Creativity and artificial intelligence

Margaret A. Boden

School of Cognitive and Computing Sciences, University of Sussex, Brighton, Sussex, BN1 9QH, England, UK

Abstract

Creativity is a fundamental feature of human intelligence, and a challenge for AI. AI techniques can be used to create new ideas in three ways: by producing novel combinations of familiar ideas; by exploring the potential of conceptual spaces; and by making transformations that enable the generation of previously impossible ideas. AI will have less difficulty in modelling the generation of new ideas than in automating their evaluation. © 1998 Elsevier Science B.V. All rights reserved.

1. Why AI must try to model creativity

Creativity is a fundamental feature of human intelligence, and an inescapable challenge for AI. Even technologically oriented AI cannot ignore it, for creative programs could be very useful in the laboratory or the market-place. And AI-models intended (or considered) as part of cognitive science can help psychologists to understand how it is possible for human minds to be creative.

Creativity is not a special "faculty", nor a psychological property confined to a tiny elite. Rather, it is a feature of human intelligence in general. It is grounded in everyday capacities such as the association of ideas, reminding, perception, analogical thinking, searching a structured problem-space, and reflective self-criticism. It involves not only a cognitive dimension (the generation of new ideas) but also motivation and emotion, and is closely linked to cultural context and personality factors [3]. Current AI models of creativity focus primarily on the cognitive dimension.

A creative idea is one which is novel, surprising, and valuable (interesting, useful, beautiful...). But "novel" has two importantly different senses here. The idea may be novel with respect only to the mind of the individual (or AI-system) concerned or, so far as we know, to the whole of previous history. The ability to produce novelties of the former kind may be called P-creativity (P for psychological), the latter H-creativity (H for historical). P-creativity is the more fundamental notion, of which H-creativity is a special case.

AI should concentrate primarily on P-creativity. If it manages to model this in a powerful manner, then artificial H-creativity will occur in some cases—indeed, it already has, as we shall see. (In what follows, I shall not use the letter-prefixes: usually, it is P-creativity which is at issue.)

2. Three types of creativity

There are three main types of creativity, involving different ways of generating the novel ideas. Each of the three results in surprises, but only one (the third) can lead to the "shock" of surprise that greets an apparently impossible idea [2]. All types include some H-creative examples, but the creators celebrated in the history books are more often valued for their achievements in respect of the third type of creativity.

The first type involves novel (improbable) combinations of familiar ideas. Let us call this "combinational" creativity. Examples include much poetic imagery, and also analogy—wherein the two newly associated ideas share some inherent conceptual structure. Analogies are sometimes explored and developed at some length, for purposes of rhetoric or problem-solving. But even the mere generation, or appreciation, of an apt analogy involves a (not necessarily conscious) judicious structural mapping, whereby the similarities of structure are not only noticed but are judged in terms of their strength and depth.

The second and third types are closely linked, and more similar to each other than either is to the first. They are "exploratory" and "transformational" creativity. The former involves the generation of novel ideas by the exploration of structured conceptual spaces. This often results in structures ("ideas") that are not only novel, but unexpected. One can immediately see, however, that they satisfy the canons of the thinking-style concerned. The latter involves the transformation of some (one or more) dimension of the space, so that new structures can be generated which could not have arisen before. The more fundamental the dimension concerned, and the more powerful the transformation, the more surprising the new ideas will be. These two forms of creativity shade into one another, since exploration of the space can include minimal "tweaking" of fairly superficial constraints. The distinction between a tweak and a transform is to some extent a matter of judgement, but the more well-defined the space, the clearer this distinction can be.

Many human beings—including (for example) most professional scientists, artists, and jazz-musicians—make a justly respected living out of exploratory creativity. That is, they inherit an accepted style of thinking from their culture, and then search it, and perhaps superficially tweak it, to explore its contents, boundaries, and potential. But human beings sometimes transform the accepted conceptual space, by altering or removing one (or more) of its dimensions, or by adding a new one. Such transformation enables ideas to be generated which (relative to that conceptual space) were previously impossible.

