A Reductio Ad Absurdum Experiment in Sufficiency for Evaluating (Computational) Creative Systems

Dan Ventura

Computer Science Department
Brigham Young University
Provo UT 84602, USA
ventura@cs.byu.edu

Abstract. We consider a combination of two recent proposals for characterizing computational creativity and explore the sufficiency of the resultant framework. We do this in the form of a *gedanken* experiment designed to expose the nature of the framework, what it has to say about computational creativity, how it might be improved and what questions this raises.

Key words: Computational Creativity, Evaluation, Metrics, Sufficiency

1 Introduction

A great deal has been written about the nature of creativity in a computational setting and how we might characterize it or measure it or detect it or justify it. Boden is generally credited with beginning this discussion [1, 2] and many others have contributed ([3–7], among others). Recently, Ritchie has refined an empirical framework for the evaluation of creativity in computational systems [8], and Colton has responded with an interesting augmentation to this proposal [9]. In particular, Ritchie has proposed 18 set-theoretic criteria by which various aspects of a potentially creative artefact might be evaluated; Colton has suggested that, additionally, three qualities must be perceived in the process that produces the artefact in question. Taken together these form an intriguing framework for characterizing (computational) creativity that has been applied to analyze computational systems for creating visual art [9], mathematical proofs [9], music [10], poetry [11] and language constructs [12].

The purpose of this paper is to explore these two proposals together, particularly from a sufficiency standpoint (Colton argues for the necessity of his qualities but makes no claims about sufficiency and Ritchie purposely avoids a discussion of either, presenting his framework as an analytical tool). In fact, we will put forth a *gedanken* experiment designed to stress the combined framework. Let us emphasize that our intention in trying to break the framework is not in any way an attempt to discredit the ideas, but rather just the opposite; we feel that the combined framework provides many excellent insights into the nature of

computational creativity. Our goal here, in trying to find counter arguments, is to explore just what we can learn from characterizing (computational) creativity as Ritchie and Colton suggest and how, if at all, their proposals might be further strengthened.

2 A *Gedanken* Experiment

Let us consider a simple task that could require creativity—producing black and white digital images of recognizable scenes. Following Ritchie's framework, the set $\mathcal{B}$ of *basic items*[1] for this task might be, for example, 450×350 pixel images with pixels taking only binary values; the *artefact class*, is a 2-tuple $(\mathcal{B}, r)$, where $r : \mathcal{B} \rightarrow [0, 1]$ is a mapping or *rating scheme* that measures the quality of recognizability—those images that are recognizable as scenes have high values of r while those that are not recognizable have low r values.[2] Note that although Ritchie defines the concept of artefact class, he immediately refines this idea to define a *value-based artefact class* as a 3-tuple $(\mathcal{B}, typ, val)$, where $typ : \mathcal{B} \rightarrow [0, 1]$ and $val : \mathcal{B} \rightarrow [0, 1]$ are mappings that capture the *typicality* and *quality* associated with a member of $\mathcal{B}$. In our *gedanken* experiment, since we are interested simply in (artefact) class membership (how recognizable an image is), the two rating schemes of the 3-tuple become redundant (that is, typicality and quality are equivalent for this task), and we can think of the 2-tuple definition as capturing this by compressing both into the single rating scheme r. Also, to begin with, we will consider the special case of no *inspiring set* I.[3] Finally, the set R of results produced by the system will be the image(s) returned by the system during one or more runs.

2.1 The RASTER System

We propose a very simple system, named, rather whimsically, *Ridiculous Artist Supporting Thought Experiment Realization* (RASTER). The RASTER system creates black and white images by employing a combination of undirected search and objective function (see Algorithm 1). The system randomly generates a binary image and computes a distance from that image to a randomly selected picture from the web that has been binarized by thresholding.[4] We emphasize

[1] Ritchie defines *basic items* as, essentially, the data type of the artefacts to be produced by a system.

[2] Though we do not formalize this notion of recognizability, we suggest that it captures an intrinsic quality of an image that is not too difficult to measure (e.g. by voting—does this look like something?)

