The transformational creativity hypothesis

Graeme RITCHIE

Department of Computing Science
University of Aberdeen
Aberdeen AB24 3UE
United Kingdom

gritchie@csd.abdn.ac.uk

Received 1 February 2006

Abstract The work of Boden on the nature of creativity has been extremely influential, particularly the hypothesis that the highest form of creativity results from *transformation of a conceptual space*. We consider how these ideas could be made more precise, and hence become amenable to empirical testing. This requires some reconsideration of foundational assumptions about computational creativity. We set down the abstract requirements for a conceptual space, review some possible types of formal model, and discuss how it might be possible experimentally to falsify (or corroborate) this hypothesis. We conclude that the central terms (conceptual space, transformation) are still too vaguely defined to support falsifiable claims, but that this is not an obstacle to writing creative computer programs.

Keywords Transformational Creativity, Conceptual Space, Exploratory Creativity, Computational Creativity.

§1 Introduction

For decades, there have been attempts within artificial intelligence to build programs which "create" artistic or scientific objects, such as representational paintings[5, pp. 135-153], music[2, 26], mathematical concepts [19, 10], stories[24, 39], jokes[3], or poems[14, 15, 16, 22, 21]. That strand of research (which continues today) focusses on methods which are applicable to artefact-creation within some particular domain (music, stories, etc.). However, in recent years, interest has grown in the

development of theoretical models or methodological frameworks for studying computational creativity in general[33, 40, 41, 30]. The questions which that work explores – such as 'can computer programs be creative?', 'how would we decide whether a computer program has been creative?', and 'what kind of computational mechanisms underly creativity?' – might previously have been regarded as philosophical, but are examined from the perspective of AI methodology.

Much of that research alludes to, or is influenced by, the widely discussed proposal by Boden[5, 6] that high levels of creativity result from the *transformation of a conceptual space*. That is, Boden offers an answer to the theoretical question 'what kinds of computational mechanism lead to genuine creativity?' Despite the large amount of attention paid to Boden's ideas, these ideas are still largely informal and unformalised (although Wiggins[40] contributes some suggestions). Moreover, if the claim about transformational creativity is to be tested empirically, some more detail is essential.

Our aim here is to contribute to this line of formal research, by considering how Boden's proposals could be framed formally, and hence how they could be tested empirically. We start by setting out our general premises (Section 2), then provide an informal summary of what we regard as the situation to be modelled (Section 3). Section 4 discusses what, in general, might constitute a space in Boden's sense, and in Section 5 we consider some possible types of formal model. After discussing various aspects of the issue in Section 6, we set out (Section 7) a possible route to testing the transformational hypothesis.

§2 Basic aims and assumptions

2.1 What do we mean by "creative"?

The existence of the words '*creative*' and '*creativity*' suggests that in society at large there is a general notion of 'being creative'. However, this does not mean that there is a single, coherent, consistent, precisely definable notion of 'creative' which is amenable to scientific scrutiny. The use of these words in ordinary discourse may be highly vague, very unsystematic and completely inconsistent[7]. It is plain, nevertheless, from the wealth of academic writing on creativity that there is a widespread belief, or perhaps hypothesis, amongst philosophers, psychologists, and others, that such a concept can be defined. Here, we do not assume a particular definition of creativity. Instead, we shall consider mechanisms which might lead to "creative" computations, in the fol-

lowing sense. We start from a loose notion of a *potentially creative program*. Such a program produces output (what we shall call *artefacts*) such that:

- the artefacts are intended to be, or to directly represent, objects or concepts (e.g. melodies, poems, pictures) which, if produced by a human, might – if good enough – be classed as demonstrating creativity (cf. Minsky's characterisation[25, p. v)] of artificial intelligence as 'the science of making machines do things that would require intelligence if done by man');
- the class to which the artefacts are intended to belong may be very large, and often ill-defined;
- the extent to which an artefact falls within the intended class (e.g. being a poem) may be a matter of subjective (human) judgement (and a computer-produced artefact may be judged as not being in the target class);
- there is some property of *quality* (loosely, how good or bad the item is) which an artefact may possess to a varying degree, and which is primarily determined by human judgement.

This excludes the behaviour of many familiar kinds of computer program: numerical calculators, syntactic parsers, database retrieval modules, etc.

2.2 Creativity and formality

Within artificial intelligence there is a recurring claim (or implicit assumption) that creativity can be defined in terms of some formal characteristics of the process involved, and that this definition would be applicable to computational models.[5)28)31)32)37)7)40)41)] An exception is Bundy[8, p.534)]: 'Creativity. . . does not correspond to some well-defined family of computational processes.' Although we share Bundy's doubts about this position, we shall accept it as a working assumption for this paper, since we are examining ways in which a particular variant of this claim (Boden's transformational hypothesis) could be tested.

It is desirable to be as precise and formal as possible about all aspects of creativity – both models of the creative process and methodologies for scrutinising computational creativity. This follows from general scientific principles, regardless of any assumptions about creative processes sharing a formal essence. Formality and precision will show up possible inconsistencies, will suggest subproblems to be considered, and will enable hypotheses to be subjected to empirical testing. Although Boden's ideas have been provocative and

influential, they are still highly informal and imprecise. We shall not be able to remedy this position entirely, but we shall make some steps in that direction.

2.3 Human creativity as the origin

If claims such as 'this program has been creative' are to have a comprehensible and plausible meaning, then our definition of 'creative' has to reflect, to a large extent, the meaning of the ordinary language term. We cannot be arbitrary or circular. If, for example, we were to define 'creative' to mean 'computationally efficient', then testing of machine creativity would be much more straightforward, but we could be accused of not addressing the real question. The precise formal definition for scientific use should have a significant amount in common with the ordinary usage, despite the messiness of the latter.

Moreover, we should be guided by the way that the word 'creative' is ordinarily used when talking of non-machine (human) creativity. This is for two reasons: firstly, that is the original, established usage; secondly, to rely on instances of machine creativity (the problem we wish to analyse) would risk circularity in claims about the nature of that process. We shall therefore, when discussing computational creativity, allude to instances of human creativity (as do most writers on creativity and AI, including – copiously – Boden).

This assumption (of being guided by how we appraise human creativity) is probably tacitly adopted by most research in this area, but, when followed up thoroughly, will take us to a position slightly different from some analyses which are more focussed on details of process[40, 30].

2.4 Observable empirical factors only

In human creative activities, there are certain aspects which are knowable, such as the attributes of the artefact created, the other comparable artefacts in existence, possibly the other artefacts the creating individual was aware of. What we usually do not know is the mental or emotional processes by which the individual produced the artefact (although we may know other aspects of the action, such as the time taken). Hence, it is routine to make judgements of creativity (in humans) on the basis of what is known, often focussing on the attributes of the artefact(s). If our formal account of creativity, for analysis of computer systems, is to mimic our judgements of humans, then it too should be based only on comparably observable factors, without adding extra information about the internal workings of the computer program. This may be our

most contentious working assumption, as some would argue that the inner workings of a computer program are critical in deciding its creativity; in particular, Boden[5, pp.39-40] advocates just such consideration of the underlying process (for both humans and computers). We suggest that this would move away from the way human creativity is normally judged.