The more fundamental the transformation, and/or the more fundamental the dimension that is transformed, the more different the newly-possible structures will be. The shock of amazement that attends such (previously impossible) ideas is much greater than the surprise occasioned by mere improbabilities, however unexpected they may be. If the

transformations are too extreme, the relation between the old and new spaces will not be immediately apparent. In such cases, the new structures will be unintelligible, and very likely rejected. Indeed, it may take some time for the relation between the two spaces to be recognized and generally accepted.

3. Computer models of creativity

Computer models of creativity include examples of all three types. As yet, those focussed on the second (exploratory) type are the most successful. That's not to say that exploratory creativity is easy to reproduce. On the contrary, it typically requires considerable domain-expertise and analytic power to define the conceptual space in the first place, and to specify procedures that enable its potential to be explored. But combinational and transformational creativity are even more elusive.

The reasons for this, in brief, are the difficulty of approaching the richness of human associative memory, and the difficulty of identifying our values and of expressing them in computational form. The former difficulty bedevils attempts to simulate combinational creativity. The latter difficulty attends efforts directed at any type of creativity, but is especially problematic with respect to the third (see Section 4, below).

Combinational creativity is studied in AI by research on (for instance) jokes and analogy. Both of these require some sort of semantic network, or inter-linked knowledge-base, as their ground. Clearly, pulling random associations out of such a source is simple. But an association may not be telling, or appropriate in context. For all combinational tasks other than "free association", the nature and structure of the associative linkage is important too. Ideally, every product of the combinational program should be at least minimally apt, and the originality of the various combinations should be assessable by the AI-system.

A recent, and relatively successful, example of AI-generated (combinational) humour is Jape, a program for producing punning riddles [1]. Jape produces jokes based on nine general sentence-forms, such as: What do you get when you cross X with Y?; What kind of X has Y?; What kind of X can Y?; What's the difference between an X and a Y? The semantic network used by the program incorporates knowledge of phonology, semantics, syntax, and spelling. Different combinations of these aspects of words are used, in distinctly structured ways, for generating each joke-type.

Examples of riddles generated by Jape include: (Q) What kind of murderer has fibre? (A) A cereal killer; (Q) What do you call a strange market? (A) A bizarre bazaar; (Q) What do you call a depressed train? (A) A low-comotive; and (Q) What's the difference between leaves and a car? (A) One you brush and rake, the other you rush and brake. These may not send us into paroxysms of laughter—although, in a relaxed social setting, one or two of them might. But they are all amusing enough to prompt wryly appreciative groans.

Binsted did a systematic series of psychological tests, comparing people's reception of Jape's riddles with their response to human-originated jokes published in joke-books. She also compared Jape's products with "non-jokes" generated by random combinations. She found, for instance, that children, by whom such humour is most appreciated, can distinguish reliably between jokes (including Jape's riddles) and non-jokes. Although they generally find human-originated jokes funnier than Jape's, this difference vanishes if Jape's

output is pruned, so as to omit the items generated by the least successful schemata. The riddles published in human joke-books are highly selected, for only those the author finds reasonably funny will appear in print.

Binsted had set herself a challenging task: to ensure that every one of Jape's jokes would be amusing. Her follow-up research showed that although none were regarded as exceptionally funny, very few produced no response at all. This contrasts with some other AI-models of creativity, such as AM [16], where a high proportion of the newly generated structures are not thought interesting by human beings.

It does not follow that all AI-modelling of creativity should emulate Binsted's ambition. This is especially true if the system is meant to be used interactively by human beings, to help their own creativity by prompting them to think about ideas that otherwise they might not have considered. Some "unsuccessful" products should in any case be allowed, as even human creators often produce second-rate, or even inappropriate, ideas. Jape's success is due to the fact that its joke-templates and generative schemata are very limited. Binsted identifies a number of aspects of real-life riddles which are not parallelled in Jape, and whose (reliably funny) implementation is not possible in the foreseeable future. To incorporate these aspects so as to produce jokes that are reliably funny would raise thorny questions of evaluation (see Section 4).

As for AI-models of analogy, most of these generate and evaluate analogies by using domain-general mapping rules, applied to prestructured concepts (e.g. [7,12,13]). The creators of some of these models have compared them with the results of psychological experiments, claiming a significant amount of evidence in support of their domain-general approach [8]. In these models, there is a clear distinction between the representation of a concept and its mapping onto some other concept. The two concepts involved usually remain unchanged by the analogy.