[3] Ritchie does not formalize this notion but rather simply describes it as "the subset of basic items that influences (either explicitly or implicitly) the construction of the program (e.g. this could be all the relevant artefacts known to the program designer, or items which the program is designed to replicate, or a knowledge base of known examples which drives the computation within the program".

[4] One could assume an inner loop that, for a given randomly generated image p, iteratively compares p to each picture on the web. However, for simplicity, and to

Algorithm 1 The (hypothetical) RASTER algorithm. `binary_threshold`() is a function that converts an image to binary representation by setting all pixel values above a threshold to 1 and all others to 0. `difference`() computes a normalized distance (e.g. Hamming) between its arguments. The fitness threshold θ_f determines the "recognizability" of the generated pattern p.

```
set δ = ∞
while δ > θ_f do
  generate a random raster pattern p
  q ← random picture from the web
  binary_threshold(q)
  δ = difference(p, q)
end while
return(p)
```

that RASTER has no inspiring set. One might perhaps argue that the set of all pictures on the web should be considered an inspiring set for the system, but we would reply that this argument is only valid if that set is used to guide RASTER's search (which it is not).[5]

Now, given enough time, it is not unlikely that the RASTER system could produce an image like that of Figure 1. Indeed, if the threshold θ_f is small enough, images like this are the only kind of output RASTER would produce. The question is, given a set R of images produced by RASTER, how would the system be characterized by the 18 criteria and three qualities?

2.2 Assessing RASTER's Creativity

We will use the 18 criteria to assess RASTER's eligibility to be considered creative. Employing Ritchie's original notation, $T_{\alpha,\beta}(X) \equiv \{x \in X | \alpha \leq typ(x) \leq \beta\}$ formalizes the subset of X in a given range of typicality, $V_{\alpha,\beta}(X) \equiv \{x \in X | \alpha \leq val(x) \leq \beta\}$, the subset of X in a given range of quality (typ and val are the rating schemes over $\mathcal{B}$ defined above), $AV(F, X) \equiv \frac{\sum_{x \in X} F(x)}{|X|}$, the average value of a function F across a finite set X, and $ratio(X, Y) \equiv \frac{|X|}{|Y|}$, the relative sizes of two finite sets X, Y. We consider the asymptotic value of each criterion as RASTER's fitness threshold $\theta_f \to 0$:

1. ~~$AV(typ, R) = AV(val, R) = 1$~~

emphasize the complete lack of an inspiring set I, we have chosen this much less efficient approach.

[5] Actually, it is difficult to make this argument formally as Ritchie's original description of the inspiring set is very informal. However, we believe that even the most conservative interpretation would admit that RASTER has an inspiring set of cardinality no greater than one (and in this case, the set can not be static and thus the argument for an inspiring set of cardinality greater than 0 is weak). Even conceding an inspiring set of cardinality one, our assessment in Section 2.2 of RASTER's creativity per the criteria remains unchanged.

2. ~~$ratio(T_{\alpha,1}(R), R) = ratio(V_{\gamma,1}(R), R) = 1$~~
3. $AV(val, R) = 1$
4. $ratio(V_{\gamma,1}(R), R) = 1$
5. ~~$ratio(V_{\gamma,1}(R) \cap T_{\alpha,1}(R), T_{\alpha,1}(R)) = ratio(T_{\alpha,1}(R), T_{\alpha,1}(R)) = 1$~~
6. ~~$ratio(V_{\gamma,1}(R) \cap T_{0,\beta}(R), R) = ratio(\emptyset, R) = 0$~~
7. ~~$ratio(V_{\gamma,1}(R) \cap T_{0,\beta}(R), T_{0,\beta}(R)) = ratio(\emptyset, T_{0,\beta}(R)) = 0$~~
8. ~~$ratio(V_{\gamma,1}(R) \cap T_{0,\beta}(R), V_{\gamma,1}(R) \cap T_{\alpha,1}(R)) = ratio(\emptyset, V_{\gamma,1}(R)) = 0$~~
9. ~~$ratio(I \cap R, I) = undefined$~~
10. $(1 - ratio(I \cap R, R)) = 1$
11. ~~$AV(typ, R - I) = AV(typ, R) = AV(val, R) = 1$~~
12. ~~$AV(val, R - I) = AV(val, R) = 1$~~
13. ~~$ratio(T_{\alpha,1}(R - I), R) = ratio(T_{\alpha,1}(R), R) = ratio(V_{\gamma,1}(R), R) = 1$~~
14. ~~$ratio(V_{\gamma,1}(R - I), R) = ratio(V_{\gamma,1}(R), R) = 1$~~
15. ~~$ratio(T_{\alpha,1}(R - I), R - I) = ratio(T_{\alpha,1}(R), R) = ratio(V_{\gamma,1}(R), R) = 1$~~
16. ~~$ratio(V_{\gamma,1}(R - I), R - I) = ratio(V_{\gamma,1}(R), R) = 1$~~
17. ~~$ratio(V_{\gamma,1}(R - I) \cap T_{\alpha,1}(R - I), R - I) = ratio(V_{\gamma,1}(R), R) = 1$~~
18. ~~$ratio(V_{\gamma,1}(R - I) \cap T_{0,\beta}(R - I), R - I) = ratio(\emptyset, R) = 0$~~

