Evaluating Machine Creativity

Alison Pease, Daniel Winterstein, Simon Colton
Division of Informatics, University of Edinburgh, 80 South Bridge,
Edinburgh, EH1 1HN, Scotland
{alisonp, danielw, simonco}dai.ed.ac.uk

Abstract

We consider aspects pertinent to evaluating creativity to be input, output and the process by which the output is achieved. These issues may be further divided, and we highlight associated justifications and controversies. Appropriate methods of measuring these aspects are suggested and discussed.

Introduction

In order to guide, measure and record progress in machine creativity a framework which outlines relevant aspects and methods by which they may be measured must be agreed. The challenge is to find a framework which is both practically useful and theoretically feasible, i.e. formal but not oversimplified. In (Ritchie, 2001) one such framework was proposed; we extend that by considering the three relevant criteria for attribution of creativity - input, output and the process by which it is achieved, and suggesting methods of measuring these criteria. Some of the criteria are necessary and some controversial. Which measure is appropriate depends on factors such as the domain (in particular whether it is arts or science), type of program (for example symbolic or subsymbolic), or the aim of the program (whether it is to understand human processes or to be more effective). Some of the measures will be calculated by the system, and others externally. We conclude by considering ways in which to attribute creativity to programs.

Assumptions

Although creativity is more easily determined (and possibly achieved) in some domains than others we assume that its essence is domain independent. Therefore our criteria (although not measurements) are general. Domain independent theories are always preferable, if they are possible, to domain specific theories, as they force us to look for a deeper structure and give more levels of understanding. If, however, it is not possible to give a domain independent analysis of creativity, then the attempt will nevertheless reveal much of the nature of creativity: which aspects are domain dependent, and why, thus guiding research. This is analogous to the development of AI in which the search to find a general essence of intelligence produced disparate branches of AI such as learning, perceiving and planning.

We aim to be true to intuitions about creativity which are largely based on but not restricted to human creativity (with evolution providing another example (Perkins, 1996)). This gives us a basis - albeit a very imprecise one - against which to judge definitions and theories. We assume that creativity in children, adults and geniuses is essentially the same phenomenon.

We base the act of creativity on a 2-stage model of *generation* and *evaluation*, referred to in (McGraw and Hofstadter, 1993) as the 'central loop of creativity'. This generalises the 4-stage model outlined in (Wallas, 1926); where generation is seen as *(i)* preparation and *(ii)* incubation, and evaluation as *(iii)* illumination and *(iv)* verification. We therefore assume that during the creative process some kind of *item* (x) is generated. In the term *item* we include artefacts such as poems, jokes or mathematical concepts, as well as less conventional 'objects' such as an interpretation, performance, aesthetic or analogy. For example in music, creativity may be achieved by the composer (who might produce a *piece of music* from various ideas and techniques), performer (who might produce a *series of sounds* from written notes, technical ability and other ideas or emotions) and listener (who might produce an *interpretation* or *aesthetic* from a series of sounds combined with her own experiences). Clearly items do not have to be material objects.

It might be thought that by including the act of producing an interpretation from perceived data as potentially creative, we risk losing the distinction between creative and non-creative acts, since we are constantly interpreting data. The distinction is upheld, however, not by excluding certain items from being *potentially* creative, but by imposing criteria such as those in the following sections, to help us to determine what is *actually* creative.

We assume that all creative items must be *novel* and *valuable*, and describe different interpretations of these words. In particular we use the distinction in (Boden, 1990) between p-creative (in which an item is new to the system which produced it) and h-creative (in which the item is historically novel).

Finally, we take it that we are aiming, through the study of machine creativity, to *(i)* further our understanding of creativity (human or other), and/or *(ii)* build programs which are useful in achieving practical goals.

Input

Creativity may be seen as output minus input. For example we may consider that a child has been creative in having an idea, and later retract our judgement if it transpires that her knowledge was more than we had assumed. The tendency in the human case is for the person attributing creativity to assume a similar degree of input (knowledge) in the creator as she herself has, unless it is clearly different (such as a much greater degree of expert knowledge). That this assumption is often unjustified shows that we are generally unqualified to make the creativity judgements that we do.

The Inspiring Set

Input to a program is more reliably identified than human input. *Explicit* knowledge, such as a database of example items, is easily determined. We should also take into account input that guided the programming. If a programmer designs her program with certain items in mind then these items are *implicit* knowledge. Ritchie calls the union of this *explicit* and *implicit* knowledge the *inspiring set*, I; defined to be the set of all items used to guide the programmer in the construction of her program(Ritchie, 2001).

If R is the set of items which have been generated by program P at some stage (in either a single or multiple runs); i.e. $\{x : x$ has been generated by $P\}$, then any item judged creative must lie in $R\backslash I$ (i.e. $R-I$). However since identifying those items used to guide a programmer ranges from difficult to impossible we extend this intuitive notion to strong and weak versions of I; I_S and I_W. I_S includes all items known to the programmer, i.e. any potentially creative item must be new to the programmer. Clearly any historically new items produced form a subset of $R\backslash I_S$. I_W is defined to be the set of all items that the programmer *knows* have influenced her program. While this set will be difficult to identify in retrospect, recording it during program construction should be feasible.

• Input Measure

Let I_S=$\{x$:programmer knows $x\}$, and I_W=$\{x$:x is *known* to have influenced the construction of $P\}$. Then any item x must be in $R\backslash I$ to be considered potentially creative. I_W could be fuzzy if the programmer recorded the *degree* of influence of x on the program construction.

Output

We take as a fundamental starting point that all creative items must be at least p-novel and valuable.