Some AI-models of analogy allow for a more flexible representation of concepts. One example is the Copycat program, a broadly connectionist system that looks for analogies between alphabetic letter-strings [11,18]. Copycat's concepts are context-sensitive descriptions of strings such as "mmpprr" and "klmmno". The two m's in the first string just listed will be described by Copycat as a pair, but those in the second string will be described as the end-points of two different triplets.

One might rather say that Copycat will "eventually" describe them in these ways. For its concepts evolve as processing proceeds. This research is guided by the theoretical assumption that seeing a new analogy is much the same as perceiving something in a new way. So Copycat does not rely on ready-made, fixed, representations, but constructs its own in a context-sensitive way: new analogies and new perceptions develop together. A part-built description that seems to be mapping well onto the nascent analogy is maintained, and developed further. One that seems to be heading for a dead end is abandoned, and an alternative begun which exploits different aspects. The model allows a wide range of (more or less daring) analogies to be generated, and evaluated. The degree to which the analogies are obvious or far-fetched can be altered by means of one of the system-parameters.

Whether the approach used in Copycat is preferable to the more usual forms of (domain-general) mapping is controversial. Hofstadter [11] criticizes other AI-models of analogy for assuming that concepts are unchanging and inflexible, and for guaranteeing that the

required analogy (among others) will be found by focussing on small representations having the requisite conceptual structures and mapping rules built in. The opposing camp rebut these charges [8].

They argue that to identify analogical thinking with high-level perception, as Hofstadter does, is to use a vague and misleading metaphor: analogical mapping, they insist, is a domain-general process which must be analytically distinguished from conceptual representation. They point out that the most detailed published account of Copycat [18] provides just such an analysis, describing the representation-building procedures as distinct from, though interacting with, the representation-comparing modules. They report that the Structure Mapping Engine (SME), for instance, can be successfully used on representations that are "very large" as compared with Copycat's, some of which were built by other systems for independent purposes. They compare Copycat's alphabetic microworld with the "blocks world" of 1970s scene analysis, which ignored most of the interesting complexity (and noise) in the real-world. Although their early models did not allow for changes in conceptual structure as a result of analogising, they refer to work on learning (using SME) involving processes of schema abstraction, inference projection, and re-representation [9]. Moreover (as remarked above), they claim that their psychological experiments support their approach to simulation. For example, they say there is evidence that memory access, in which one is reminded of an (absent) analog, depends on psychological processes, and kinds of similarity, significantly different from those involved in mapping between two analogs that are presented simultaneously.

The jury remains out on this dispute. However, it may not be necessary to plump absolutely for either side. My hunch is that the Copycat approach is much closer to the fluid complexity of human thinking. But domain-general principles of analogy are probably important. And these are presumably enriched by many domain-specific processes. (Certainly, psychological studies of how human beings retrieve and interpret analogies are likely to be helpful.) In short, even combinational creativity is, or can be, a highly complex matter.

The exploratory and transformational types of creativity can also be modelled by AI-systems. For conceptual spaces, and ways of exploring and modifying them, can be described by computational concepts.

Occasionally, a "creative" program is said to apply to a wide range of domains, or conceptual spaces—as EURISKO, for instance, does [16]. But to make this generalist program useful in a particular area, such as genetic engineering or VLSI-design, considerable specialist knowledge has to be provided if it is not to generate hosts of nonsensical (as opposed to merely boring) ideas. In general, providing a program with a representation of an interesting conceptual space, and with appropriate exploratory processes, requires considerable domain-expertise on the part of the programmer—or at least on the part of someone with whom he cooperates. (Unfortunately, the highly subject-bounded institutional structure of most universities works against this sort of interdisciplinarity.)

For example, EMI (experiments in musical intelligence) is a program that composes in the styles of Mozart, Stravinsky, Joplin, and others [6]. In order to do this, it employs powerful musical grammars expressed as ATNs. In addition, it uses lists of "signatures": melodic, harmonic, metric, and ornamental motifs characteristic of individual composers.

Using general rules to vary and intertwine these, it often composes a musical phrase near-identical to a signature that has *not* been provided. This suggests a systematicity in individual composing styles.