There is a fine but important distinction between the production of the artefact, and the devising of the production-method which gave rise to the artefact. If we happen to be aware of the method, we can treat the method itself as an abstract artefact, created by whoever (or whatever) devised the production-method. That would then permit a judgement about whether the production-method, as an artefact, displayed evidence of having been devised in a creative manner. In this way, we still predicate creativity only on the basis of observable data, but widen the scope of our data when circumstances permit it.

Boden discusses the very general question 'can computer programs be creative?' (which she labels as 'Lovelace Question' 2, or perhaps 4), and, as noted earlier, tries to answer the slightly more particular question 'which computational mechanisms lead to creativity?' (not a 'Lovelace Question'). Neither question can be answered empirically unless we have a way to answer the more specific question 'has this computer program been creative (on this occasion)?'

All of these questions are usually treated, within the AI literature, as empirical questions which can be falsified or corroborated by building programs which generate artefacts (physical or abstract), and then studying these programs. However, this appearance of being empirical is illusory unless we can define what we mean by 'creative/creativity', and define it in terms of factors which are (at least in principle) observable (for example, Ritchie[33] has proposed some such factors). Moreover, our definition(s) should describe what behaviour we would regard as creative without building in, prematurely, proposals about *how* that behaviour might be achieved. If we can maintain a separation between our observational vocabulary and our theoretical models of possible mechanisms, then we can, without circularity, treat questions such as 'which computational mechanisms lead to creativity?' as empirical issues. If we incorporate our hunches about the best way to achieve creativity into our definitions of what observable behaviour constitutes creativity, then we have, to a large extent, undermined the empirical nature of the investigation.

2.5 The primacy of human judgements

It follows from the arguments above that this line of AI research is at least partly trying to model human judgements about creativity. That is, writers such as Boden base their intuitions and proposals about creativity on examples (from art or science) which are deemed to be creative (and occasionally on examples which are not creative). Potential notions of creativity are implicitly assessed according to how well they match this (human-judged) data on what is or is not creative. (There is a loose analogy here with linguists aiming to construct language models which account for human intuitions about grammaticality or meaning.) It follows that if we have a hypothesis of the form 'Mechanism X leads to creativity', then the ultimate test of the creativity of the action of Mechanism X must be human verdicts (of creativity).

§3 An informal overview of creative activity

We start by summarising briefly (and informally) what we believe to be a reasonable approximation to the notion of creative activity, simplified in ways which should be conducive to subsequent formalisation. This should give some idea of the level of abstraction that we are adopting. Notice that the next few paragraphs set out a set of *simplifying assumptions*, not empirical claims; the use of blunt declarative statements should not imply that these are assertions of fact. However, making them explicit allows others to challenge their plausibility, particularly as a basis for theoretical arguments. Also, these statements are intended as a summary of the phenomena to be accounted for, not a theory of how these phenomena come about. All these assumptions are generally derived from consideration of what goes on in human creative activities.

We shall regard a (potentially) creative action as resulting in a specific item, an *artefact*, although this need not be a concrete object (Wiggins[40] uses the term *concept* in a comparable role). The action takes place within the context of a society, in which there are various individuals. The judgement that an artefact manifests creativity, or that an action constitutes creativity, is made by individual(s) within the society. In particular, each individual may have idiosyncratic opinions or criteria about the type of artefact being produced, so these judgements are always relative to the one making the judgement.

Within the society, there is a finite (and small) set of what we will call *medium types* and also a finite and small set of *genres*. An artefact belongs to a medium type; that is, a medium type is a class of artefact, and that class is determined by the basic form of the artefact – whether it is a sequence of words

and punctuation, or a two-dimensional array of coloured pixels, or a sequence of musical notes, etc. These medium types embody as few claims or assumptions about the higher-level analyses that might be imposed upon the data – they do not represent the metre of the poem, the possible scenes depicted by a painting, or the key of a melody. Nor does belonging to a medium type imply that an entity is a member of any recognisable type of cultural artefact, such as a poem or melody – the membership merely indicates the raw data type of the medium, such as being a finite sequence of words and punctuation marks. Whether or not an item is a member of a medium type is trivially decidable, as this is inherent in its representation. Several different sorts of artefact (viewed culturally) might use the same basic medium type; for example, finite sequences of words and punctuation might be the representation used for poems, for jokes, for stories (or more accurately, for items which could then be judged as possibly belonging to one of these culturally-defined classes).

A genre is a culturally-defined type of artefact, either very broad or quite narrow (e.g. an impressionist painting, a symphony, a story). It will have an associated medium type, indicating the basic data which represents the artefact. Artefacts do not belong absolutely or objectively to genres: instead, each individual can make a judgement about the extent to which an artefact conforms to the norms of a genre; for example, not all sequences of words count as poems. We distinguish between the genre, which is shared (in some sense) amongst the members of the society, and an individual's assessment of a particular artefact with respect to that genre. Individuals also make judgements about the *quality* of an artefact, and these judgements are relative to some particular genre (a text may be a poor story but an excellent poem).

The word '*creative*' is sometimes applied to a person, sometimes to an action, sometimes to an artefact; Boden[4, p. 170)] refers to a *thought* as being creative. Here, driven by our desire to consider only observable data, we shall standardise to regarding it as a property of an artefact relative to a set of other artefacts of the same medium type (and usually of the same genre). That is, a judgement would be of the general form 'artefact A displays creativity relative to artefacts $\{A_1, \ldots, A_n\}$'. The choice of which artefacts are relevant in context for this comparison is not simple. For a human-created artefact, it might be the other exemplars that the creator was already familiar with, although, in actual situations, human judges might well make comparisons with exemplars that they themselves are familiar with, overlooking Boden's distinction between

P-creative (based on novelty for the creator) and *H-creative* (based on novelty with respect to history). For a computer program, the comparison set might be some corpus of examples available to the program designer[33] or some knowledge base used for case-based generation [32].

§4 Spaces and transformation

4.1 Boden's position

Central to Boden's approach is the notion of *conceptual space*, as the abstract location of the entities (our 'artefacts') produced by creative acts. She does not define this term precisely, although she asserts the need for it to be elaborated[5, p. 73] Subsequent debate in this area has speculated on what this notion might mean. For example, could it be the traditional *search space* of AI problem-solving[31, 40]? Boden[5, p. 77] says a search space is 'one example... of a conceptual space'. (See discussion in Section 6.4).

Boden argues for the importance of *changes* to the space causing, or even constituting, creative acts, and the first mention of conceptual space in her book is: '...changing the existing rules to create a new conceptual space'[5, p. 46]; that is, she does not first establish the notion of space and then consider changes to spaces, but takes the idea of conceptual space as given or self-explanatory. In the terminology we are using here, it appears that where we have a "genre" (a class of artefacts loosely defined by cultural norms), Boden might posit a conceptual space which defines the limits and the internal layout of that genre. If so, one difference would be that our "genres" are relatively simple labels used (by humans) to classify artefacts (e.g. "poem", "story" "melody"), while Boden's conceptual spaces are an attempt to explain the underlying mechanisms and the relationships between items within a space. Our term "genre" is an informal term intended to refer (as in Section 3) to the items under investigation; it is not intended to describe or explain any underlying structures or mechanisms.

Boden says that, although working within an existing space may produce interesting results (*exploratory creativity*), a higher form of creativity can result from making changes to the space: *transformational creativity*. This concept, and the claim that it can lead to the highest form of creativity, has been very influential. (For example, Wiggins' model[40] is designed to define a transformational mechanism.)