Novelty

For an item x to be considered novel it must be in $R\backslash I$. However this is not sufficient: for example suppose that x_1 and x_2 are books that differ by one word in the 7^{th} chapter, where $x_1 \in I$ and x_2 has been produced by the system. Then although $x_2 \in R\backslash I$, it is too similar to be considered novel. At the other extreme suppose that an item $x_3 \in R\backslash I$ was generated by the system, where x_3 was a book written in a representation which no human understands. This would be considered novel, but bizarre. We want to measure novelty in items in order to specify an intermediate range of variation which excludes both the boring and the bizarre.

Novelty relative to a body of knowledge

In (Boden, 1990) the idea of a concept space which represents a context is outlined. This is mapped (the boundaries, key landmarks etc. found), explored (using the map) and transformed (the mapped boundaries transcended). Novelty is explained according to an item's position in the space; in particular whether it is comfortably within a well explored area (not novel at all), in a little explored area but still within the boundaries ('merely' novel), outside but close to the boundaries ('fundamentally' novel) or outside and far from the boundaries (too chaotic to be considered novel). The degree of novelty with respect to the distance of an item from the nearest well explored area of a concept space is represented in Figure 1. This explanation is supported by

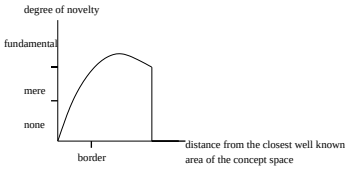

Figure 1: Boden's concept of novelty with respect to a concept space

evidence that both p-creativity and h-creativity are usually achieved by someone familiar with the domain who has changed it in some way. For example in (Bundy, 1994) it is pointed out that knowledge of the axioms of group theory may be learned in a few minutes, but it is not until one is familiar with their use and potential that any creative work may be carried out, such as extending them to ring theory. Another advantage is that it accounts for the element of surprise that usually comes with creative thoughts, since they could not have been had according to familiar rules. It also accords with our belief that creativity is partially determined by value since typically a transformation is of much greater value to a domain than exploration (as it opens up a whole new area or genre).

Boden's stress on transformation (fundamental novelty) over exploration (mere novelty) is controversial. Firstly, in (Bundy, 1994) and (Pind, 1994) it is claimed that exploration can result in more creative items than transformation. For example an unusual but legal chess move might be more creative than simply changing the rules of the game, i.e. transforming the *assumed* rather than *actual* boundary. This could be achieved by applying the suggestion in (Polya, 1962) to *consider the negative* to the heuristic *protect your queen* (which bounds the well explored area) rather than the legal queen's move (which bounds the actual concept space). Although novelty is not the sole criteria for creativity, this example shows that if the boundary to be transformed is the *actual* boundary then Boden's emphasis on transformation can be counterintuitive. However if the significant boundary is instead the *assumed* boundary then the exploration/transformation distinction is extremely useful. Boden does in fact allow that transforming *familiar* boundaries is sufficient for creativity; for example she says that "the surprise that we feel on encountering a creative idea

often springs from our recognition that it simply *could not* have arisen from the generative rules (implicit or explicit) *which we have in mind*." - p.41, (Boden, 1990).

Another controversy concerns the significance in terms of creativity of the difference between exploration and transformation. Bundy argues that transformation on the object level (breaking the rules which define the space) is equivalent to exploration on the meta level (the space of possible rules to break and ways to break them). In humans this meta level is often difficult to explore, since our assumptions which define the assumed object level space are rarely explicit (indeed the hardest part of the creative process is often identifying these boundaries). This leads to a third problem with Boden's theory; the condition that bodies of knowledge can be seen as concept spaces. In some fields, for example natural language, it is very difficult to see what rules might define the theoretical boundaries (rules of grammar are not always precise or universally agreed). Different genres often do not fit neatly into different concept spaces. For example are Impressionism and Expressionism two explored areas within the concept space of Post-Renaissance painting, or do they constitute distinct concept spaces? Boden suggests in (Boden, 1994) that the answer to this problem is to define a range of putative concept spaces and use them to evaluate the worth of her analogy.

• Transformation Measure

We define a *program* P to be the set of generating procedures; i.e. $P = \{k : k$ is a procedure for generating or evaluating an item, a procedure for generating or evaluating the processes used to generate or evaluate an item, a search strategy, a piece of input data, etc.$\}$. The *bag* of all pieces of knowledge in the program which contribute to the generation and evaluation of item x is: $H_x = [k \in P : k$ is used to generate $x \in R]$. (A bag includes repetitions, unlike a set.) Note that this bag is not necessarily unique as there may be different ways to generate the same item. We define a set of those object level procedures used to generate or evaluate an item x: $O_x = \{k \in P : k$ is directly used to generate or evaluate $x\}$, and a set of meta level procedures used to generate or evaluate a member of O_x: $M_x = \{k \in P : k$ is used to generate or evaluate a member of $O_x\}$. Then:
novelty(x) = fundamental if $H_x \cap M_x \neq \emptyset$ and $\nexists H'_x$ such that $H'_x \cap M_x \neq \emptyset$ (i.e. if x could not have been generated without the use of meta level procedures); *mere* if $H_x \cap M_x = \emptyset$ (i.e. x has been generated using only object level procedures) and the complexity measure below is satisfied; and *none* otherwise.

This can only be applied to systems in which it is possible to distinguish M_x from O_x.