Individual musical style has been addressed also in a pioneering program that improvises jazz in real time, though the technique can be applied to other types of music [10]. The most highly developed version, at present, generates jazz in the style of Charlie Parker—and (ignoring the lack of expressiveness, and the quality of the synthesized sound) it actually sounds like Parker. Besides strong (and relatively general) knowledge of musical dimensions such as harmony and rhythm, and of musical conventions characteristic of jazz, the system has access to a large set of Parker-specific motifs, which can be varied and combined in a number of ways. (The programmer is an accomplished jazz-saxophonist: without strong musical skills, he would not be able to identify the relevant motifs, or judge the aptness of specific processes for using them.) In exploring this conceptual space, the program often originates interesting musical ideas, which jazz-professionals can exploit in their own performance. However, in its present form it never moves outside Parker-space: its creativity is merely exploratory, not transformational.

Architectural design, too, has been formally modelled. For instance, a shape-grammar describing Frank Lloyd Wright's Prairie houses generates all the ones he designed, as well as others he did not [14]. To the initiated eye, every one of these novel (exploratory-creative) structures falls within the genre. The grammar not only identifies the crucial dimensions of the relevant architectural space, but also shows which are relatively fundamental. In a Prairie house, the addition of a balcony is stylistically superficial, for it is a decision on which nothing else (except the appearance and ornamentation of the balcony) depends. By contrast, the "addition" of a fireplace results in overall structural change, because many design-decisions follow, and depend upon, the (early) decision about the fireplace. Exploring this space by making different choices about fireplaces, then, can give rise to surprises more fundamental than can adding balconies in unexpected places.

Perhaps the best-known example of AI-creativity is AARON, a program—or rather, a series of programs—for exploring line-drawing in particular styles [17] and, more recently, colouring also [5]. Written by Harold Cohen, an artist who was already an acclaimed professional in the 1960s, AARON explores a space defined with the help of rich domain-expertise.

AARON is not focussed primarily on surfaces, but generates some representation of a 3D-core, and then draws a line around it. Versions that can draw many idiosyncratic portraits use 900 control points to specify the 3D-core, of which 300 specify the structure of the face and head. The program's drawings are aesthetically pleasing, and have been exhibited in galleries worldwide. Until very recently, coloured images of AARON's work were hand-painted by Cohen. But in 1995, he exhibited a version of AARON that can do this itself. It chooses colours by tonality (light/dark) rather than hue, although it can decide to concentrate on a particular family of hues. It draws outlines using a paintbrush, but colours the paper by applying five round "paint-blocks" of differing sizes. Some characteristic features of the resulting painting style are due to the physical properties of the dyes and painting-blocks rather than to the program guiding their use. Like drawing-AARON, painting-AARON is still under continuous development.

The drawings (and paintings) are individually unpredictable because of random choices, but all the drawings produced by a given version of AARON will have the same style. AARON cannot reflect on its own productions, nor adjust them so as to make them better. It cannot even transform its conceptual space, leaving aside the question of whether this results in something "better". In this, it resembles most current AI-programs focussed on creativity.

A further example of exploratory AI-creativity is the BACON suite designed to model scientific discovery [15]. The heuristics used by the BACON system are carefully pre-programmed, and the data are deliberately prestructured so as to suit the heuristics provided. New types of discovery are impossible for BACON. It is therefore misleading to name such programs after scientists remembered for noticing relations of a type never noticed before. Even the notion that there may be (for instance) some linear mathematical relation to be found was a huge creative leap.

Almost all of today's "creative" computers are concerned only with exploring pre-defined conceptual spaces. They may allow for highly constrained tweaking, but no fundamental novelties or truly shocking surprises are possible. However, a few AI-systems attempt not only to explore their conceptual space but also to transform it, sometimes in relatively unconstrained ways.

Transformational systems include AM and EURISKO [16], and certain programs based on genetic algorithms. Some of these have produced valued structures that the human experts say they could never have produced unaided: the sculptor William Latham, for example, has generated 3D-forms of a type which he could not have imagined for himself [22].

Most GA-programs only explore a pre-given space, seeking the "optimal" location within it. But some also transform their generative mechanism in a more or less fundamental way. For example, GA-work in graphics may enable superficial tweaking of the conceptual space, resulting in images which, although novel, clearly belong to the same family as those which went before [22]. Or it may allow the core of the image-generating code to be lengthened and complexified, so that the novel images may bear no family-resemblance even to their parents, still less to their more remote ancestors [21]. Similarly, some work in evolutionary robotics has generated novel sensory-motor anatomies and control systems as a result of GAs that allow the length of the "genome" to be altered [4].

