Categorising creative systems

Geraint A. Wiggins
Centre for Computational Creativity
City University, London
geraint@city.ac.uk; www.soi.city.ac.uk/˜geraint

Abstract

I present some simple reasoning about creative behaviour based on a framework derived from the work of Boden [Boden, 1990; 1994]. The aim is to move further towards a model which allows detailed comparison, and hence understanding, of systems which exhibit behaviour which would be called "creative" in humans. The work paves the way for the description of more natural, multi-agent creative AI systems.

1 Introduction

Boden [Boden, 1990; 1994] presents a broad philosophical framework for the description of creative systems. Wiggins [Wiggins, 2001] developed Boden's informal descriptive framework into a more formal (through preliminary) descriptive system, with the intention that the resulting formalism could be used as a framework within which to analyse, evaluate and compare creative systems. The framework has been used (perhaps before it was quite ready!) in this way by Gervás [Gervás, 2002b; 2002a].

Gervás demonstrated very clearly that the application of such frameworks is not simply objective, but is heavily dependent on the viewpoint of the person doing the applying. In this paper, I propose some further details of the framework, in particular focussing on how it can facilitate the description of various sub-categories of creative behaviour not explicitly introduced by Boden.

In §2, I briefly outline the framework, and make one very small but helpful refinement. In §3, I specify and name a number of situations, which might be called "significant" within a creative process, and explain how they can be modelled in the formalism. In §4, I suggest how the outcomes of §3 can be used to propose solutions to the problems involved. In §5, I discuss the consequences of these solutions.

2 Background

The central plank of the formalism presented by Wiggins [Wiggins, 2001] is that an exploratory creative system, in Boden's terms [Boden, 1990], may be abstractly represented by a septuple, thus:

$$\langle \mathcal{U}, \mathcal{L}, [\![.]\!], \langle\!\langle . \rangle\!\rangle, \mathcal{R}, \mathcal{T}, \mathcal{E} \rangle.$$

Here, I shall use a septuple as follows, because of a small but significant detail, the point of which will become clear later:

$$\langle \mathcal{U}, \mathcal{L}, [\![.]\!], \langle\!\langle ., . \rangle\!\rangle, \mathcal{R}, \mathcal{T}, \mathcal{E} \rangle.$$

The symbols here are defined, as before, as follows. The function of each is briefly explained below (see [Wiggins, 2001] for more detail). $\mathcal{S}^\star$ is the set of sequences generable from the elements of $\mathcal{S}$.

$\mathcal{U}$	a universe of possible concepts, both partial and complete
$\mathcal{L}$	an alphabet from which to build rules
$\mathcal{L}^\star$	a language, derived from $\mathcal{L}$, in which to express rules
$[\![.]\!]$	a function generator, which maps a subset of $\mathcal{L}^\star$ to a function which selects elements of $\mathcal{U}$
$\langle\!\langle ., . \rangle\!\rangle$	a function generator, which maps two subsets of $\mathcal{L}^\star$ to a function which generates new elements of $\mathcal{U}$ from existing ones
$\mathcal{R}$	a subset of $\mathcal{L}^\star$
$\mathcal{T}$	a subset of $\mathcal{L}^\star$
$\mathcal{E}$	a subset of $\mathcal{L}^\star$

$\mathcal{U}$ is the (abstract) set of all possible partial and complete concepts. $\mathcal{R}$ is a set of rules, expressed using the language $\mathcal{L}^\star$, generated from the symbols in $\mathcal{L}$, which select an "acceptable" or "relevant subset" of $\mathcal{U}$ referred to by Boden [Boden, 1990] as the *conceptual space*. This space contains all the concepts which are in some sense appropriate to the creative outputs in question, and broadly characterise an area of creative output (*e.g.,* a musical or literary style or a branch of mathematics). In Wiggins' formulation, this is extended to include partial concepts, which means concepts, not all of whose defining properties are specified, but which might be instantiated to be concepts in the conceptual space in principle. So applying a selector function generated from $\mathcal{R}$ by $[\![.]\!]$ gives Wiggins' equivalent of Boden's conceptual space:

$$[\![\mathcal{R}]\!](\mathcal{U}).$$

$\mathcal{T}$ is a set of rules which, when interpreted by $\langle\!\langle ., . \rangle\!\rangle$, describe the behaviour of a creative agent as it traverses the conceptual space from known concepts to unknown ones (much as does a standard AI search engine). The first argument of $\langle\!\langle ., . \rangle\!\rangle$ takes a constraining rule set, such as $\mathcal{R}$, above, and the second a rule set such as $\mathcal{T}$. For analytical purposes, it is convenient to separate these two out *a priori*, rather than mixing them together as in the former version of the framework.