Novelty relative to complexity

In (Bundy, 1994) a complexity requirement for creativity is described. An item is considered creative if *(i)* the concept space in which it lies is large and complex and *(ii)* it is generated from a little explored area; i.e. exploration, rather than transformation is important. The more complex it is to find an item the more novel it is deemed. The problem with such a criteria is that there may be items that we want to call novel yet which are simple to generate. For example the HR program(Colton, 2001a) discovered the concept of *refactorable integers*, an integer whose number of divisors is itself a divisor (such as 8 which has 4 divisors - 1, 2, 4 and 8 - including 4). This concept was new to many mathematicians, including Colton (the programmer) and yet its generation history was simple.

• Complexity Measure

These measures judge the complexity of an item x, based on the size of the domain and how unusual and complicated the process for generating x was.
(i) Given a program P, we define its concept space (CS) to be the set of all items which it is possible (in principle) to generate: $CS_P = \{x : x$ may be generated by $P\}$. We then use the size of this set to denote the complexity of a CS of a program:
$complexity_i1 = |CS_P|$. Note that in many cases $|CS_P|$ will be infinite.
Since this may be difficult to calculate we define the *restricted concept space* of an item x to be the set of items which could be generated using the same (or less) *amount* of knowledge: $RCS_1(x) = \{x' \in CS_P : |H_{x'}| \leq |H_x|\}$. Then:
$complexity_i2 = |RCS_1(x)|$.
Alternatively we define the set $RCS_2(x)$ of all items which can be generated using a subset of the knowledge used to produce x: $RCS_2(x) = \{x' \in CS_P : H_{x'} \subseteq H_x\}$. Then:
$complexity_i3 = |RCS_2(x)|$.

(ii) We want a measure which will capture exactly those items which lie in a little-explored area of the search space. We assume that such items will have been generated using knowledge not often used.

We define bag_R and set_R as the bag and set of all pieces of knowledge (k) used to generate and evaluate all items x in R: $bag_R = \mathcal{B}_{x \in R} H_x$, and $set_R = \bigcup_{x \in R} H_x$. We then define a frequency measure $\mathtt{f}(k)$ for each $k \in bag_R$:
$\mathtt{f}(k)$ = the number of occurrences of k in bag_R.
Using this we define two further frequency measures:
average frequency(bag_R) = $\mathtt{af}(bag_R) = \frac{|bag_R|}{|set_R|}$ (this gives the mean number of times an individual k is used in the generation of all items in R), and
relative frequency(k) = $\mathtt{rf}(k) = \frac{\mathtt{f}(k)}{\mathtt{af}(bag_R)}$ (the number of times a specific k is used in the generation of all items in R as compared to the mean). If $\mathtt{rf}(k_i)$ returns a value less than 1 then k_i is used less frequently than the average k, and if the value is more than 1 then k_i is used more frequently than the average). We define three measures which consider all of the pieces of knowledge used in the generation and evaluation of an item x and their relative frequencies. Firstly we define a set which contains those pieces of knowledge not usually used: $Rare = \{k : \mathtt{rf}(k) < 1\}$. Then:
$complexity_{ii}1(x) = |H_x \cap Rare|$ (the number of pieces of knowledge used to generate x whose relative frequencies are less than the average frequency).

Secondly, we sum the relative frequencies of all k's used in the generation and evaluation of x, where the higher this sum, the lower the novelty of x:
$complexity_{ii}2(x) = \frac{1}{\Sigma_{\mathtt{rf}(x)} H_x}$

Thirdly, of all the k's used to generate and evaluate x, we take the number of times that the k with the lowest relative frequency was used. We define *minimum relative frequency*(H_x) = $min(\mathtt{rf}(H_x))$ = minimum of $\mathtt{rf}(k)$ over all $k \in H_x$. The higher this number, the lower the novelty of x:
$complexity_{ii}3(x)$=$\frac{1}{min(\mathtt{rf}(H_x))}$ (the amount of times which the least used overall piece of knowledge is used in the generation of x).

The final novelty measure may be binary, on the condition that both complexity measures (i) and (ii) above are higher than fixed values α and β, i.e.:
$novelty1(x) = 1$ if $complexity_i(x) > \alpha$ and $complexity_{ii}(x) > \beta$; 0 otherwise.

Alternatively it could be some function of the two, eg:
$novelty2(x) = \alpha complexity_i(x) + \beta complexity_{ii}(x)$.
Arguably the number of rules and amount of knowledge needed to generate x should also be included in the complexity measure. This could be done by adding a term such as $\gamma|H_x|$ to $novelty2(x)$.

These complexity measures can only be applied to systems where the relationship between k and x is clear (which excludes many sub-symbolic systems, where all knowledge is involved to some extent in the creation of each item). Although a useful measure of novelty within a system it is hard to see how the measures could be used to compare novelty of items generated by different systems. Objective measures such as those in (Ming, 1997) might be relevant in this regard. An upper bound could be defined in order to prevent bizarre items from scoring highly.

Novelty relative to an archetype
In (Ritchie, 2001) a fuzzy *typicality* set T is defined, which is based on a weighted vector of properties which characterise the items in a domain. These properties may be quantitative or qualitative, objective or subjective. For example a certain typical type of poem might contain a specific number of syllables, lines, etc. (syntactic features are not enough, but are the easiest to describe). We use this idea to define novelty of an item as the degree to which it belongs to the set $R\backslash T$. A related idea considers identifying one or more archetypes within an n-dimensional space and then defining the novelty of an item as the distance between it and its closest archetype. In both, the choice of properties/dimensions is restricted by what can be measured; properties such as 'humour' or 'passion' would probably have to be ignored or crudely estimated.
• Archetypal Measure
novelty1(x) = $\mu_{R\backslash T}(x)$ where $\mu_{R\backslash T}(x)$ denotes the degree to which x belongs to the fuzzy set $R\backslash T$. $\mu_{R\backslash T}(x) \in [0, 1]$, where 0 = *completely familiar* and 1 = *completely novel*)
Let $A = \{a : a$ is an archetype of a specified kind$\}$. Then
novelty2(x) = minimum(distance(x, a)) over $a \in A$

These measures (which are essentially the same) could both incorporate Boden's idea of going too far by specifying a minimum value for both the degree of membership of T and the maximum distance between x and a, after which novelty decreases.

