Lakatos and Machine Creativity

Alison Pease[1] and **Simon Colton**[2] and **Alan Smaill**[3] and **John Lee**[4]

Abstract. We argue that Lakatos' work on the history and philosophy of mathematics is of key relevance to machine creativity as it suggests ways in which to explore and transform concept spaces, re-represent knowledge and change evaluation criteria. We describe approaches to implementing methods which Lakatos identifies, including our own approach, which extends Colton's HR and has enabled us to automatically generate mathematical conjectures, concepts and examples which were previously impossible in HR - including Goldbach's conjecture. The methods are of general importance as they can be applied to many domains - we describe their theoretical application to game plans, two-dimensional geometry, moral philosophy, philosophy of mind, political argument and meta-level reasoning.

1 INTRODUCTION

Our thesis in this paper is that *Lakatos's* Proofs and Refutations [17] *has important and exciting implications for the field of machine creativity*. This book provides a rational reconstruction of the development of the proof of Euler's conjecture, in which new concepts, conjectures and 'proofs' are invented. It spans 200 years of in-depth development in this area, describing methods which were used to generate and evaluate ideas, and the evolution of the methods themselves, thus operating on both object and meta-level. We argue that the methods which Lakatos identifies:
(i) are extremely relevant to machine creativity (key issues in creativity research are described in §2 and Lakatos' methods in §3);
(ii) can be automated (§4), and
(iii) apply to other domains (§5).
If this is the case then they will be powerful tools in machine creativity.

Our motivation is the cognitive scientific goal of understanding *intelligent action*, aiming to elucidate creativity by modelling the external process of discovery via interaction. This differs from trying to understand the *mind* by modelling internal cognitive processes which lead to creative output. A secondary motivation is to develop new techniques of general use in AI. If successfully automated, Lakatos's methods will be relevant to concept formation, representation and modelling interaction in mathematics.

[1] University of Edinburgh, Division of Informatics, 80 South Bridge, Edinburgh, EH1 1HN, Scotland, email: alisonp@dai.ed.ac.uk
[2] University of Edinburgh, Division of Informatics, 80 South Bridge, Edinburgh, EH1 1HN, Scotland, email: simonco@dai.ed.ac.uk
[3] University of Edinburgh, Division of Informatics, 80 South Bridge, Edinburgh, EH1 1HN, Scotland, email: smaill@dcs.ed.ac.uk
[4] University of Edinburgh, Division of Informatics, 2 Buccleuch Place, Edinburgh, EH8 9LW, Scotland, email: john@cogsci.ed.ac.uk

2 MACHINE CREATIVITY

2.1 Exploration and Transformation

Boden's theory of creativity [3], in which she aims both to state what it is, and suggest how it is possible, has had a major impact on the field of machine creativity. She uses the metaphor of a concept space which is mapped, explored, and transformed (METCS) to provide an account of novelty. An item in a little explored area but still within the boundaries is 'merely' novel, and outside but close to the boundaries 'fundamentally' novel. While considering both types of novelty to be potentially creative (subject to a value criterion), she regards exploratory creativity to be less creative than transformational creativity. Criticisms have been levelled in broadly two areas; *(i)* questioning Boden's analysis of the role that each aspect of METCS plays in creativity (for example Bundy's claim that working *within* a mathematical space may be more creative than changing the space [5]), and *(ii)* questioning the value of the concept space metaphor, i.e. whether it is always possible to define concept spaces (for example it is difficult to state which rules define theoretical boundaries of natural language). Boden acknowledges the criticism and in her later writings the status of exploratory creativity is higher, going from "mere novelty" to a kind of creativity which is "not to be sneezed at" [2]. She suggests that the answer to the problem of defining concept spaces is to identify a range of putative concept spaces and use them to evaluate the worth of her analogy [1].

2.2 Re-Representation

An important way in which a domain may be explored or transformed is by re-representing items within it. Kuhn[15] wrote that a paradigm shift (or transformation), in which we perceive something in a totally different way, is one of the main stages in the cycle of scientific development. James[13] (cited in [11]) wrote that we represent an object according to the property which is most useful to us; if that changes then we need to change our representation. For example I would normally see a newspaper as essentially for reading, however if my principle need is warmth then its key feature becomes its flammability. This is an example of *radical reshaping*, one of the factors identified by Hofstadter[11] which influence the way in which we perceive something; other factors are *belief* - who do I think wrote it?, *goals* - why am I reading it? and *external context* - what is its immediate environment?

Karmiloff-Smith[14], referring to Boden's METCS metaphor, proposes that the ability to explore a domain comes after a process of representational redescription, in which implicit knowledge is initially acquired. Key details are then abstracted and the knowledge redescribed - this is now explicit knowledge and inter-domain connections may be made. This process continues with each stage building on the last, and the knowledge being made more explicit and flexible

(for example imagining counterfactuals). All levels of knowledge are retained for efficiency in any given situation, where creativity results from declarative knowledge (which may be *explored*), as opposed to procedural knowledge (which may only be *used*). (She supports her argument with results of experiments in which children of different ages were asked to draw 'funny' houses or people. Younger children (4 - 6 years old) were unable to deviate from the normal pictures, and any differences consisted of adding elements to a completed drawing, whereas the drawings of the older children (8 - 10 years old) showed much more variation.)