Note that all these equalities hold independent of the choice of α, β, γ.[6]

Also, note that because of our use of a single rating scheme, criterion 1 reduces to criterion 3, criterion 2 reduces to criterion 4, criterion 5 vacuously evaluates to 1, and criteria 6-8 all vacuously evaluate to 0. And, because of our decision to have $I = \emptyset$, criterion 9 is undefined, criteria 11 and 12 reduce to criterion 3, criteria 13-17 reduce to criterion 4 and criterion 18 reduces to criterion 6, leaving just three applicable criteria (those not struck out in the list). Ritchie actually proposes his criteria as predicates, parametrized by a threshold $0 < \theta < 1$, and it should be obvious that in the limit, RASTER would be characterized creative by all three of the applicable criteria, regardless of the choice of θ.

[6] Actually, this is not quite true if one insists on maintaining the three different thresholds α, β, γ while using a single rating scheme; however, we consider the statement functionally true, dismissing this odd pathological case.

Fig. 1. Hypothetical figure created by the RASTER program

To justify our (asymptotic) values for criteria 3, 4 and 10, consider that as $\theta_f \to 0$, $\delta \to 0$ for any p returned as an artefact. This in turn implies that $\texttt{difference}(p, q) \to 0$ and therefore that in the limit $p = \texttt{binary\_threshold}(q)$, a recognizable[7] binary image. Therefore, since this argument holds for each p produced, and since the set R is composed of p returned by RASTER, in the limit, the ratio of highly rated (recognizable) artefacts to total artefacts, $ratio(V_{\gamma,1}(R), R)$, will approach unity (criterion 4). *A fortiori*, the average rating (recognizability) of R will also approach unity (criterion 3). And, since by design $I = \emptyset$, $I \cap R = \emptyset$, $ratio(I \cap R, R) = 0$, and therefore $1 - ratio(I \cap R, R) = 1$ (criterion 10).

To address Colton's qualities, we argue that the system possesses *imagination* due to its undirected random search of the set $\mathcal{B}$, and that the system possesses *appreciation* by nature of its fitness measure that is applied to the result of said search. Finally, we argue that the system possesses *skill* because it (only) produces black and white images (with a high fitness value).

2.3 Imbuing RASTER with an Inspiring Set

Now, there are two obvious objections to this thought experiment: the majority of Ritchie's criteria have been rendered irrelevant, and an inordinate amount of time is likely required for RASTER to produce even a single artefact.

As to the first objection, on the one hand this invalidation of 15 of the 18 criteria is allowable under Ritchie's original framework, and so the criticism can not be fairly leveled at our thought experiment. On the other hand, this is an interesting observation and is also perhaps a valid argument for why we might dismiss RASTER as not creative. Indeed, we might consider this indicative of a meta-heuristical application of the 18 criteria—if most of the criteria *can not* be applied, then the system in question may not be a likely candidate for attribution of creativity. The basis for the second objection is directly attributable to the structure of the RASTER system itself, and perhaps suggests the need for an addition to Ritchie's criteria or Colton's qualities in the form of some measure of *efficiency*.