One should not assume that transformation is always creative, or even—in the present state of the art—that AI-systems that can transform their rules are superior to those which cannot. Significantly, some AI-modellers deliberately avoid giving their programs the capacity to change the heart of the code. That is, they prevent fundamental transformations in the conceptual space, allowing only exploration and relatively superficial tweaking. One reason for this is the human may be more interested, at least for a time, in exploring a given space than in transforming it in unpredictable ways. A professional sculptor such as Latham, for instance, may wish to explore the potential (and limits) of one particular family of 3D-structures, before considering others [22]. Another reason for avoiding rampant transformation in AI-models of creativity is the difficulty of automating evaluation.

4. The evaluation of new ideas

A main reason why most current AI-models of creativity attempt only exploration, not transformation, is that if the space is transformed then the resulting structures may not have any interest or value. Such ideas are novel, certainly, but not creative. (We saw in Section 1 that "creativity" implies positive evaluation.)

This would not matter if the AI-system were able to realize the poor quality of the new constructions, and drop (or amend) the transformation accordingly. A truly automatic AI-creator would have evaluative mechanisms sufficiently powerful to do this. At present, this is very rarely so (an exception is artificial co-evolution in which the fitness function evolves alongside the several species involved [19]). Notoriously, AM produced many more useless items than powerful mathematical ideas, and although it did have heuristics of "interestingness" built into it, its evaluations were often mistaken by human standards. And some "adventurously" transformational programs embody no evaluative criteria at all, the evaluation being done interactively by human beings [21].

There is no reason in principle why future AI-models should not embody evaluative criteria powerful enough to allow them to transform their conceptual spaces in fruitfully creative (including H-creative) ways. But for such computerized self-criticism to be possible, the programmers must be able to express the values concerned sufficiently clearly for them to be implemented. Even if the values are not predetermined, being represented instead as an evolving fitness function, the relevant features must be implemented in and recognized by the (GA) system.

To some extent, this can be achieved implicitly, by defining a culturally accepted conceptual space so successfully that any structure that can be generated by the program will be accepted by humans as valuable [5,14]. But the structures generated within newly transformed spaces will need types of evaluation different (at least in part) from those implicit within the original space, or previously provided in explicit form.

It is even more difficult to express (verbally or computationally) just what it is that we like about a Bach fugue, or an impressionist painting, than it is to recognize something as an acceptable member of one of those categories. And to say what it is that we like (or even dislike) about a new, or previously unfamiliar, form of music or painting is even more challenging.

Identifying the criteria we use in our evaluations is hard enough. Justifying, or even (causally) explaining, our reliance on those criteria is more difficult still. For example, just why we like or dislike something will often have a lot to do with motivational and emotional factors—considerations about which current AI has almost nothing to say.

To make matters worse, human values—and therefore the novelties which we are prepared to approve as "creative"—change from culture to culture, and from time to time. In some cases, they do so in unpredictable and irrational ways: think of the fashion-industry, for example, or of rogue memes like the back-to-front baseball-cap. Nor are value-shifts confined to trivial cases such as these: even Bach, Mozart, and Donne were ignored and/or criticized in certain periods.

The scientific criteria of theoretical elegance and coherence, and of experimental verification, are less variable than artistic values. But that's not to say they are easy to

define, or to implement. (An attempt to do so, for certain sorts of mathematical symmetry, has been made by the BACON team.)

Moreover, science too has its equivalent of fad and fashion. Even the discovery of dinosaurs was not a cut-and-dried event, but the culmination of a process of scientific—and political-nationalistic—negotiation lasting for several years [20]. The important point is that what scientists count as "creative", and what they call a "discovery", depends largely on unarticulated values, including social considerations of various kinds. These social evaluations are often invisible to scientists. For sure, they are not represented in AI-models.

5. Conclusion

Some H-creative ideas have already been generated by AI-programs, though usually by merely exploratory (or combinational) procedures. Transformational AI-originality is only just beginning.