$\mathcal{E}$ is a set of rules which define the evaluation of the creative outputs resulting from the agent's activity, appropriately contextualised. The formalism does not specify what this context is; it might be the subjective judgement of the agent, or of other agents, or comparison with some objective measure. These issues are not the focus of the current paper.

A further useful mechanism is the function $^\diamond$, defined such that

$$\mathcal{F}^\diamond(X) = \bigcup_{n=0}^{\infty} \mathcal{F}^n(X),$$

where $\mathcal{F}$ is a set-valued function of sets. A useful constant will be $\perp$, the null (or completely undefined) concept.

A brief example may help to clarify the usage of this mechanism. Consider the familiar example (*e.g.,* [Ebcioğlu, 1988]) of the harmonisation of 17th Century German hymn tunes in the style of J. S. Bach. We can model this case as follows (but note that there are other ways, depending on what one wants to achieve). $[\![\mathcal{R}]\!]$ selects a subset of $\mathcal{U}$ which might be described as the set of all partial and complete harmonisations of the canon in question. $\mathcal{E}$ then selects those which are considered good. To see why there is a difference between $\mathcal{R}$ and $\mathcal{E}$, consider the comparison between the harmonisations produced by J. S. Bach himself, and those produced by a 1st-year music student: the latter are not usually valued as highly as those of the former, because even the best student is unlikely to produce music of the same quality as those Bach harmonisations which have been selected by history at this early stage in his or her career.

This same pair of subjects can help understand the need for $\mathcal{T}$, also. An extremely competent and experienced composer and improviser such as Bach will normally have the ability to "see" a harmonisation which is correct in syntactic terms and of high quality in value terms more or less without conscious effort. This is rarely true of beginning composers, who need to develop their intuitions over a period of time, usually through a kind of problem-solving approach. $\mathcal{T}$ allows us to model these behaviours individually, and to study their interactions with the externally defined $\mathcal{R}$ and $\mathcal{E}$.

Wiggins [Wiggins, 2001] shows how *transformational* creativity [Boden, 1990] can be cast as exploratory creativity at the meta-level, where the conceptual space is the set of possible rule sets.

The substantive difference between Boden's formulation and that of Wiggins is the addition of the rule set, $\mathcal{T}$, which describes the actual behaviour of a creative agent as it goes about its business – Boden is not concerned with this level of detail. The difference gives Wiggins' formulation more power to describe the behaviour of *implemented* creative systems. In particular, there is the question of how $\mathcal{R}$ and $\mathcal{T}$ interact, and how both interact with $\mathcal{E}$. This question is the focus of this short paper.

3 Characterising some creative circumstances

The apparent supposition in Boden's work is that creative agents will be well-behaved, in the sense that they will either stick within their conceptual space, or alter it politely and deliberately by transformation. It can be argued, however, that this is not adequate to describe the behaviour of real creative systems, natural or artificial, either in isolation or in societal context. This section looks at some situations not covered by the assumption of good behaviour, and gives names to them. The important point is that some of these situations may appropriately trigger particular events, such as a step of transformational creativity, so it is useful to be able to identify them in the abstract. This leaves us with several general classes of small-scale conditions which might be observed in AI systems; we can then assess their creative potential.

3.1 Uninspiration

There are various ways that a supposedly creative agent can fail to be creative in a valued way. These ways can be characterised through the rule set $\mathcal{E}$.

Hopeless uninspiration

The simplest case, *hopeless uninspiration*, is where there are no valued concepts in the universe:

$$[\![\mathcal{E}]\!](\mathcal{U}) = \emptyset.$$

This system is incapable, by definition, of creating valued concepts, and as such might be termed ill-formed (if such creativity is the intention).

Conceptual uninspiration

Another form, *conceptual uninspiration*, arises when there are no valued concepts in the conceptual space:

$$[\![\mathcal{E}]\!]([\![\mathcal{R}]\!](\mathcal{U})) = \emptyset.$$

I label this form of uninspiration "conceptual" because it entails a mismatch between $\mathcal{R}$ (which defines the conceptual space) and $\mathcal{E}$(which evaluates concepts within it, and, more broadly, within $\mathcal{U}$). This condition is contradictory to the purpose of the two rule sets: if $\mathcal{R}$ is supposed to constrain the domain of a creative process, then it is inappropriate for $\mathcal{E}$ not to select some of the elements it admits. As such, like the hopeless case, conceptual inspiration indicates ill-formation of the intended-creative system.