Boden is adamant that 'transformational creativity' involves radical

changes to the space, not minor adjustments, and implies that this is a qualitative difference ("transformation" versus "tweaking"), rather than merely a matter of the degree of change. When we come to look at classes of models (Section 5), we shall consider whether one could characterise such a distinction formally.

Our aim is to examine the hypothesis that transformational creativity allows a higher form of creativity than exploratory creativity. As will become clearer, it is not feasible, in the present state of the field, to set out actual evidence for or against this claim. Instead, we will discuss how to gather relevant evidence.

It is not entirely clear whether Boden regards her statements about the superiority of transformational creativity as a *hypothesis about*, rather than a *definition of*, what counts as creative. This blurring continues in a review[38)] of Boden's book, which refers to the idea as a 'thesis' and a 'definition' in adjacent sentences. Here, we adopt the interpretation that it is a hypothesis, but we shall later return briefly to the possibility that it is a definition (Section 7.5).

4.2 Conceptual spaces and the set of possibities

If transformation is to be possible, the conceptual space must be embedded within some wider or more general system of possibilities, otherwise change could consist only of reduction in the set of artefacts in the space. This wider, more general space could simply be the medium type (i.e. anything that can be represented using the basic data type) or there might – in the case of certain genres – be some intermediate definition of *well-formed and logically possible item.* This would give a hierarchy of inclusion:

conceptual space (typical items as currently defined) $\subseteq$
well-formed and logically possible items $\subseteq$
all items in the medium type

Transformation could then extend the conceptual space out into the set of logically possible items.

In genres where there is a clear definition of a set of logically possible items, the question arises of whether this set is identical to the conceptual space or properly includes it. Suppose the set of chess games consists of all valid move-sequences, then anything which transcended that definition would not be a game of chess – that all-encompassing set of games is the "logically possible" set. There are then two possible positions:

(i) There is a set of "typical games", a proper subset of the logically possible games, which constitutes the conceptual space, and this space is available for transformation within the larger logically possible set.

(ii) The logically possible set of games is exactly the conceptual space, so no valid game of chess can display transformational creativity – the only way that a chess player can be transformationally creative is to invent new rules (cf. the widely accepted story that rugby football was conceived when a player of association football picked up the ball and ran with it).

Similarly, if we take the definition of a haiku to be something along the lines of *any syntactically valid sequence of words segmented into three lines of 5, 7 and 5 syllables*, this would be the "logically possible" set; if it were also to be the conceptual space for the haiku genre, this would rule out the possibility of transformationally creative production of haikus (only devising a new poetic form would constitute transformational creativity).

If we identify the logically possible set with the conceptual space, but still admit the possibility of transformation, then transformation is limited only by the medium type. (If there were to be some more limited set of possibilities for transformations, this would raise the question: why was this not the original "logically possible" space?)

We conclude from this that although some (but not all) genres may come with a ready-made definition of a logically possible set of objects (a proper subset of the medium type), this space cannot simultaneously be the conceptual space and the space within which transformations may occur. We must either exclude transformational creativity within this genre, or posit a conceptual space which is a proper subset of the logically possible set.

Hence, in the cases of interest (i.e. where transformational creativity may occur), there will not usually be a prior, well-defined, explicit definition of *the space which is available to be transformed*.

4.3 Transformation as interpretation?

The 'transformation of spaces' is very much an *analysis* that is posited by Boden (and others) of the implications of particular artefacts – it is not part of the raw data. An art scholar might characterise an early Cubist painting as "transforming the space", but what the artist has actually done is produce a painting. The spaces are not given to us (see discussion in Section 4.2), nor

is the transformation. If we are to have a formal description of creative space-transformation, we have to show how such an analysis can be derived from the actual artefacts. More cautiously, we should formalise the situations where a transformational analysis is *possible*, so that we can then weigh up the transformational account against alternative descriptions of what is going on.

Given our assumptions, the scenario we are interested in analysing is, informally, "artefact A causes individual P to adopt a different space"; this could be either by creating a new genre, or by restructuring an existing genre. That is, we are trying to elucidate the relationships between several things: an artefact, an individual, and two spaces (before and after); the genre should perhaps be included.

The artefact must be sufficiently similar to previous artefacts for it to be relatively clear which norms or spaces are relevant. This may mean that it belongs at least in some peripheral way to an existing space (for a known genre).

However, this supposes that the assessing individual seeks a revised space. If not, then no transformation occurs, and the artefact's ratings remain the same. This touches on issues within the psychology and sociology of art and science: when does the individual feel the need to transform? That is, the criteria for transforming a space are not solely formal. The current paper, however, considers only the formal aspects.

4.4 What is a space?

Let us consider, more abstractly, the requirements for a "conceptual space" in Boden's sense.

We start with a minor terminological point. In one sense, the space is just a set of artefacts. Although it may sometimes be sufficient to consider this simple perspective, we shall more often want to consider how the space is *characterised*; that is, what finite rules or structures indicate how artefacts fall within the space. Only when there is some *intensional* characterisation independent of the actual artefacts (i.e. something other than a simple extensional list of known artefacts) can we consider issues such as whether a new artefact does or does not fall within the space, or how a space may or may not be altered. A set of artefacts may be infinite, yet have a finite (and computationally malleable) definition. We shall use the terms *artefact-set* for the actual set, and *space-definition* for the more compact definition of possible artefacts; where the sense is obvious, we may just use the term 'space' for either of these.

Arguably, we have to consider *two* distinct spaces: that imposed by judgements of the extent to which an artefact conforms to a genre (what has been called[33] *typicality*), and a further space induced by judgements of the *quality* of the artefact. We shall return to this in Section 6.2.

In Boden's discussions, it seems that the 'conceptual space' is either an informal metaphorical construct, or is taken to be whatever the computer program uses to structure its computations. The problem with using the latter (leaving aside concerns expressed earlier about inspecting the workings of a program) is that all outputs would, by definition, be within the program's space. This would make it logically impossible for a program to produce output which transcended its own space. Hence, programs could never carry out anything beyond 'exploratory' creativity (a consequence with which some sceptics might agree!). We will therefore, in the discussions below, not make the limiting assumption that the conceptual space is to be identified with the abilities of a given program. Instead, we shall consider more abstract "spaces", which might be associated with the genre in general rather than with a single program.

There are five crucial functions that a space must fulfil if it is to support this analysis of creativity.

Membership. It must be possible to determine the membership of an artefact with respect to a space-definition. This membership may be a binary decision, or graded in some way. Alternatively, membership may be better represented as the positioning of the artefact relative to other artefacts in various ways (for example, along multiple dimensions). However, there must be some notion of "membership" or "positioning within the space", and this must be decidable: some parts of the definition may rely on vague or subjective terms, but the computational structure of the definition must not be circular or otherwise flawed.

Similarity. Discussions of creativity, both informal and formal, usually involve (sometimes tacitly) some notion of "similarity" between artefacts. If all the artefacts of a given genre were completely incomparable, every distinct artefact would, by definition, be 100% novel. The very idea of new but unnovel artefacts assumes some form of similarity. It is conceivable that the metric of similarity could be formally unrelated to the space-definition, but this seems unnatural and unlikely. A more elegant approach would be to have a definitional apparatus which directly led to some distribution of the artefacts within the space, with some means of comparison. As with membership, this comparison could result in a score of some kind, or could be a qualititative statement of a set of

differences (e.g. along various dimensions). In this way, the attributes of an artefact that affect membership would also affect similarity. Also, any useful notion of similarity must rely on some higher-level representation of the data than the rudimentary data types we used to distinguish medium types. A distance metric which simply compared word-strings, or pixel arrays, would miss the kinds of similarity that are important for questions of creativity. For example, a pixel-by-pixel comparison of a portrait and a landscape by the same painter, or even two landscapes, would probably be classed as very distant (dissimilar), whereas two portraits in different styles might be classed as similar.

