; and Liao, L. 2005. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In Proceedings of the 2005 Conference on Uncertainty and Artificial Intelligence. AUAI Press.

Itti L. and Baldi P. (2004). A Surprising Theory of Attention, *IEEE Workshop on Applied Imagery and Pattern Recognition.*

Maher, M. L., and Fisher, D. 2012. Using AI to evaluate creative designs. In A. Duffy, Y. Nagai, T. Taura (eds) Proceedings of the 2nd International Conference on Design Creativity (ICDC2012), 45-54.

Oman, S., and Tumer, I. 2009. The potential of creativity metrics for mechanical engineering concept design. In Bergendahl, M. N.; Grimheden, M.; Leifer, L.; P., S.; and U., L., eds., Proceedings of the 17th International Conference on Engineering Design (ICED'09), Vol. 2, 145–156.

Ranasinghe, N., and Shen, W.-M. 2004. A surprising theory of attention. In IEEE Workshop on Applied Imagery and Pattern Recognition.

Ranasinghe, N., and Shen, W.-M. 2008. Surprise-based learning for developmental robotics. In Proceedings of the 2008 ECSIS Symposium on Learning and Adaptive Behaviors for Robotic Systems.

Rissland, E. (2009). Black Swans, Gray Cygnets and Other Rare Birds. In L. McGinty and D.C. Wilson (Eds.): ICCBR 2009, LNAI 5650, pp. 6–13, 2009. Springer-Verlag Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007%2F978-3-642-02998-1_2?LI=true#page-1

Runco, M. A. 2007. Creativity: Theories and Themes: Research, Development and Practice. Amsterdam: Elsevier. Shah, J.; Smith, S.; and Vargas-Hernandez, N. 2003. Metrics for measuring ideation effectiveness. Design Studies

24:111–134.

Wiggins, G. 2006. A preliminary framework for description, analysis and comparison of creative systems. Knowledge-Based Systems 16:449–458.