Novelty as surprise
It is noted in (Macedo and Cardoso, 2001) that a characteristic of creative results is that at some step of their generation they surprise the creator. A measure of surprise therefore may be relevant in an assessment of novelty. We hold that surprise is a *reaction* to fundamental novelty, so fundamental novelty implies surprise. This is not weakened by the fact that trivially novel results (such as the number of blades of grass on the lawn) are not surprising, nor that surprising events (such as a sudden loud noise) may not be novel. The measure is useful therefore in showing that an item is *not* fundamentally novel (if it is not surprising) but cannot be used to prove that it is. However it may be used as a guide if it is argued that surprise partly captures what is meant by novelty. If used in conjunction with other novelty measures in this section this could prove a powerful measure.

We hold that surprise is dependent on *context* as well as *probability*. For example an event E may be surprising to one observer but not to another. Context refers to the way in which an individual classifies information. For example most people would classify 7^{th} June as a normal day, however a few will classify it as special (for instance if it is their birthday). If I say that my birthday is 7^{th} June then most will find it unsurprising, but the few who classify it as special will be surprised. This is because, supposing that a person has 10 special days in a year, classified as $Special$, and the others are classified as $Rest$, then:
$prob$(my birthday is in $Special$)= $\frac{10}{365}$ (a low probability and so it is surprising if it happens), and
$prob$(my birthday is in $Rest$)= $\frac{355}{365}$ (a high probability and so it is unsurprising if it happens).

This example shows that the surprisingness of the date of my birthday is dependent on both probability and context (how a person has classified the days in the year).
• Surprise Measure
$surprise(x)$ is dependent on the individual probabilities of an event which is classified as similar to x happening. We define it as:
$surprise(x) = 1 - \frac{\Sigma_{y \in Y} sim(x,y) * prob(y)}{\Sigma_{y \in Y} sim(x,y)}$, where Y is the set of comparable events and $sim(x, y)$ denotes the similarity of x and y.

Clearly this now opens the question of 'what is a similar event?'. However although this is difficult to formalise, in practice there may be obvious ways of categorising events.

Perceived novelty
Although it is difficult to formalise our notion of novelty, humans are often capable judges of what is novel. Therefore a principled test in which humans are asked to judge the degree of difference between items in I and those in R may be preferable to a more formal analysis of the items. In (Pearce and Wiggins, 2001) a series of experiments are described which test whether pieces of music produced by a program (R) are perceived to be in the same style as those input to the program (I). In particular a *discrimination test* is outlined in which subjects are asked to distinguish

items in R (system generated compositions) from those in I (human generated compositions). In post-experimental analysis subjects stated that they categorised items which seemed overly conservative or chaotic as system generated (R). It is suggested that if subjects are unable to distinguish system from human generated pieces then the system must be generating items with an appropriate degree of variation (novelty) from those in I.

• Perceived Novelty Measure

Following the ideas above, experiments in which subjects are asked to:

(i) distinguish items in I from those in R,

(ii) rate the degree of variation between items which are all from I; and the degree of variation between items which are all from R,

(iii) rate the degree of variation between a pair of items which are both from I; and the degree of variation between a pair of items, one of which is in I and the other in R

should be performed and analysed statistically to determine perceived novelty. Items in R are considered novel if: *(i)* subjects are unable to reliably identify them as system generated, i.e. the mean proportion of items in R which are correctly classified is less than or equal to 0.5 (a mean of 0.5 would be expected if subjects classified the items randomly); *(ii)* there is a comparable degree of variation between items which are either all from I or all from R, i.e. there is no difference between the mean perceived degree of variation in subsets of I and subsets of R; *(iii)* there is a comparable degree of variation between items in I and R, i.e. there is no difference between the mean perceived degree of variation between pairs of items in which both come from I and pairs in which one item is from I and the other from R.

Pearce and Wiggins have performed experiments *(i)* and *(ii)* in the music domain but suggest that they could generalise to other domains such as painting. It is also possible to see how scientific results could be analysed in this way. For example experts may have some intuition about the comparability of the concepts *prime*, *amicable* and *odious* numbers[1] without being able to formalise that intuition. This might be captured in the above experiments. As suggested in (Pearce and Wiggins, 2001), using experts as subjects may be preferable to novices as they are better able to gauge the degree to which items differ. Asking experts which items they prefer might also be fruitful as they may prefer items which are atypical, as opposed to novices who may be more likely to prefer prototypical examples of a genre.

Quality

The difficulty of measuring the quality of creative items is reflected by the large number of examples of work that was not valued at the time it was produced. Examples arise in artistic domains, for example Van Gogh's paintings, as well from fields we might expect to be more objective: consider the initial reaction to group theory, immunisation, or the jet engine. The reason for the mistakes is that h-creative work by definition cannot be subject to familiar criteria. Indeed the ability to measure quality in a field without mistakes would imply that that field was incapable of any further transformation. Therefore we cannot expect to ever measure quality without making some mistakes. Despite our fallibility, however, it is useful to examine our notion of quality in order to develop practical measures.