Determining the space. As noted earlier, all that is available to the theoretician or analyst (of the creative genre under consideration) is a set of artefacts. The space-definition is not usually given, but must be induced on the basis of the evidence (the artefacts). As argued in Section 4.2, there may be genres where there is a readily available definition of "logically possible artefact", but the space amenable to transformation – the currently typical artefacts of the genre – is unlikely to be explicitly and formally defined. That discussion also suggested that we might need to abstract two different "space definitions": the currently typical set of artefacts for the genre, and the range of possibilities into which transformation might change the current space, unless the latter is simply the medium type. This is a significant task, even if the analyst narrows the search by opting for one particular type of formal model. In discussions of the creativity of computer programs, the step of figuring out what the relevant space is for a given output set is rarely if ever considered, even though this is logically prior to notions of changing a space. Whatever a space is, formally, it must be something which can be abstracted from a set of artefacts. Although in real-life cases of human creativity, this step may be performed intuitively by people, here we are aiming for a formal computational model, so we must have a decidable mechanism. In essence, this constitutes machine learning.

Exploration. Boden's *exploratory creativity* consists of following an organised path through the artefact-set. Most supposedly creative programs can be viewed as doing this, which is unsurprising, as the standard way (perhaps the only manageable way) to construct any generating program is to have a well-defined set of possibilities and move through them systematically. The formal definition of conceptual space must allow this. (It might be hard to contrive a definition which did not.) Once again, the basic level of the medium type (word string, pixel array, etc.) offers an uninteresting way to organise exploration. Instead,

exploration should be channelled by whatever space is postulated for the set of artefacts. (Inspection of discussions in the literature on computational creativity suggests that adopting this more relevant kind of space has been an implicit assumption of most authors, but it is helpful to make the point explicit.)

Change. This is the central topic for discussion here: how can a space-definition be altered ("transformed"), particularly in response to a new artefact? More precisely, how can the space-definition be altered in a way that will have suitable changes for the associated notions of membership, similarity, and exploration?

These seem to be the most critical desiderata for a conceptual space. The basic operations that can be performed on a space S, corresponding to the five aspects listed above, therefore appear to be:

(i) **Locate artefact A within S** (this could be a numerical rating of membership, or could be a more complex rating where the space-definition ascribes attribute-values, or positions on dimensions, to an artefact).

(ii) **Rate artefacts A, B for similarity (w.r.t. S)** (where the space-definition leads to degrees of similarity)

(iii) **Induce a space-definition from artefacts $A_1, \ldots, A_n$**

(iv) **Given (existing) artefacts $A_1, \ldots, A_n$, generate a (new) artefact A** (where A is in some suitable sense "within" the space)

(v) **From S, create a revised space-definition S' to include artefact A.**

This set of operations is not the only possible set. Rather than ascribing to the audience (interpreter) an ability to transform a space S to a space S', we could instead postulate an ability to determine whether two spaces were transformationally related; i.e. replace the last operation with:

(v′) **Given two space-definitions S_1, S_2, determine whether S_2 is a transformed version of S_1.**

Then, when confronted with a new artefact A, the audience would first group this with some known artefacts $A_1, \ldots, A_n$ of that genre, induce a suitable space S' for $A, A_1, \ldots, A_n$, and then see whether S' is a transformation of the space S previously attributed to the genre. As an account of an individual's reaction to a cultural artefact, this is at least as plausible as having the interpreter (audience) transforming the space directly.

From this it can be seen that the requirements for a conceptual space are very underdetermined by the writings of Boden and others.

§5 Some possible formal structures

The world of formal models contains a wide variety of structures that could be used to construct space-definitions. We shall consider a few of these here, and comment upon their suitability for supporting the operations that spaces can undergo (particularly change).

5.1 State-Transition/Derivation models

This category covers, abstractly, both formal rewrite grammars and traditional automata[17], and heuristic search programs [27]: various symbolic rules define possible choices, and there is a notion of a *derivation* – a combination of, or sequence through, the rules – which defines the valid items. Whether this is what Boden[5] means by a *generative rule* approach is hard to say.

In such a model, membership is a binary decision, depending on whether the artefact is the result of a derivation. Similarity is not a natural feature of such a system, but some measure could perhaps be contrived based on the rules used within a derivation. (For example, rules could be treated as having varying weight, or being at different levels in a hierarchy of importance, so that the degree of similarity between two derivations would depend on the weight or level associated with those choices which differ between the two derivations; something of this sort is used in the General Theory of Verbal Humour[1].)

Systematic exploration (indeed, exhaustive generation) of possibilities is straightforward in such a model, and there are natural points where additional domain-specific heuristic information could be injected: in the choice of rule for each step in the derivation.

Alterations to a space-definition could be made in various ways, but generally adding a construct (e.g. a transition or expansion rule) would lead to an *expanded* artefact-set, and removing a construct would *shrink* the artefact-set.

5.2 Prototypes

There is a great deal of work within cognitive psychology on categorisation and concept formation, which discusses at length the notions of class membership, similarity between concepts, and typicality[9, 23]. This line of research is typified by the highly influential ideas of Rosch[36, 35] on the way that humans categorise objects in everyday life, which has led to *prototype theory*. The core idea is that some objects are clearer, or more typical, exemplars of a particular category or class. For example, for the category *bird*, a robin is a

highly typical example, but an ostrich or a penguin would be less typical. Thus membership is not a simple binary distinction, but may be graded, depending on how close an item is to the class's *prototype*: an object which embodies the most typical features of the category. This raises another important aspect of prototype theory: the idea of *distance*, or its converse *similarity*, between two entities. Osherson and Smith[29] observe that there is no single agreed account of prototype theory, but offer a formalisation which they say captures the essence of most versions. In their definition, a concept (or category) is represented as a quadruple $\langle A, d, p, c \rangle$ where A is a set of objects (the *domain*), d is a distance metric on A, p is a particular member of A (the prototype), and c is a function $A \rightarrow [0, 1]$, the *characteristic (membership) function* of the concept.

If we adopt Osherson and Smith's formalisation as the typical prototype theory, then some of our requirements (Section 4.4) are met immediately: membership (graded) and similarity are provided directly. No obvious structure is supplied for exploring the set of exemplars of a concept, but solutions could perhaps be devised using the similarity measure (e.g. working outwards from the concept's prototype). Transformation is not defined. It is possible to imagine changes to any of p, c or d, or perhaps even to A, but there is no obvious distinction between "transformational" and "non-transformational" change. If the transformation were to involve adding further elements to A, then there would have to be some larger superset U of which A was a subset (see Section 4.2 above). Although this superset could quite naturally be the set which we have called the *medium type* – finite word sequences, arrays of pixels, etc. – its existence is a minor adjustment to the basic prototype formalism.

5.3 Multiple dimensions

Perhaps the most intuitively natural structure for a conceptual space, judging by writings on this topic, is a set of *dimensions*, where an artefact can lie at specific (ordered) points on each dimension.