Conceptual uninspiration can only be remedied by transforming $\mathcal{R}$, by modifying $\mathcal{E}$ or by aberration (see below), which in itself requires transformation. How $\mathcal{R}$ and/or $\mathcal{E}$ should be modified is an open question whose answer is presumably domain-dependent.

Generative uninspiration

In *generative uninspiration*, the technique of the creative agent does not allow it to find valued concepts within the space constrained by $\mathcal{R}$:

$$[\![\mathcal{E}]\!](\langle\!\langle \mathcal{R}, \mathcal{T} \rangle\!\rangle^\diamond(\{\perp\})) = \emptyset.$$

This kind of uninspiration is less serious than the other two, and does not necessarily indicate an ill-formed creative system: it merely indicates that a creative agent is looking in the wrong place. This raises the question of *why* there is such a mismatch. Boden's underlying assumption seems to be that the conceptual space is in some sense definitive, and, certainly, in a multi-agent environment, it is the only place in the formalism where the consensus about a creative domain can logically be represented. Therefore, I propose that the usual solution to generative uninspiration would be transformation of $\mathcal{T}$, for the agent concerned, but that transformation of $\mathcal{R}$ (instead, or as well) may also be a valid response, noting that such transformation may be non-trivial in a multi-agent environment.

3.2 Aberration

Now, consider the following more interesting scenario, which also concerns the relationship between $\mathcal{R}$ and $\mathcal{T}$. A creative agent, **A**, is traversing its conceptual space. From any (partial) concept in the conceptual space, **A**'s technique will enable it to create another concept. Suppose now that the new concept is not in the expected style (note that there is no guarantee that it should be so – there is only an assumption in Boden's work), and is therefore not selected by $\llbracket\mathcal{R}\rrbracket$. In this case, the set $\mathcal{A}$ given by

$$\mathcal{A} = \langle\langle \mathcal{R}, \mathcal{T} \rangle\rangle^{\diamond}(\{\bot\}) \setminus \llbracket\mathcal{R}\rrbracket(\mathcal{U})$$

is non-empty. I term this *aberration*, since it is a deviation from the notional norm as expressed by $\mathcal{R}$. The choice of this rather negative terminology is deliberate, reflecting the hostility with which changes to accepted styles are often met in the artistic world.

The evaluation of this set of concepts is actually slightly more complicated than the single-concept motivating case outlined above. The aberrant but valued subset, which I call $\mathcal{V}_{\mathcal{A}}$ here, is calculated thus:

$$\mathcal{V}_{\mathcal{A}} = \llbracket\mathcal{E}\rrbracket(\mathcal{A}).$$

Because we are working in the extensional limit case, with all the created concepts notionally elaborated, we have to consider the possibility that all aberrant concepts, some aberrant concepts or no aberrant concepts may be valued. I term these *perfect* ($\mathcal{V}_{\mathcal{A}} = \mathcal{A}$), *productive* ($\mathcal{V}_{\mathcal{A}} \subset \mathcal{A}$) and *pointless* ($\mathcal{V}_{\mathcal{A}} = \emptyset$) aberration, respectively.

4 Responding to creative circumstances

The characterisations in §3 are only descriptively useful if appropriate responses, categorised by condition, can be specified. This section does so. I assume some appropriate learning mechanism(s) which can adapt the rules expressed in $\mathcal{L}$, from positive and/or negative training sets.

4.1 Uninspiration

Hopeless uninspiration

There is no solution to hopeless uninspiration within the specified universe; there is no capacity within the system to solve the problem. Therefore, it is up to the system designer to remedy the problem, like a *deus ex machina*.

Conceptual uninspiration

The only means within the system to address this issue is the transformation of $\mathcal{R}$. In Boden's terms, this would probably not be appropriate, since the $\mathcal{R}$ set is rather more definitive than in Wiggins' terms. However, in the general (multi-agent) case, where $\mathcal{R}$ must really reflect some kind of consensus among agents about a particular domain, it would clearly be appropriate to modify $\mathcal{R}$ in some way. Because of the multi-agent aspect, which has not been rigorously considered here, I leave the nature of such a modification for a future discourse.