Indeed, one might consider imposing some sort of temporal limit by which time the system must have produced an artefact to be considered as possibly creative. Under such a condition, the RASTER system is likely to be disqualified; but can we produce an acceptable substitute that is not? The *inspired* RASTER (iRASTER) system may be such an alternative (see Algorithm 2). iRASTER differs from RASTER by having an inspiring set I of exemplar binary images and by its method for generating candidate images, which is done by applying suitably defined genetic operators to members of I.

Depending on what kind of temporal constraint is imposed, it is perhaps not clear that iRASTER will satisfy the constraint; however, it should be clear that iRASTER will produce artefacts similar to (at least some of) those produced by

[7] We are assuming the existence of a binarization threshold that produces a recognizable image—perhaps something like the mean pixel value for the image.

Algorithm 2 The (hypothetical) iRASTER algorithm, a modification of the RASTER algorithm that incorporates an inspiring set I.

Input: a set I of exemplar binary images
set $\delta = \infty$
while $\delta > \theta_f$ **do**
  generate a raster pattern p by applying genetic operators to members of I
  $q \leftarrow$ random picture from the web
  `binary_threshold`(q)
  $\delta =$ `difference`(p, q)
end while
return(p)

RASTER in a much shorter period of time. And, it is certainly true that any imposed temporal constraint can not be overly strict without disqualifying many viable candidates for creative attribution as, in general, creative productivity contains a significant stochastic element.

Now, for iRASTER we can revisit the criteria scoring, first noting that criteria $1 - 8$ do not depend upon I and so will have the same values in the limit for iRASTER as for RASTER. For the remainder, we again consider the asymptotic value of each criterion as iRASTER's fitness threshold $\theta_f \to 0$:

9. $ratio(I \cap R, I) = 0$ (although, $ratio(I \cap R, I) = 1$ as $|R| \to \infty$)
10. $(1 - ratio(I \cap R, R)) = 1$
11. ~~$AV(typ, R - I) = AV(val, R - I) = 1$~~
12. $AV(val, R - I) = 1$
13. ~~$ratio(T_{\alpha,1}(R - I), R) = ratio(V_{\gamma,1}(R - I), R) = 1$~~
14. $ratio(V_{\gamma,1}(R - I), R) = 1$
15. ~~$ratio(T_{\alpha,1}(R - I), R - I) = ratio(V_{\gamma,1}(R - I), R - I) = 1$~~
16. $ratio(V_{\gamma,1}(R - I), R - I) = 1$
17. ~~$ratio(V_{\gamma,1}(R - I) \cap T_{\alpha,1}(R - I), R - I) = ratio(V_{\gamma,1}(R - I), R - I) = 1$~~
18. ~~$ratio(V_{\gamma,1}(R - I) \cap T_{0,\beta}(R - I), R - I) = ratio(\emptyset, R - I) = 0$~~

Again, all these equalities hold independent of the choice of α, β, γ.

With $I \neq \emptyset$, criterion 9 is no longer undefined, criterion 12 no longer reduces to 3 (though criterion 11 does reduce to criterion 12), criteria 14 and 16 no longer reduce to criterion 4 (though criterion 13 reduces to 14 and criteria 15 and 17 reduce to criterion 16) and criterion 18 no longer reduces to criterion 6 but still vacuously evaluates to 0 for the same reason. So the count of applicable criteria increases to 7. And considering the predicate case, by the same argument as before, in the limit, iRASTER would be characterized creative by 6 of 7 applicable criteria (criteria 3, 4, 10, 12, 14 and 16, but not criterion 9), regardless of the choice of θ.[8]

[8] Actually, an argument can be made for the seventh as well, if we allow the set R to grow very large, but since our motivation for this second system is to reduce time required to produce R, it is appropriate to disqualify this criterion.