If each artefact can be allocated values for the dimensions, then membership of a space can be defined by specifying some subspace of the full multi-dimensional space. A further possibility would be to have a *weight* associated with each dimension (intuitively, reflecting the importance of that factor)[33]. This would map each basic n-tuple to another (n-dimensional) vector, but in a space with a different distribution of artefacts. In this case, membership could also be computed as a numerical value by aggregating the weighted vector components.

(Formally, such arrangements, and variations of them, have been explored in multi-attribute decision theory [18]).

Of various possible similarity measures, an obvious one would be the Euclidean distance between the two vectors, either in the original n dimensions, or in the weighted space.

The space would set bounds on exploration, but would not offer any particular structure for carrying it out, beyond exhaustively trying every valid value on every dimension.

Change could consist of revising one or more of the criteria for allocating values (on dimensions) to artefacts. This could be thought of as altering the assessment (perhaps unconscious) of the various attributes of the artefact by an individual audience. (Whether this describes a perceptual or a conceptual change is unclear.) For example, it could be argued that the rise of impressionist painting involved a change in the public's notion of a "realistic" painting, or that society's idea of "obscenity" has varied over the centuries. Such a change, of any size, leaves the dimensionality, and arguably the dimensions, of the space intact, but could change the location of artefacts within the space, so that some previously highly typical items might become peripheral, or vice versa.

If weights are associated with dimensions, then these could change, altering the importance assigned to each property. This is a plausible match to the notion of changes in "taste" or "fashion". For example, acceptance of abstract paintings does not necessarily involve a change in what counts as "realistic", but could instead be a change in the importance attached to this property. Again, changes could be small or great, and would not alter the dimensions of the possible space, but would in general change the distribution of artefacts through the space, perhaps quite dramatically, and might or might not count as "transformation".

A less straightforward change might be to add dimensions to the space. (Removing a dimension is also a logical possibility, but the effects would be indistinguishable from assigning zero weight to the dimension.) While this may sound more radical than the previous two types of change, whether it makes a real difference depends on how the assessment function(s) associated with the new dimension(s) rate the relevant artefacts, and how much weight is attached to these factors. New dimensions could be added without having much impact on the overall landscape of the genre (in terms of the abstractions of Section 4.4 above), if they were allocated very low weights.

Gärdenfors, in work originally published[11] around the time of Boden's book[5], apparently completely independently, proposes a knowledge representation scheme which he calls *conceptual spaces*. This is a detailed and sophisticated form of multi-dimensional model, which Gärdenfors offers as a model of human categorisation. He emphasises its relevance to empirical findings such as those of Rosch (Section 5.2, above) and shows how prototypes can be defined as regions within his conceptual spaces [13, 12], with "natural concepts" corresponding to convex regions. This means that there is a formal difference between Boden and Gärdenfors as to which type of entity is labelled as a "conceptual space". For Boden, it seems that an individual culturally defined class (e.g. poems, jokes, symphonies) would correspond to a conceptual space; for Gärdenfors, a conceptual space is the whole set of multiple dimensions, within which particular *concepts* can be defined as different sub-regions.

5.4 Constraints

Boden mentions changes in constraints as a possible kind of space-transformation, so we should consider a constraint-based specification, as in certain problem-solving representations[20]. Although constraint-solving is an elegant and efficient mechanism for determining the values for a related set of variables, it still leaves open the structures which the variable values characterise. A constraint-solver could be imposed upon a basic representation which is declaratively defined in some other way (e.g. by generative rules). In a multi-dimensional formalisation (see above), the subspaces of interest could be stated by imposing constraints upon values of coefficients. Indeed, if the variable domains are ordered sets (more especially, if they are numerical), then the constraint-based model is inherently multi-dimensional. Only if the variable domains do not lend themselves to interpretation as dimensions is it formally distinct.

Conventionally, any assignment of values to the variables which satisfies all the constraints counts as a solution. Here, we could regard the constraints as specifying the artefact-set. To have a looser notion of the artefact-set (i.e. not simply an all-or-nothing decision), we could allow value-assignments which merely satisfy *most* of the constraints to count as members of the genre. As in the multi-dimensional case, weights could be attached to constraints, allowing a combined rating of how well a value-assignment meets the constraints.

There is no obvious, natural definition of similarity. In the version

where not all the constraints have to be satisfied, then perhaps a measure could be based on a comparison of the set of constraints which the two artefacts satisfy.

The constraints may license many combinations of values for variables, so exploration would consist of enumerating these in some order. Some constraint models also contain *preferences* (or *soft constraints*), which do not make rigid stipulations (which would eliminate some possible values) but instead indicate an ordering on possible values, making some more "preferred". These could be used to impose order upon the enumeration of value combinations.

Change could occur by the addition or removal of constraints. Unlike the state-transition/derivation models, where addition of components expands the artefact-set (or leaves it untouched), here addition *shrinks* the artefact-set. Conversely, removing constraints can only *widen* (or leave unchanged) the artefact-set. Boden's informal discussions give the impression that constraint removal is the route to creativity. However, some forms of artistic innovation can be seen more naturally as the imposition of further constraints: the *dogme* film movement, or pointillist painting, or (according to Buchanan[7]) haiku.

5.5 Connectionist networks

There is not univeral agreement on the usefulness of connectionist models in creativity:

> To peg the definition [of creativity] on a nonconnectionist model is to hitch it to a fading star. [28, p.547]

> ...it is somewhat futile to look to connectionism for useful insights about creative insight. [37, p. 137]

Nevertheless, connectionist models typically afford very natural notions of the facilities needed.

A network can very easily represent degrees of membership and can show degrees of similarity between inputs. Exploration is perhaps less obvious, but could be organised in a generate-and-test manner.

It is not clear what the natural notion of change would be for a network representation. Changes in weights are easily accommodated, being the main currency of network computation, and could reflect space changes of both small and large magnitudes. Various kinds of alterations to the internal topology are also possible. In both cases, the effects on the space (as embodied in the membership and similarity judgements) would be indirect and probably hard

to predict, and there would be no clear intuitive mapping from changed formal entities to manifested properties in the artefacts (e.g. increasing the importance attached to "realism").

Boden[5] writes approvingly of connectionist models, but does not explain how they fit into the transformational perspective. She mentions, briefly, that a hybrid (connectionist/symbolic) model might be needed.

§6 Discussion

6.1 Transformation vs. tweaking

In all the types of formal model reviewed above, there is no obvious formal distinction between minor and major changes. Any of the adjustments suggested could vary in their extent. A highly novel artefact will manifest a high degree of difference from previous artefacts, but this could arise within the same formal space, for example by a change in the weights attached to the components of an multi-dimensional space.

6.2 The role of quality

As mentioned in Sections 3 and 4.4, an important part of judging creativity is an assessment of the "quality" of the artefacts. In a formal description, there are various ways in which quality might be structured, and in which typicality and quality might be interlinked. Whereas the discussion of typicality (Section 5) considered binary membership (yes-no), graded degrees of membership (a fuzzy set) and some form of structured allocation of a "position" within the space, in the case of quality only the latter two would be plausible – a yes-no decision on quality is unrealistic, even for this simplified discussion.