Generative uninspiration

From the point of view of the creative agent itself, that is, within the descriptive framework, only generative uninspiration can be remedied. Transformational creativity is required. To transform the set $\mathcal{T}$ in a useful way, we need to identify a valued concept, in the conceptual space constrained by $\mathcal{R}$ (otherwise, we may have aberration), and to use it to guide the transformation. However, there is a methodological problem here: there is no clear way to pick the concept automatically, except by use of an oracle. The "oracle" might in fact be systematic search of $\mathcal{R}$ (assuming this is possible in finite time), or, again, the *deus ex machina* of user intervention.

There are some interesting issues to be considered here about the dynamics of this aspect of a creative system. There are obvious possibilities in analogy with the development of creative thinking through education. These, however, are outside the scope of the current paper.

4.2 Aberration

In the case of aberration, there is a choice as to whether to view the result as acceptable or not, and therefore we have the three categories, perfect, productive, and pointless. Acceptability is determined in terms of evaluation by whatever audience the agent, **A**, is playing to. If a new concept is accepted, then a sensible solution might be to revise the notion of what the correct domain (as constrained by $\mathcal{R}$) is, so as to include the new concept. This, of course, might have consequences: other new concepts might be included and/or existing ones might be excluded along the way. If the new concept is not accepted under evaluation, then a reasonable recourse would be to adapt **A**'s technique, $\mathcal{T}$. This may have similar consequences with respect to added and existing concepts available to **A**: valued concepts may be lost, and new aberrant behaviour may be made possible.

We can now use the sets $\mathcal{A}$ and $\mathcal{V}_{\mathcal{A}}$ to generate training examples to modify $\mathcal{R}$ and $\mathcal{T}$, using our learning mechanism(s), as follows. Note that there are open questions here about some of the training sets required, since that choice is a major factor in the behaviour of the system. The main issue here is a standard one for AI: how much of what an AI program does is simply programming a computer directly to do something, and how much is emergent behaviour which was not directly programmed? In particular, if we simply train $\mathcal{T}$ to match $\mathcal{R}$ first, we might be "coaching" our creative agent too directly, instead of allowing it to develop, and, second, in doing so we might be restricting its creative capability.

Perfect aberration

In perfect aberration, all the new concepts are valued, and so should be added to $\mathcal{R}$. $\mathcal{T}$ has enlightened us as to new possibilities. We therefore attempt to revise $\mathcal{R}$, by whatever learning methods are available, in such a way that all the concepts in $\mathcal{A}$ (and $\mathcal{V}_{\mathcal{A}}$) are included, so $\mathcal{V}_{\mathcal{A}}$ is a positive training set, and the negative training set is either $\emptyset$ or $\mathcal{U} \setminus \llbracket\mathcal{R}\rrbracket(\mathcal{U}) \setminus \mathcal{A}$ or some subset of the latter, depending on the effect desired.

This, of course, is subject to the same caution as conceptual inspiration above: if $\mathcal{R}$ is a representation of an agreed domain between multiple agents, then we need agreement on changing it; the same issue arises in the definition of (any concrete) $\mathcal{E}$. Again, however, these issues are beyond the scope of the current paper.

Productive aberration

In productive aberration, we need to transform both $\mathcal{R}$ and $\mathcal{T}$, because we wish valued concepts to become accepted, and unvalued ones not to be generated. $\mathcal{V}_{\mathcal{A}}$ and $\mathcal{A} \setminus \mathcal{V}_{\mathcal{A}}$ constitute positive and negative training sets for $\mathcal{R}$, since $\mathcal{R}$ needs to expand just enough to enclude only the valued concepts in $\mathcal{A}$. $\mathcal{T}$, on the other hand, needs to be transformed to restrict its coverage: $\mathcal{A} \setminus \mathcal{V}_{\mathcal{A}}$ is a negative training set for $\mathcal{T}$, while, again, a positive training set might be $[\![\mathcal{R}]\!](\mathcal{U})$, or simply $\emptyset$.

Pointless aberration

In pointless aberration, we need to transform $\mathcal{T}$ only, so as to prevent the unvalued aberrant concepts from being generated. There is a negative training set: $\mathcal{A}$. Again, the nature of the positive training set is an open question.

5 Discussion

These labels allow us to characterise the behaviour of a given creative system and to identify broad classes of response. This, in turn, will allow comparison of behaviours both between and within the classes defined above, and thus allow better understanding of the field.