Quality relative to emotional response

The quality of a piece of music or poem is often judged by the extent to which it evokes emotions in an audience. We consider two types of emotional response; firstly *any* emotional response at all, and secondly any *positive* emotional response. A good item in the first sense will make at least one person feel joyful, sad, hopeful etc., as contrasted with a worthless item which evokes only indifference[2]; and in the second will affect at least one person positively, as contrasted with affecting people negatively. Therefore we consider the quality of an item to be dependent on the number of people it affects (either at all or positively) and the extent. The self-centred version of the latter interpretation - does x positively or negatively affect *me*? - is arguably the most common judgement of quality.

We can measure both criteria by conducting experiments in which subjects are asked to record their emotional reaction to an item. The question 'is x a good poem?' is taken to be asking 'does x evoke an emotional response from anyone, and if so is it a positive affect?' An emotive agent might apply this measure to give an internal evaluation of its work.

• Emotional response measure

Let S be a sample of subjects who have been asked to evaluate an item x according whether it affects them positively, negatively or neither, and to what degree.

Let $Pos(x) = \{d_s : s \in S$ rates x positively to degree $d\}$, and $Neg(x) = \{d_s : s \in S$ rates x negatively to degree $d\}$. (Note that someone may add a (nonzero) degree to both sets since an item may affect her both positively and negatively. An indifferent reaction would score zero in either set.)

Also let $TotalPos(x) = \sum_{d_i \in Pos(x)} d_i$,

$TotalNeg(x) = \sum_{d_i \in Neg(x)} d_i$,

and $TotalAff(x) = TotalPos(x) + TotalNeg(x)$.

We define the following measures:

(i) if the criteria is *intensity* of emotional response:

$quality1_i(x) = TotalAff(x)$ (absolute quality)

$quality2_i(x) = \frac{TotAff(x)}{|S|}$ (average quality)

(ii) if the criteria is the *type*, as well as intensity, of emotional response:

$quality1_{ii}(x) = TotalPos(x)$ (absolute net quality)

$quality2_{ii}(x) = TotalPos(x) - TotalNeg(x)$ (absolute gross quality)

$quality3_{ii}(x) = \frac{TotalPos(x)}{|S|}$ (net quality relative to the number of people asked)

[1] Two numbers are *amicable* if the sum of the divisors of each equals the other (for example 220 and 284), and an *odious* number is one whose binary representation consists only of 1's (for example 15 has binary expansion 1111).

[2] We exclude surprise from this discussion as methods for measuring it have already been outlined, and it is seen as indicative more of the novelty than quality of an item.

Although these measures may seem successively more sophisticated and therefore 2_i and 3_{ii} preferable, they are all worthy of empirical testing. We may not want to say that the more people hear but do not react to a difficult piece of music, the less good it is (i.e. $quality1$ may capture our notion better than $quality2$ or 3). Both *(i)* and *(ii)* will favour popular work, so Mills & Boon books might well score higher than the classics (although they would score low on novelty). A more sophisticated approach would take into account the different types of audience. Instead of one number, this would give a quality distribution, allowing flexible evaluations such as 'good for teenagers but not parents'.

Quality relative to an aim

Quality may refer to the extent to which an item solves a problem, or achieves the aim for which it was produced. The question 'is x a good poem?' is asking 'does x express the ideas or evoke the emotions which were intended?'.

• Pragmatic measure

In some domains we can list items which are known to satisfy an aim. For example in concept formation the aim might be to generate interesting concepts. The quality of a concept x can be tested by whether it belongs to a set of known interesting concepts (within, eg. number theory). Below we define Sat as $\{s{:}s$ satisfies $aim\}$. In other domains it is not feasible to define such a set but it may be possible to write a marking criteria (MC) which rates the extent to which x satisfies the aim. In order to be as objective as possible this would consist of a weighted sum of largely objective criteria. For example in music the aim might be to compose a piece of music in the style of Mozart and x be a composition. A MC would include criteria such as musicological and stylistic aspects (MC used by examiners provide a useful source). We define:

$quality1(x) = 1$ if $x \in Sat$, and 0 otherwise.

$quality2(x) = m$ where m is the mark awarded to x according to the MC.

Clearly it is often difficult to assess quality, both as the extent to which it evokes an emotional response and in terms of satisfying an aim. Both senses may be meant in creativity, and overall quality is likely to be a function of both senses (note that an item regarded as good in one sense may not be good in the other). The emotional sense excludes work which is written and then lost before anyone has been exposed to it (which might be good according to the second sense). In general we hold that the first measure is more appropriate to evaluating the worth of items in artistic domains and the second to items in science. Scientific results can evoke emotional responses but these are not necessarily indicative of their worth. For example a mathematician may become very excited by a proof later shown to be faulty (which may be of value if it leads to the correct proof - but may simply be misleading and a waste of time). On the other hand Newton's mechanics made people feel worried because they perceived the idea as threatening to their notion of free will, yet the fact that it elicited emotion is incidental to it being a good theory. In the artistic domain however, a general aim might be to elicit emotion (of any kind), so even if an item fails on a *specific* aim (for example the aim of portraying the feeling produced when watching lambs play) it may still be considered good if it evokes an emotional response.

It should also be noted that quality(x) (and hence the creativity of x) is a continually changing function, depending on the environment. The degree to which a poem conveys its intended message, for instance, depends on who reads it.