The simplest approach might be to assume that quality is stated in terms of whatever components make up the underlying conceptual space defining the genre. This would mean that an attribute of an artefact is potentially relevant to determining the quality of that artefact only if it is relevant to deciding membership/location within a genre – no other attributes can be considered. This does not mean that quality values cannot be assigned to items which, although outside the subspace which counts as typical of the genre, are nevertheless within the formal (available) space. Discussions of creativity, both informal (Boden) and formal (Wiggins, Ritchie, Pease et al.) assume that highly atypical artefacts can be assigned quality ratings, even high ratings. In fact, the general message

from these authors is that the combination of low typicality and high quality is a sign of real creativity, possibly meriting a space-transformation.

Using the same attributes for both typicality and quality is unnecessarily restrictive from the point of view of imitating human judgements in artistic domains, where there may be certain minimum standards for membership of the genre, but other, more subtle, criteria for separating good from bad instances. In certain forms of poetry (e.g. in British nineteenth century culture) a text is a poem if it meets certain basic standards of syntactic and semantic coherence, and conforms to some clear metric structure; rhyming is a further desirable feature. (There are exceptions – some works of William MacGonagall give priority to rhyme over metre.) However, the quality of a poem may depend on further factors, such as its emotional effect, or the profundity of its content. Similarly, the attributes which ensure that a text is a joke may not be sufficient to determine whether or not it is a funny joke[34, p.16]. So, in general, judgements of typicality may be based on different factors from judgements about quality. This means we have to allow for a "quality space" logically distinct from "typicality space". (The formal models discussed earlier would not all be equally plausible for a quality space: the state-transition model seems less likely than the multi-dimensional model, for example.)

This means that we have to consider not one but possibly two space transformations. When an artefact stimulates transformation, it could be that change occurs in the typicality space, or the quality space, or both. If both, the magnitude of the two changes need not be the same.

6.3 Space-induction and properties

For all the types of formal model discussed above, there are learning procedures which will induce definitions given a set of examples, *assuming the set of relevant properties is known*. This need for pre-selected relevant properties might seem to defer the solution still further, or even result in circularity (needing to know the space before computing it), but there is a distinction between knowing the basic properties relevant to the genre and knowing what combinations of these properties constitute the space-definition for that genre. Also, in a potentially creative activity, such as writing, painting, composing, there will be an established culture which will make available a basic set of attributes for artefacts (e.g. hue, pitch, prosodic stress). If an artefact manifests an entirely new property that had not even been thought of as a possible attribute

(i.e. was outside the logically possible space), then the process we are sketching may have difficulty even representing the nature of that artefact, which is indeed a problem. However, the situation we are analysing (following the cases discussed by Boden and others) is where an artefact can be seen to be significantly different from other examples. If the difference is detectable, the basic properties on which it is based must be within the system. This highlights another aspect of the analysis of such creative domains: when developing a formal model, it is reasonable to assume that the data under consideration will not just be the raw artefacts, but will include some human ratings of the artefacts in some way. Either this will be in terms of some pre-defined attributes, or it might be in terms of similarity judgements. In the latter case, there are techniques (e.g. multi-dimensional scaling) which can compute an underlying multi-dimensional space from a collection of similarity judgements about a set of objects.

6.4 How many spaces?

So far we have focussed mainly on the typicality space, but Section 6.2 indicates that there may also be some sort of quality space, with its own options for change.

Even in the area of typicality, it is not clear whether advocates of transformation have in mind changes to the boundaries of the space (the set of possible artefacts within the genre), or whether the alterations might be to the way that the artefacts are distributed within that space: contrasts and similarities, degrees of variation, etc. A radical rearrangement of artefacts within the existing space might reflect the kinds of changes that Boden and others have offered as instances of transformation. Distinguishing between these empirically may be exceedingly difficult.

If, as in Section 5.3, a weight were attached to each dimension of a space, this would map items into another space (of the same number of dimensions), the layout of which would change with changes to the weights.

As mentioned in Section 4, the rules governing a system's search for artefacts affect what might be produced[31, 40]. The term "search space" could be taken to mean simply "the set of items which are available (to the searching program)", or it could mean "a representation indicating the items available and also showing routes which are available for traversing this set of items". By the former perspective, the mere set of possible games in chess defines a search

space, without indications of which board-states can be reached from which others, thus not showing how exploration strategies could be defined; by the latter perspective, the search space is more of a route-map between states, so that (for example) depth-first and breadth-first exploration of the routes can be defined. (Wiggins[40] separates these aspects, by having a set of available concepts which is distinct from the way that a creative agent traverses the concepts.) Whatever a conceptual space is, it will constitute a search space in the weaker sense, since it defines (somehow) a set of items. If, as in Section 4.4, one characteristic of a conceptual space is that it supports exploration, then it will also provide a search space in the stronger sense. However, that ready-made search space may not be the only conceivable one for that set of artefacts, and the search space used by a computation need not be based solely on dimensions or relations which exist naturally between items in the conceptual space. For example, Manurung's evolutionary algorithm for poetry generation[22, 21] searches through possible texts in an order which does not reflect the natural domain properties such as metre and rhyme. So even where the conceptual space provides links between concepts/artefacts, the search space is logically distinct, and is yet another area where change might occur. Perhaps imposing a search route which is radically different from the underlying structure of the conceptual space could be regarded as a "transformation" of that space?

Hence, transformation could affect any of the above "spaces" (quality, typicality boundary, typicality layout, weighted space, search space). These degrees of freedom are not helpful in trying to pin down exactly what types of space-change give rise to creativity.

For simplicity, we shall continue to talk as if there was only one space involved, since some of the methodological issues are the same for all of them, but the more complicated arrangement should be borne in mind.

6.5 The metalevel

It has been suggested that genuine creativity involves processing at a metalevel, that transformational creativity consists of metalevel computation and even that it may consist of an exploration at the metalevel comparable to that which goes on at the object level (the main space)[8, 7, 40]. If we take the artefacts and the conceptual space, however formalised, to constitute the object level, then any change to this which involves non-trivial computation (e.g. to select a suitable change to the space) necessarily requires metalevel processing.

Whether this is in any interesting and substantive way similar (but at a different level) to what goes on in creative exploration without space-change is an open question. Wiggins[40] has shown that it is possible to state the processing at the two levels in a similar formal way, but his account is so general and abstract (search within an unstructured space with an evaluation function) that it does not prove very much (and is formal rather than empirical). Wiggins considers the actions of the creating agent. We have taken the perspective of the individual assessing the artefact (with the most minimal assumptions about actual creation), and argued that *if a new space must be found, then this demands some form of metalevel (outside the space) processing*. Thus, whether one attributes the transformation to the creator (Wiggins, Boden, Bundy) or to the individual appreciating the artefact (as here), computing a new structure for the object level necessitates metalevel work.

§7 Empirical testing

It is highly implausible to claim that any transformation whatsoever of a conceptual space results in high creativity, so we shall assume that Boden's hypothesis is not that transformation is a *sufficient* condition for high creativity; rather, we assume that Boden is offering transformation as a *necessary* condition for high creativity (i.e. all – or most – highly creative acts involve transformation). If this claim is to be empirical, then a great deal needs to be done. The hypothesis has not even been substantiated for human creativity (even in a single genre, let alone all genres), despite the fact that this should be a prerequisite for applying it to machine creativity.

7.1 The formal model

The first stage is to develop a precise formal model. This involves various essential steps.