To justify our (asymptotic) values for criteria 10, 12, 14 and 16 (3 and 4 still hold as argued above), consider the following. A probabilistic argument suggests that it is very unlikely that $p \in I$ and since we are assuming a small, finite R, it is therefore very likely that $|I \cap R| \approx 0$, and it follows that $1 - ratio(I \cap R, R) = 1$ (criterion 10). As before, as $\theta_f \to 0$, $\delta \to 0$ for any p returned as an artefact. This implies that $\texttt{difference}(p, q) \to 0$ and therefore that in the limit $p = \texttt{binary\_threshold}(q)$. Therefore, since this argument holds for each p produced, and since the set R is composed of p returned by RASTER, and since $(|I \cap R| \approx 0) \Rightarrow ((R - I) \approx R)$, in the limit, the ratio of highly rated (recognizable) novel artefacts to total artefacts, $ratio(V_{\gamma,1}(R - I), R)$, will approach unity (criterion 14). And therefore, *a fortiori*, the average rating (recognizability) of $R - I$ will also approach unity (criterion 12) and the ratio of highly rated (recognizable) novel artefacts to total novel artefacts, $ratio(V_{\gamma,1}(R - I), R - I)$, will also approach unity (criterion 16).

Revisiting Colton's qualities, we argue that the new system possesses imagination due to the stochastic nature of its search of the set $\mathcal{B}$, and that the new system still possesses appreciation by nature of its fitness measure and that the system still possesses skill because it still (only) produces black and white images (with a high fitness value).

3 Analysis and Discussion

The RASTER system is by design nearly equivalent to the proverbial room full of monkeys pounding on typewriters. The iRASTER system is an attempt to maintain this absurdity while introducing additional "knowledge" to improve "time to artefact production". Nevertheless, according to the framework, the RASTER system can be attributed with three characteristics of creativity: the average rating of artefacts is high, the variability of the rating is low (that is, artefacts are uniformly rated high), and the artefacts produced are novel (they do more than replicate an inspiring set).[9] Further, we have argued that the system has skill, imagination and appreciation. The iRASTER system is attributed with three additional characteristics of creativity: a large proportion of the artefacts are novel, the average rating of novel artefacts is high and there is little variability in the ratings of novel artefacts (a large proportion of them are rated highly), and again we argue that it possesses skill, imagination and appreciation.

The salient question is should these two systems be considered creative? In the case of RASTER, we assert that the answer is a definite no and have identified two possible improvements to the framework that would justify this: a notion of variety (of creative attributes) and a notion of efficiency. The first of these can be assessed with a meta-criterion defined as $ratio(P(\mathcal{C}, R), \mathcal{C}) > \theta$, where $\mathcal{C}$ is a set of (normalized) criteria for characterizing creativity, $P(X, Y) = \sum_{x \in X} pred(x, Y)$, $pred(x, Y) = 1$ if criterion $x = \texttt{TRUE}$ when evaluated over the set Y and

[9] Though in our analysis we have considered the asymptotic case $\theta_f \to 0$, in "practice" it should always be true that $\theta_f > 0$, and it is possible to employ a second threshold θ_g such that $1 > \theta_f \geq \delta \geq \theta_g > 0$ to emphasize artefact novelty.

$pred(x, Y) = 0$ otherwise. (Note that depending upon how one interprets this meta-criterion, RASTER may *still* be considered creative. Because our problem definition admits no inspiring set and makes use of a single rating scheme, the majority of Ritchie's criteria become redundant. As mentioned before, Ritchie's framework neither disallows nor penalizes this scenario, so, in a very real sense the set $|\mathcal{C}| = 3$ and $ratio(P(\mathcal{C}, R), \mathcal{C}) = 1$; that is, RASTER is creative in as many ways as we have to measure it. A similar argument for iRASTER gives $ratio(P(\mathcal{C}, R), \mathcal{C}) = 6/7$.)

For the second idea, one might measure time to artefact completion or one might, instead, measure some function of the number of intermediate steps or comparisons or rough drafts or discarded versions. Looking at the problem in this way suggests a set-theoretic formulation that complements the existing framework. For example, we can define a new set D of discarded or incomplete artefacts generated during the artefact production process, and we can suggest a 19th criterion such as $ratio(R, D \cup R) > \theta$.[10]

Interestingly, the simple addition of an inspiring set to the RASTER algorithm (iRASTER) has the effect of increasing both the variety of applicable creativity criteria and the efficiency of artefact production, and we are therefore somewhat less certain of our position on the question of iRASTER's creativity. However, it is still possible that it will fail to meet a reasonable standard for the efficiency requirement, and one could argue its disqualification on those grounds.