The two major bottlenecks are:
(1) domain-expertise, which is required for mapping the conceptual space that is to be explored and/or transformed; and
(2) valuation of the results, which is especially necessary—and especially difficult—for transformational programs.

These two bottlenecks interact, since subtle valuation requires considerable domain expertise. Valuation, thus far, is mostly implicit in the generative procedures used by the program, or interactively imposed by a human being. Only a few AI-models can critically judge their own original ideas. And hardly any can combine evaluation with transformation.

The ultimate vindication of AI-creativity would be a program that generated novel ideas which initially perplexed or even repelled us, but which was able to persuade us that they were indeed valuable. We are a very long way from that.

References

[1] K. Binsted, Machine humour: an implemented model of puns, Ph.D. Thesis, University of Edinburgh, 1996.
[2] M.A. Boden, The Creative Mind: Myths and Mechanisms, Basic Books, New York, 1990.
[3] M.A. Boden (Ed.), Dimensions of Creativity, MIT Press, Cambridge, MA, 1994.
[4] D. Cliff, I. Harvey, P. Husbands, Explorations in evolutionary robotics, Adaptive Behavior 2 (1993) 71–108.
[5] H. Cohen, The further exploits of AARON, painter, in: S. Franchi, G. Guzeldere (Eds.), Constructions of the Mind: Artificial Intelligence and the Humanities, Special edition of Stanford Humanities Review 4 (2) (1995) pp. 141–160.
[6] D. Cope, Computers and Musical Style, Oxford University Press, Oxford, 1991.
[7] K.D. Forbus, D. Gentner, K. Law, MAC/FAC: A model of similarity-based retrieval, Cognitive Science 119 (1994) 141–205.
[8] K.D. Forbus, D. Gentner, A.B. Markman, R.W. Ferguson, Analogy just looks like high level perception: why a domain-general approach to analogical mapping is right, Journal of Experimental and Theoretical AI, in press.
[9] D. Gentner, S. Brem, R.W. Ferguson, A.B. Markman, B.B. Levidow, P. Wolff, K.D. Forbus, Conceptual change via analogical reasoning: a case study of Johannes Kepler, Journal of the Learning Sciences, in press.

[10] P. Hodgson, Modelling cognition in creative musical improvisation, Ph.D. Thesis, University of Sussex, in preparation.

[11] D.R. Hofstadter, FARG (The fluid analogies research group), Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought, Basic Books, New York, 1995.

[12] K.J. Holyoak, P. Thagard, Analogical mapping by constraint satisfaction, Cognitive Science 13 (1989) 295–355.

[13] K.J. Holyoak, P. Thagard, Mental Leaps: Analogy in Creative Thought, MIT Press, Cambridge, MA, 1995.

[14] H. Koning, J. Eizenberg, The language of the prairie: Frank Lloyd Wright's prairie houses, Environment and Planning B 8 (1981) 295–323.

[15] P. Langley, H.A. Simon, G.L. Bradshaw, J.M. Zytkow, Scientific Discovery: Computational Explorations of the Creative Process, MIT Press, Cambridge, MA, 1987.

[16] D.B. Lenat, The role of heuristics in learning by discovery: three case studies, in: R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), Machine Learning: An Artificial Intelligence Approach, Tioga, Palo Alto, CA, 1983, pp. 243–306.

[17] P. McCorduck, Aaron's Code, W.H. Freeman, San Francisco, CA, 1991.

[18] M. Mitchell, Analogy-Making as Perception, MIT Press, Cambridge, MA, 1993.

[19] T.S. Ray, An approach to the synthesis of life, in: C.G. Langton, C. Taylor, J. Doyne Farmer, S. Rasmussen (Eds.), Artificial Life II, Addison Wesley, Redwood City, CA, 1992, pp. 371–408. Reprinted in: M.A. Boden (Ed.), The Philosophy of Artificial Life, Oxford University Press, Oxford, 1996, pp. 111–145.

[20] S. Schaffer, Making up discovery, in: M.A. Boden (Ed.), Dimensions of Creativity, MIT Press, Cambridge, MA, 1994, pp. 13–51.

[21] K. Sims, Artificial evolution for computer graphics, Computer Graphics 25 (4) (1991) 319–328.

[22] S. Todd, W. Latham, Evolutionary Art and Computers, Academic Press, London, 1992.