(i) **Choose type of formal model.** To proceed concretely, we have to at least provisionally adopt a formal mechanism for the study. However, the question of which formal model is best suited to analysis of creative processes is itself an open research question. That is why this choice of a type of formalisation is only tentative. We can look upon this as a first stage in a methodological "generate-and-test" approach, whereby we select a formal framework and see where it takes us. Then we can "backtrack" to choose another type of model, and see

how well that works. Logically, we could compare all possible formal models at once (a "breadth-first" approach), but that might prove cumbersome in practice. For the exposition here, we shall assume that one class of model is chosen initially.

(ii) **Define basic space notions.** Using some chosen type of formalisation (e.g. connectionist networks), we would have to specify what would constitutes space-definition: what parameters to instantiate, what constructs to define, etc. This would almost certainly including defining the first of the five operations listed in Section 4.4: membership.

(iii) **Define space-induction.** Of the five operations listed in Section 4.4, the next crucial one is a decidable way of assigning a space-definition to a given set of artefacts (the induction operation). (It is not essential to define exploration at this stage, and possibly not even similarity, unless these are implicit in, or needed for, the other steps). It is conceivable that the chosen machine learning method would compute more than one possible definition for given data. In that case, either some further criteria would have to be defined, to choose between the competing analyses, or when, using the space induction operator, all the possibilities would have to be investigated.

(iv) **Define transformation.** Transformation could be handled in either of the two ways outlined earlier: as a mapping from a space to another space, or as a binary predicate over spaces. What is crucial is that there is a set of criteria for deciding when an artefact manifests (or demands) a transformation of typicality space, quality space, or both.

If the case is to be made for the transformational hypothesis across a range of medium types and genres, the definitions and criteria listed above will have to be extremely general and abstract, so there will also have to be information about how these abstract principles are made concrete in particular genres.

None of the above is trivial, but if proponents of the transformational creativity hypothesis cannot at least make these steps, then empirical testing of the hypothesis would appear to be impossible.

7.2 Human judgements

We argued in Section 2 that human judgements about creativity are the ultimate test of whether some program or mechanism has indeed led to cre-

ativity. So far, this has been tacitly assumed in the literature, as authors cite occasional examples from human art or science to illustrate claims about computational creativity (e.g. Boden (passim), or the review of musical composition by Wiggins[40]). This is comparable to the way in which linguists offer isolated examples (often artificially constructed) to support claims about grammatical or semantic rules. It is, in effect, a very informal form of empirical testing. For serious scientific testing of a hypothesis, something more rigorous is needed.

Superficially, the necessary steps appear to be as follows.

(i) **Build a data set.** Firstly, compile a suitably varied collection of human-created artefacts, aiming to include items which are likely to be regarded as examples of high creativity, and (preferably) comparable examples of lesser degrees of creativity.

(ii) **Analyse the artefacts.** For these items, use the chosen model (Section 7.1) to place these artefacts in a suitable space, and decide which of these artefacts are examples of space-transformation. If necessary, revise the data set to ensure that there is a spread of transformational and non-transformational items.

(iii) **Collect creativity ratings.** Using studies with human subjects, collect creativity ratings for the artefacts in the data set. This constitutes a "gold standard" for assessing measures of creativity.

However, the collection of human creativity ratings is not completely straightforward. The most obvious approach would be simply to construct a set of items, where each item is an exemplar of some genre, and ask subjects to rate each item for "creativity". Various questions arise about such an approach. Should subjects judge single artefacts in isolation, or should this happen relative to some context, perhaps a collection of other artefacts (in keeping with Section 3, earlier)? How can we control for subjects basing judgements on their own (subjectively variable) knowledge of the genre? Will subjects be able to give verdicts on 'creativity' separately from other properties such as 'technical competence' or 'being attractive'?

We do not offer to resolve these questions here, but merely observe that some experimental framework must be devised if the transformational hypothesis is to connect with human judgement.

7.3 Compare predictions with data

Once we have data annotated with human judgements, and a formal model which defines which items count as transformational, then we can proceed to compare the predictions of the model with the actual data. This should either falsify or be compatible with the hypothesis that instances of higher creativity are the result of transformation.

A single study of this sort would be interesting, but not conclusive. However, if repeated studies in various genres failed to falsify the hypothesis, we could have some confidence that the claim was probably true.

7.4 Testing the hypothesis on computer output

If we are to test whether space-transformation leads to improved creativity when embodied in a program, then we have to make a decision about methodological strategy: on what do we base our judgement about whether transformation has been involved:

(i) an analysis of what provides the most suitable description of the output of the program?

(ii) an examination of the processing implemented in the program?

If we opt for (i), then the situation is exactly parallel to the human case, differing only in the source of the artefacts for the judgement step (Section 7.2): the items should include output from the computer program.

In case (ii), if we decide that the program's computations do count as transformation, then we then have the further question to consider: is this transformational processing central to the program's processing, or could an elegant and adequate non-transformational account could be given of the computation? It might be that the computation could have been organised in a number of ways, some of which are "more transformational" than others. That is, we have to consider whether the same behaviour (including production of artefacts) could result from two variants of the program, one transformational and the other not.

7.5 Transformation as a definition

Let us return (briefly) return to the idea that Boden was not making an empirical claim about creativity, but *defining* creativity. There are two variants to consider. The weaker form is that, like the hypothesis, the statement lays down only necessary, not sufficient, conditions. This would be a stipulation, *a priori*, that an activity can be deemed highly creative only if it involves trans-

formation (this does seem to be the way Boden judges the cases she considers). This version could be applied to determining the creativity of a program only by eliminating non-transformational computations from consideration as instances of profound creativity – it could not tell which transformational computations were deeply creative.

The stronger version is a true definition, in which transformation is a necessary *and* sufficient condition (for high creativity). This could be employed directly to gauge the creativity of a program, by virtue of the sufficiency clause. We could directly determine whether a computer program had been truly creative by checking whether or not it had transformed a space. That could be checked in two possible ways, as in Section 7.4 above: analyse the output, or dissect the internal workings. This would immediately answer the question: 'has this computer program (on this occasion) been creative (by performing space-transformation)?'

However, there is a problem with either of these variants of taking transformation as defining high creativity. With the first version (necessary condition), what if a non-transformational program produced output which human judges unanimously rated as highly creative? With the second version (necessary and sufficient conditions), what if some program were to be labelled "creative" by virtue of being transformational, but human judges did not rate its output as being at all creative? In either of these cases, would we have to tell the judges they were wrong? Relying on a formal definition of "true creativity" would detach us from our foundations in human judgements of creativity (Section 2). It seems unlikely that this is what Boden intended.

§8 Conclusion

None of the above is to deny, or play down, the importance of other factors in creativity. For example, Buchanan[7] mentions the relevance of the creator's background knowledge and skills, and the effects of experience. The very limited brief we have set ourselves here is to consider how the question of space-transformation might be formalised, in order that various claims about the necessity or effectiveness of transformation might be made empirical.

The approach outlined in in Section 7 may not be the only route by which claims about transformational creativity can be made concrete and testable, but it is at least, in sketch form, one possibility (albeit one involving some potentially problematic subtasks). Any advocate of transformational creativity as

a superior form of creativity who does not offer some comparable route to falsification/corroboration is in a weak position from an empirical point of view. We may have to resign ourselves to treating this hypothesis as unfalsifiable.