Process

The process by which an item has been generated and evaluated is intuitively relevant to attributions of creativity. Consider the story of Euler's (p-creative) discovery of Arithmetic Series. A class of unruly pupils was told by their exasperated teacher to add up all the numbers between 1 and 100. All of the pupils calculated the answer 5050, but everyone except Euler laboriously added each of the numbers. Euler realised that if they were written in ascending order and then underneath in descending order, the sum of each of the pairs was 101, and there were 100 pairs. Therefore twice the required sum was 10100, and the answer was 5050.

	1	2	3	...	99	100
+	100	99	98	...	2	1
	101	101	101	...	101	101

If we consider 5050 to be the output, and the formula $\frac{n*(first+last)}{2}$ to be the process, then it appears that process *is* relevant to the attribution of creativity, and that it can (partially) be known. We know how Euler produced the answer using the formula, although not how he produced the formula. However it could be argued that since the value lies in the *formula*, rather than the specific *number*, it is the formula which is the output. Since we do not know how Euler arrived at the formula, his method cannot be part of our creativity judgement.

We often do make judgements regarding the creativity of others with little (if any) knowledge of the processes behind an output. Three viewpoints attempt to justify this. Firstly there is the view that *process is irrelevant*, creativity judgements can be made purely on the basis of output. The twins described in (Sacks, 1985) who produced results in number theory by 'seeing' patterns of prime numbers (but were unable to articulate how they had done it) were just as creative as someone producing the same results by more conventional methods. Secondly, there is the pragmatic view that although *process may be theoretically relevant* to judgements of creativity, since we cannot know underlying human processes we must use other criteria which we can know, such as novelty and quality. Finally there is the belief that *process is relevant* to attribution of creativity, and we can at least partially know it. In (Hofstadter, 1994) it is argued that we can probe underlying processes to an arbitrary level if we examine external behaviour sufficiently carefully. However since we *can* know the processes that occur within a program these beliefs must be reconsidered. If process does matter then we must argue *why* and specify *how*. This will be determined by whether our aims come from engineering or cognitive science.

We assume a two stage model, generation and evaluation. Further questions regarding process include *who* carries out these stages and *when*. For instance the two stages may have been carried out by different systems (consider brainstorm-

ing techniques in which one person generates and another evaluates, or an Interactive Genetic Algorithm in which a program generates and a human evaluates). If so, then we consider the group of contributing systems an entity and any resulting creativity is attributed to it as a whole. (All evaluation in this section, therefore, is carried out internally.) 'When' might refer to the order of the stages; for example only valuable items may be generated, each item evaluated immediately, or many items generated before evaluation.

Methods of generation

Randomness

The level of randomness used to produce an item may be high (if it is generated mainly or completely due to random procedures, for example throwing a pot of paint at a canvas); low (if randomness plays a small role in generating it, for example accidentally buying the wrong colour paint and then finding that it looks better anyway); or none (if it is generated entirely according to systematic rules). The majority held view is that items with a high level of randomness are less creative than those with a lower level[3], and items which are completely determined are not creative.

Low level randomness is certainly present in human creativity. In (Boden, 1990) examples are listed including serendipity (eg. the low level randomness described above), coincidence, and unconstrained conceptual association (brainstorming). She stresses that randomness does *not* necessarily mean undetermined; and defines *relative randomness* to be a lack of order relative to a specific (relevant) consideration. If an event (eg. a die landing on a 6) has an *independent* causal history from another (eg. my game of craps), then it is relatively random to it. The die landing on 6 *is* determined, but by factors which are independent of my game, i.e. gravity and the way in which I throw it.

Low level randomness has also been found useful in computational models of creativity. For instance in COPYCAT (McGraw and Hofstadter, 1993) microexploration is random (they consider it to be the most efficient method of initially generating ideas since it is equivalent to non-biasedness). If there is no undetermined randomness, some would claim that since the process may be completely traced and all element of mystery eliminated, it cannot be called creative. However if the concept of relative randomness is accepted then the fact that an item is determined is not sufficient to exclude it from being potentially creative. This idea parallels the hard determinist argument which states that since every event is caused there is no free will (I may cause event E but since I am myself caused I am not responsible and therefore I cannot be free). It is contrasted with soft determinism, which argues that there is free will *because* every event is caused. (If I cause event E then I am responsible for it and therefore free, and the fact that I am myself caused is irrelevant.)

We can measure the degree of randomness in program which generates an item x according to its replicability; i.e. the likelihood of generating an item like x, given the same input conditions as those from which x was produced.

• randomness

Let $p(output = x|input = i)$ be the probability that, given $input\ i$, item x is generated, and $distance(x, x')$ be a difference measure of the distance between two items x and x'. CS is the concept space of the program which generated x. Then we define:

$randomness1(x)=\frac{\Sigma_{x' \in CS} p(output=x'|input=i)*distance(x,x')}{\Sigma_{x' \in CS} distance(x,x')}$

This measures the probability that an item which is similar to x will be produced given the same input conditions as those with which x was produced. Its validity depends on the distance measure defined. It can be applied to items in a finite or countably infinite concept space. If the concept space is continuous then the density instead of probability function should be taken; i.e.

$randomness2(x)=\int d(output=x'|input=i)*distance(x, x')$

where $d(output = x'|input = i)$ = density function for the distribution of x' given $input = i$.

This may be categorised into *none* if $randomness(x) = 0$; and *low* or *high* if a number $\alpha \in (0, 1)$ is introduced such that $randomness(x)$ is low if $0 < randomness(x) < \alpha$, and high if $randomness(x) > \alpha$ (although it would be hard to justify any specific α). The sets *none*, *low* and *high* could be fuzzy. There would be no single correct definition for these terms, but definitions should largely agree on the categorisation of most items.