The emphasis in this work is on the further definition and understanding of the three sets, $\mathcal{R}$, $\mathcal{T}$ and $\mathcal{E}$, and their relationships to each other, to the creative domain and to the activity they are intended to describe. In any case, what does become clear when one looks in detail at these proposals is that Boden's originals were (intended to be) rather broad-brush, and that when one focusses in, the relationships between the conceptual space, evaluation, and the universe (albeit only implicit in Boden's work) become less, not more, simple.

Three clarifications do seem to emerge naturally from this discussion. First, to be interesting, $\mathcal{R}$ must define a set which is in some sense external to a given creative agent; second, $\mathcal{T}$ is the primary characterisation of the agent, and in this context, $\mathcal{R}$ is secondary (as in aberration, above); and, third, $\mathcal{E}$ needs to be independent of $\mathcal{R}$. This last needs a little elucidation, since, at first sight, it sounds like a contradiction. The point is that, for transformational creativity to occur, there needs to be aberrant behaviour (unless we allow arbitrary spontaneous behaviour from our agents, which seems inappropriate). Otherwise, unless $[\![\mathcal{R}]\!](\mathcal{U})$ is infinite, the creative behaviour will stagnate, and the system will develop no further. While this is, of course, likely to be true of AI creative systems in the foreseeable future, it would be unfortunate if one were condemned to be so for all eternity. We can explain the apparent contradiction as follows: the set $[\![\mathcal{R}]\!](\mathcal{U})$ is specific to the domain, and effectively defines it. But the set constrained by $\mathcal{E}$ need be only the extension in $\mathcal{U}$ of *those properties of* $[\![\mathcal{R}]\!](\mathcal{U})$ *which are valued.* Thus, $[\![\mathcal{E}]\!](\mathcal{U})$ could be very large, but only a small part of it might be explored, due to the restrictions in $\mathcal{R}$ and $\mathcal{T}$.

The issue of multi-agent creative systems is becoming increasingly important, in the current line of reasoning. The aim of Boden's and Wiggins' frameworks is to describe the behaviour of creative systems, but no natural creative systems exist in isolation (and, indeed, one might argue that neither do artificial ones). Therefore, the generalisation of these ideas, which has been informally mentioned above on several occasions, to multi-agent systems seems crucial and urgent. Only in this context will the distinctions highlighted above become really clear, as the shared and individual content in the system will need to be made explicit.

6 Summary and Conclusion

This paper has presented a very small step on the road to a more precise understanding of creative systems. I have presented six categorisations of creative behaviour, which can be identified directly from the behaviour of creative systems as described using Wiggins' formalism, and suggested how the needs of each category of system can be met, from within or from outside the system itself. This raises many questions, not least the issue of interaction between multiple creative agents. This question will be addressed in future work.

Acknowledgments

I am very grateful to my colleagues in the Centre for Computational Creativity at City University, London, for on-going discussion about this and other work. Graeme Ritchie supplied some particularly helpful comments on earlier related work, during his sabbatical at City in 2002. Three anonymous reviewers gave some very helpful feedback.

References

[Boden, 1990] M. Boden. *The Creative Mind.* Abacus, 1990.

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

[Ebcioğlu, 1988] K. Ebcioğlu. An expert system for harmonizing four-part chorales. *Computer Music Journal*, 12, 1988.

[Gervás, 2002a] P. Gervás. Exploring quantitative evaluations of the creativity of automatic poets. In C. Bento, A. Cardoso, and G. A. Wiggins, editors, *Proceedings of the ECAI'02 Workshop on Creative Systems: Approaches to Creativity in AI and Cognitive Science*, pages 39–46, Lyon, France, 2002. ECAI'02.

[Gervás, 2002b] P. Gervás. Linguistic creativity at different levels of decision in sentence production. In A. Cardoso and G. A. Wiggins, editors, *Proceedings of the AISB'02 Symposium on AI and Creativity in Arts and Science*, pages 79–88, `www.aisb.org.uk`, 2002. AISB.

[Wiggins, 2001] G. A. Wiggins. Towards a more precise characterisation of creativity in AI. In C. Bento and A. Cardoso, editors, *Proceedings of the ICCBR2001 workshop on Creative Systems*, Naval Research Laboratory, Code 5510, Washington, DC., 2001. Navy Center for Applied Research in Artificial Intelligence.