However, it may be that published discussions of transformational creativity do not intend to set out an empirical scientific hypothesis. If so, we can lay aside the concerns discussed in this paper. In particular, devising computational architectures which might be particularly useful in building creative programs in specific genres is a *different* problem, and one that could be addressed without bothering with any of the issues we have discussed here. Those who wish to build generators of music, poetry, art, jokes, concepts, etc. can move ahead regardless of the status of claims about transformations. Perhaps the conjecture that "transformational is better" should be left as a loose slogan to inspire program designers, rather than viewed as a strict hypothesis.

Acknowledgment Thanks are due to the Aberdeen NLG group for helpful discussion, and to the journal referees for useful suggestions.

References

1) Attardo, S., and Raskin, V. "Script theory revis(it)ed: joke similarity and joke representation model". *Humor: International Journal of Humor Research*, 4(3), pp. 293–347, 1991.

2) Baggi, D. ed. *Readings in Computer Generated Music*. IEEE Computer Society Press, New York, 1992.

3) Binsted, K., and Ritchie, G. "Computational rules for generating punning riddles". *Humor: International Journal of Humor Research*, 10(1), pp. 25–76, 1997.

4) Boden, M. "Modelling creativity: reply to reviewers". *Artificial Intelligence*, 79, pp. 161–182, 1995.

5) Boden, M. A. *The Creative Mind*. Abacus, London, 1992. First published 1990, second edition Routledge, London, 2003.

6) Boden, M. A. "Creativity and Artificial Intelligence". *Artificial Intelligence*, 103, pp. 347–356, 1998.

7) Buchanan, B. "Creativity at the metalevel". *AI Magazine*, (Fall 2001), pp. 13–28, 2001. AAAI-2000 Presidential Address.

8) Bundy, A. "What is the difference between real creativity and mere novelty?". *Behavioral and Brain Sciences*, 17(3), pp. 533–534, 1994. Open Peer Commentary on 5).

9) Cohen, B., and Murphy, G. L. "Models of concepts". *Cognitive Science*, 8, pp. 27–58, 1984.

10) Colton, S. *Automated Theory Formation in Pure Mathematics*. Distinguished Dissertations. Springer-Verlag, London, 2002.

11) Gärdenfors, P. "Induction, conceptual spaces and AI". *Philosophy of Science*, 57, pp. 78–95, 1990.

12) Gärdenfors, P. "Conceptual spaces as a framework for knowledge representation". *Mind and Matter*, 2(2), pp. 9–27, 2004.

13) Gärdenfors, P., and Williams, M.-A. "Reasoning about categories in conceptual spaces", in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 385–392. Morgan Kaufman, 2001.

14) Gervás, P. "WASP: Evaluation of different strategies for the automatic generation of spanish verse", in *Proceedings of the AISB 00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science* (Wiggins, G. A. ed.), pp. 93–100. Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2000.

15) Gervás, P. "Generating poetry from a prose text: Creativity versus faithfulness", in *Proceedings of the AISB 01 Symposium on Artificial Intelligence and Creativity in Arts and Science* (Wiggins, G. A. ed.), pp. 93–99. Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2001.

16) Gervás, P. "Exploring quantitative evaluations of the creativity of automatic poets", in *2nd Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science, ECAI 2002* (Bento, C., Cardoso, A., and Wiggins, G. eds.), Lyon, France, 2002.

17) Hopcroft, J., and Ullman, J. *Introduction to automata theory, languages, and computation*. Addison-Wesley, Reading, Mass, 1979.

18) Keeney, R. L., and Raiffa, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, 1976.

19) Lenat, D. "On automated scientific theory formation: a case study using the AM program", in *Machine Intelligence 9* (Hayes, J., Michie, D., and Mikulich, L. eds.), pp. 251–283. Ellis Horwood, Chichester, 1979.

20) Mackworth, A. K. "Consistency in networks of relations". *Artificial Intelligence*, 8, pp. 99–118, 1977.

21) Manurung, H. M., Ritchie, G., and Thompson, H. "A flexible integrated architecture for generating poetic texts", in *Proceedings of the Fourth Symposium on Natural Language Processing (SNLP 2000)*, pp. 7–22, Chiang Mai, Thailand, 2000.

22) Manurung, H. M., Ritchie, G., and Thompson, H. "Towards a computational model of poetry generation", in *Proceedings of the AISB 00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science* (Wiggins, G. A. ed.), pp. 79–86. Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2000.

23) Medin, D. L., and Smith, E. E. "Concepts and concept formation". *Annual Review of Psychology*, 35, pp. 113–138, 1984.

24) Meehan, J. *The metanovel : writing stories by computer*. PhD thesis, Yale University, Department of Computer Science, 1976.

25) Minsky, M. ed. *Semantic Information Processing*. MIT Press, Cambridge, Mass., 1968.

26) Miranda, E. *Composing Music with Computers*. Focal Press/Elsevier, Amsterdam, 2001.

27) Nilsson, N. J. *Problem-solving methods in artificial intelligence*. McGraw-Hill, New York, 1971.

28) O'Rourke, J. "The generative-rules definition of creativity". *Behavioral and Brain Sciences*, 17(3), p. 547, 1994. Open Peer Commentary on 5).

29) Osherson, D. N., and Smith, E. E. "On the adequacy of prototype theory as a theory of concepts". *Cognition*, 9, pp. 35–58, 1981.

30) Pease, A., Winterstein, D., and Colton, S. "Evaluating machine creativity", in *Case-Based Reasoning: Papers from the Workshop Programme at ICCBR 01* (Weber, R., and von Wangenheim, C. G. eds.), pp. 129–137, Vancouver, 2001.

31) Perkins, D. "An unfair review of Margaret Boden's *The Creative Mind* from the perspective of creative systems". *Artificial Intelligence*, 79, pp. 97–109, 1995.

32) Ram, A., Wills, L., Domeshek, E., Neressian, N., and Kolodner, J. "Understanding the creative mind: a review of Margaret Boden's *Creative Mind*". *Artificial Intelligence*, 79, pp. 111–128, 1995.

33) Ritchie, G. "Assessing creativity", in *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in Arts and Science*, pp. 3–11, York, England, 2001.

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

35) Rosch, E., and Mervis, C. "Family resemblances: Studies in the internal structure of categories". *Cognitive Psychology*, 7, pp. 573–605, 1975.

36) Rosch, E. "On the internal structure of perceptual and semantic categories", in *Cognitive development and the acquisition of language* (Moore, T. E. ed.), pp. 111–144. Academic Press, New York, 1973.

37) Schank, R. C., and Foster, D. A. "The engineering of creativity: a review of Boden's *Creative Mind*". *Artificial Intelligence*, 79, pp. 129–143, 1995.

38) Turner, S. "Margaret Boden, *The Creative Mind*". *Artificial Intelligence*, 79, pp. 145–159, 1995.

39) Turner, S. R. *The Creative Process: A Computer Model of Storytelling*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1994.

40) Wiggins, G. "Towards a more precise characterisation of creativity in AI", in *Case-Based Reasoning: Papers from the Workshop Programme at ICCBR 01* (Weber, R., and von Wangenheim, C. G. eds.), Vancouver, 2001. Navy Center for Applied Research in Artificial Intelligence.

41) Wiggins, G. "Categorising creative systems", in *Proceedings of Third (IJCAI) Workshop on Creative Systems: Approaches to Creativity in Artificial Intelligence and Cognitive Science*, 2003.