Alternatively, $randomness3(x)$ could be a function of each random procedure used in generating x. For example if x is generated by a GA then this function might be $IR+nG$ where IR = the level of randomness in generating the initial population, n the number of generations it took to produce x, and G the randomness present in each generation (determined by the level of randomness in the selection, crossover, mutation and insertion procedures).

Methods of evaluation

Two kinds of evaluation are relevant; the evaluation of the item, and evaluation of the processes used to generate it. We assume that the former must occur, following Boden who states that in order to be creative one must recognise the worth of an idea as well as actually have it. The latter may or may not occur (the question of whether one has to be aware of which rules one is breaking and how, is seen as controversial). Both types may take place either during or after generation. We can split this into *process*, i.e. *how* the system evaluates, and *output*, i.e. how *good* its evaluations are. Here we only consider the latter.

Evaluating the item

The extent to which an evaluation carried out by the creative system corresponds to that carried out externally may be tested by statistical methods. This was demonstrated in (Steel, 1999) who evaluated output from a run of the HR program (rating concepts as highly, quite, potentially or not at all interesting) and then compared his evaluation with that of HR. It shows the external judgement of how reliable

[3]The role of evaluation may effect the judgement. For example if an artist did one painting randomly and did not evaluate it, this would be less creative than if he had done many paintings randomly and picked out the best one.

the creator's evaluation is. A graph may be plotted in which each point represents the creator's and external evaluation for a given item. The level of correlation, r, and the degree of confidence in r can then be calculated.

• Evaluation of Item Measure

Let E_C = $\{c_x$: c_x is the creator's evaluation of $x \in R\}$, E_E = $\{e_x$: e_x is the external evaluation of $x \in R\}$, and $\overline{c}$ and $\overline{e}$ respectively be the mean creator's and external evaluations over all $x \in R$.

Let $r_1 = \frac{\sum d_c d_e}{\sqrt{\sum {d_c}^2 \sum {d_e}^2}}$ where $d_c = c - \overline{c}, d_e = e - \overline{e}$. This is the product-moment coefficient (in which items are given a numerical measure of value).

Also let $r_2 = 1 - \frac{6 \sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$ where $d_i = c_i - e_i$ and n is the size of sample (i.e. $n = |E_C| = |E_E|$). This is the rank correlation coefficient (in which items are ranked according to value). Then:

$evaluation1(R) = r_1$ if $r_1 > 0$, and 0 otherwise.

$evaluation2(R) = r_2$ if $r_2 > 0$, and 0 otherwise.

This can be measured to any given confidence level (which depends on the size of the set E_C). Clearly this measure is only as reliable as the external evaluation which is likely to be subjective. For domains in which this evaluation is harder to measure the rank correlation coefficient might be preferable.

Evaluating the process

Boden argues that process does matter, stating that a program is creative only if it produces items *in the right way* - by transforming the boundaries of a conceptual space. This, she claims, can only be done if the program contains reflexive descriptions which mark its own procedures and is capable of varying them. The program should contain a meta-level which assesses methods of transforming a space and considers when and how to apply them. This is supported by (Bundy, 1994) who advocates a self reflection criteria for creativity.

Evaluation of the procedures used to generate and evaluate items may be judged by considering a procedure as an item and testing for correlation between the creator's and external evaluation, although the external evaluation of procedures may be more difficult than other domains. Alternatively it can be judged by the quality of the items it outputs and the extent to which this is effected by the inclusion of a meta-level. Comparing items generated both with and without a meta-level reflects this criteria. For example the quality of items produced by META-DENDRAL may be compared to that of items produced by DENDRAL(Buchanan and Feigenbaum, 1978); those of EURISKO(Lenat, 1983) to those of AM(Lenat, 1976); and those of HR with meta-level capabilities (Colton, 2001b) to those of HR without (Colton, 2001a).

• Evaluation of Process Measure

We consider two output sets to be comparable if they have been produced using the same CPU time or contain the same number of items, as appropriate. The quality of a set, $qual(S)$ is defined to be $\sum_{x \in S} quality(x)$ where $quality(x)$ is one of the measures outlined above. Now let R_M and R_O be two comparable output sets produced with and without access to a meta-level respectively. We define:

evaluation($processes$) = $qual(R_M) - qual(R_O)$

Note that the same measure of quality must be used in both calculations. If this is 0 or negative then the meta level is not contributing positively to the quality of the output and therefore its evaluations of its generating procedures are not satisfactory. If it is greater than 0 then its evaluations are having a beneficial effect. If CPU time is used then it may be appropriate to divide both $qual(R_M)$ and $qual(R_O)$ by the number of items each program produced. A weakness of this measurement is that it presupposes that two programs exist which are the same in every respect other than the existence of a meta level.

Evaluation of machine creativity

How to measure success

We stated at the start of this paper that we are aiming, through the study of machine creativity, to *(i)* further our understanding of creativity (human and other), and/or *(ii)* build programs which are useful in achieving practical goals. In order to assess whether this has been achieved we must firstly consider the success of *(i)* or *(ii)*. For *(i)* we may argue that we now have better concepts such as the concept of a *heuristic* (which was due to (Polya, 1962) but has since been developed computationally), or of a *generative system*. Examples of the second aim include *(iia)* tools for enhancing human creativity and *(iib)* a creative program which performs better than a non-creative one. For *(iia)* we could show that there is someone who is better able to create with the help of a program. For example some musicians use programs as an aid to their compositional process. The more people and the greater the extent that they use the tool, the more successful we have been. *(iib)* can be assessed according to how well the creative program competes with non-creative programs. This raises the question of what differentiates creative from non-creative systems. That is, once we have shown that our aims have been achieved, we must then show that they have been achieved in the way that we claimed, i.e. *through the study of machine creativity*.