Also of interest is the fact that even with iRASTER increasing (over RASTER) the number of applicable criteria, the existence of a single rating scheme in the problem definition invalidates the majority of the criteria. Put another way, having a single rating scheme and no inspiring set implies that creativity per the Ritchie criteria is extremely limited at best (or, more critically, that the Ritchie criteria are not sufficient measures of creativity). Does this gibe with our current understanding of creativity? Or, perhaps, does it expose an unintentional (or intentional) bias in the set of criteria?

One might consider the possibility of quantifying Colton's qualities in a set-theoretic way to facilitate a cohesive framework, but this idea misses the point of Colton's thesis—it is not enough that a system possess skill, appreciation and imagination: to be considered creative, the system must be *perceived* to possess these traits. This suggests that (self-)promotion is an important aspect of creativity; indeed, it even suggests the possibility that inducement of perception could supplant actual possession of the qualities as the defining characteristic of creativity (one could certainly argue cases where this appears true for human artists, for example). Following this line of thinking further, we can ask whether any of Ritchie's criteria can be "lobbied" as suggested by Colton, and if so, is this positive or negative? This in turn leads us to ask whether any rating scheme that is subjective (human-evaluated) is not *ipso facto* subject to a No-Free-Lunch type argument that demonstrates that it will be uninformative on average. And,

[10] An anonymous reviewer suggests tying efficiency to value—the longer the time to completion, the greater the value must be for the artefact to be considered creative. This might be done with a criterion such as $ratio(V_{\alpha,1}(R), D) > \theta$.

if this is the case, is "perception engineering" the solution? It is certainly true that changes to *typ−* and *val−* type rating schemes do happen (e.g. posthumous increase in fame of an artist, acceptance of a new painting style over time, etc.), so why not admit [intentional] influences, originating either from an "agent" (like the system designer) or, perhaps more impressively, from the creative system itself? This finally leads to the question of whether self-promotion may be a sufficient (if not necessary) characteristic of creative entities.

References

1. Boden, M.: The Creative Mind. Abacus, London (1992)
2. Boden, M.: Creativity and artificial intelligence. Artificial Intelligence **103** (1998) 347–356
3. Colton, S., Pease, A., Ritchie, G.: The effect of input knowledge on creativity. In: Case-based Reasoning: Papers From the Workshop Programme at ICCBR '01. (2001)
4. Pease, A., Winterstein, D., Colton, S.: Evaluating machine creativity. In: Case-based Reasoning: Papers From the Workshop Programme at ICCBR '01, Technical Report SS-08-03 (2001) 129–137
5. Koza, J., Keane, M., Streeter, M., Mydlowec, W., Yu, J., Lanza, G.: Genetic Programming IV: Routine Human-competitive Machine Intelligence. Kluwer Academic Publisher/Springer (2003)
6. Wiggins, G.: Searching for computational creativity. In: Proceedings of the IJCAI-05 Workshop on Computational Creativity, Technical Report 5-05 (2006) 68–73
7. Wiggins, G.: A preliminary framework for description analysis and comparison of creative systems. Knowledge-Based Systems **19** (2006) 449–458
8. Ritchie, G.: Some empirical criteria for attributing creativity to a computer program. Minds and Machines **17** (2007) 76–99
9. Colton, S.: Creativity vs. the perception of creativity in computational systems. In: Creative Intelligent Systems: Papers from the AAAI Spring Symposium, Technical Report SS-08-03, AAAI Press (2008) 14–20
10. Haenen, J., Rauchas, S.: Investigating artificial creativity by generating melodies using connectionist knowledge representation. In: Proceedings of the 3rd Joint Workshop on Computational Creativity, ECAI. (2006) 33–38
11. Gervás, P.: Exploring quantitative evaluations of the creativity of automatic poets. In: Proceedings of the 2nd Workshop on Creative Systems, Approaches to Creativity in Artificial intelligence and Cognitive Science, ECAI, C. Bento and A. Cardoso and G. Wiggins (2002)
12. Pereira, F.C., Mendes, M., Gervás, P., Cardoso, A.: Experiments with assessment of creative systems: An application of Ritchies criteria. In: Proceedings of the Workshop on Computational Creativity, IJCAI, Technical Report 5-05 (2005) 37–44