How to attribute creativity

Boden suggests that it is helpful to regard aspects such as novelty, quality and process as dimensions of creativity. Instead of asking 'is x creative?' (assuming a boolean judgement) or 'how creative is x?' (assuming a linear judgement) we should ask 'where does x lie in creativity space?' (assuming an n-dimensional space for n criteria where we can measure each dimension). Ritchie also suggests a catalogue of criteria which can be combined in different ways according to different interpretations of creativity.

Analogy to the attribution of intelligence

After fifty years of study we do not expect to be able to attribute a program with x amount of intelligence. Progress is measured instead by subdividing it into different aspects (ability to perceive and react, plan, reason etc.) which are each measured according to appropriate methods. Although

these aspects are seen as different dimensions of intelligence (which is assumed to emerge through the interaction of the aspects) little attempt is made to order the dimensions or compare systems which incorporate different aspects. The question which faces us now is whether we should see creativity as analogous to intelligence or as a particular aspect of it. In either case we need to subdivide it further, but what these sub-aspects mean in terms of overall creativity would differ. If it is seen as analogous to intelligence then evaluation of creativity may turn out to be based on our best yet test for intelligence - the Turing Test (Turing, 1950). If it were considered one aspect of it then attribution would be some function of all its sub-aspects. In such a young field it would be premature to decide which is the more fruitful analogy. Instead we must structure our research by developing theoretical (which are the key aspects and why?) and practical (how can we measure them in our programs?) approaches to machine creativity. It is currently more appropriate to ask 'is $P1$ or $P2$ more creative in aspect A?' (particularly if $P1$ and $P2$ are versions of the same program) rather than 'where does x lie in creativity space?'.

Conclusion and further work

We have outlined what we consider to be the key aspects of assessing creativity, and highlighted associated justifications and controversies. Methods of measuring the aspects have been outlined, which are somewhat arbitrary and are therefore intended as suggested starting points rather than definitive standards. These should now be assessed according to two criteria. Firstly, to what extent do they reflect human evaluations of creativity, and secondly, how applicable are they? The first should be tested empirically. If examining our concept leads to another, more refined concept - $creativity2$ - then theoretical reasons should be put forward for modelling this new concept (and then empirical grounds for using the measures to capture the new concept). The measures can be assessed according to the second criterion by attempting to apply them to AI programs, which will show whether they are practically feasible.

We expect that both of the above methods of assessment will either provide evidence for the measures suggested or lead to more sophisticated ones. In this way research on evaluating machine creativity can proceed scientifically, by a series of falsifiable claims. We hope that this will lead to a deeper understanding of the nature of creativity.

Acknowledgements

We are grateful to the participants of the AISB-01 symposium on creativity for a most inspiring discussion, led by Geraint Wiggins. We would also like to thank Alan Smaill, Greame Ritchie and two anonymous reviewers for their helpful comments on earlier drafts. This work was supported by EPSRC grants GR/M45030, 00304543 and GR/M98012. The third author is also affiliated with the Department of Computer Science, University of York.

References

Boden, M. (1990). *The Creative Mind: Myths and Mechanisms*. Weidenfield and Nicholson, London.

Boden, M. A. (1994). Creativity: a framework for research. *Behavioural and Brain Sciences*, 17(3):558–556.

Buchanan, B. and Feigenbaum, E. (1978). Dendral and metadendral: Their applications dimension. *AI*, 11.

Bundy, A. (1994). What is the difference between real creativity and mere novely? *Behavioural and Brain Sciences*, 17(3):533 – 534. Open peer commentry on (Boden, 1990).

Colton, S. (2001a). *Automated Theory Formation in Pure Mathematics*. PhD thesis, University of Edinburgh.

Colton, S. (2001b). Experiments in meta-theory formation. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science*.

Hofstadter, D. (1994). *Fluid Concepts and Creative Analogies*. HarperCollins, New York, USA.

Lenat, D. (1976). *AM: An Artificial Inteligence approach to discovery in mathematics*. PhD thesis, Stanford University.

Lenat, D. (1983). Eurisko: A program which learns new heuristics and domain concepts. *Artificial Intelligence*, 21.

Macedo, L. and Cardoso, A. (2001). Creativity and surprise. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science*.

McGraw, G. and Hofstadter, D. (1993). Perception and creation of diverse alphabetic styles. *AISBQ*, (85):42 – 49.

Ming, L. (1997). *An introduction to Kolmogorov complexity and its applications*. Springer, New York, USA.

Pearce, M. and Wiggins, G. (2001). Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on AI and Creativity in Arts and Science*.

Perkins, D. N. (1996). Creativity: Beyond the darwinian paradigm. In Boden, M., editor, *Dimensions of Creativity*, pages 119–142. MIT Press, Cambridge, MA, USA.

Pind, J. (1994). Computational creativity: What place for literature? *Behavioural and Brain Sciences*, 17(3):547 – 548. Open peer commentry on (Boden, 1990).

Polya, G. (1962). *Mathematical Discovery*. John Wiley and Sons, New York, USA.

Ritchie, G. (2001). Assessing creativity. In *Proceedings of the AISB'01 Symposium on AI and Creativity in Arts and Science*, pages 3 – 11. SSAISB.

Sacks, O. (1985). *The man who mistook his wife for a hat*. Duckworth, London.

Steel, G. (1999). Cross-domain mathematical concept formation. Master's thesis, University of Edinburgh.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59:433–460.

Wallas, G. (1926). *The Art of Thought*. Harcourt Brace, New York, USA.