Boden also considers the importance of the process of representation in creativity, in [3]. A new representation may be a helpful or even necessary step in a creative solution. Consider the sequence $2, 1, 2, 2, 2, 3, 2, ...$ [11, p. 16]. What is the next number? Now consider the sequence represented below[5]. An answer is now much easier to find, hence the step from the first to the second representation may be creative. As an example Boden cites Niels Bohr's representation of the atom as a solar system, which suggested questions about orbits of electrons, leading to the discovery of new and useful knowledge.

(Re)representation is a long standing problem in AI, with much work carried out, eg. [19]. It consists of two problems; *(i)* how to generate different representations, and *(ii)* how to choose between them.

2.3 Dynamic evaluation criteria

Many theories of creativity are based on a 2-stage model of *generation* and *evaluation*, referred to in [19] as the 'central loop of creativity'. These stages are thought to run concurrently or cyclically rather than sequentially. While Boden expounds her theory of novelty (§2.1) in depth she has not developed a theory of value (i.e. an evaluation stage). A value criterion is clearly necessary in evaluating creativity and Boden emphasises this point. However she believes that value is not definable in scientific terms, nor constant (being influenced by unpredictable factors such as nationality, fashion, rivalry and commercialism). Certainly creative work which is historically new (Boden's h-creativity[3]) by definition cannot be subject to familiar criteria. As pointed out in [22], the ability to measure quality in a field without mistakes would imply that that field was incapable of any further transformation. This is reflected by the large number of examples of work which was not valued at the time it was produced, for example Van Gogh's paintings, group theory, or immunisation. The area of dynamic evaluation criteria then, while little developed and requiring much further research, is essential to theories of creativity.

2.4 Key issues in machine creativity

In view of the research above we are interested in:

- identifying a putative concept space;
- exploration (object-level) of this concept space;
- transformation (meta-level) of this concept space;
- determining *(i)* ways in which knowledge may be re-represented, and motivations for doing so, and *(ii)* ways in which to use the new representation to further explore or transform a domain, and
- developing an account of dynamic evaluation criteria.

We show how Lakatos' methods [17] suggest ways in which to achieve these goals.

[5] $(2, 1), (2, 2), (2, 3), (2, ...$

3 LAKATOS-STYLE REASONING

Figure 1. Examples of polyhedra and their Euler characteristic $V - E + F$

Lakatos describes the evolution of the conjecture and proof that for all polyhedra, the number of vertices (V) minus the number of edges (E) plus the number of faces (F) equals 2. It is presented as a discussion between a group of students and their teacher. The initial conjecture (C) is found by *induction*; i.e. since C holds for all regular polyhedra (eg. the cube in Figure 2) the class guess that it might hold for all polyhedra. They start with Cauchy's 'proof' below [17, p. 7]:

(1) remove one face of the polyhedron and stretch it flat onto a blackboard. (If C holds, we now have $V - E + F = 1$);
(2) cut all polygons on the board into triangles by drawing more edges. ($V - E + F$ stays constant as each new edge forms a new face);
(3) remove each triangle on the board one by one. ($V - E + F$ continues to remain constant since we either remove 2 edges, a vertex and a face, or one edge and one face). At the end we are left with a single triangle, for which $V - E + F = 3 - 3 + 1 = 1$.

The methods below are summarised in Table 1.

3.1 Exploring the polyhedra domain

Figure 2. Counterexamples to the conjecture that for all polyhedra, $V - E + F$

Counterexamples are soon found, such as the hollow cube, and some students use these to reject C (the method of *surrender*). This is presented as a naive reaction, in which many valuable ideas may be lost by subjecting conjectures to overly harsh judgement too soon. Another reaction is to modify concept definitions to exclude the counterexamples (the method of *monster-barring*). Counterexamples are seen as monsters, which should not be allowed to upstage a theorem which brings harmony to the field. With the emergence of more and more counterexamples the definition of polyhedron shrinks from *a solid whose surface consists of polygonal faces* to *a surface consisting of a system of polygons* (thus excluding the hollow cube) to *a system of polygons arranged in such a way that (1) exactly two polygons meet at every edge and (2) it is possible to get from the inside of any polygon to the inside of any other polygon by a route which never crosses any edge at a vertex* (to exclude the twin polyhedron). More counterexamples lead to arguments over the meaning of terms in the definitions, and *polygon*, *area*, and *edge* are further defined, with the strictest definition being taken in each case.

Rather than see counterexamples as monsters and modify concept definitions to exclude them, one student suggests that they are valid examples of polyhedra but can be seen as exceptions, with the appropriate reaction being to modify the conjecture to exclude them. This can be done in two ways. Those properties which counterexamples share and positive examples lack should be found and then either a whole class or particular examples excluded; such as generalising from hollow cubes to *polyhedra with cavities*, the picture frame to *polyhedra with tunnels* and twin tetrahedra to *polyhedra with multiple structure*, and modifying C to 'for any polyhedra without cavities, tunnels or multiple structure, $V - E + F = 2$' (the method of *exception-barring by piecemeal exclusion*). Alternatively since the class cannot be certain to have listed all the exceptions (above) some students advocate withdrawing into a much smaller domain for which C seems certain to hold. This may be done by looking at the properties which all positive examples share and counterexamples lack. For instance all the positive (and none of the negative) examples are *convex*. The domain of application is then restricted accordingly, i.e. C becomes 'for any convex polyhedra, $V-E+F = 2$' (the method of *exception-barring by strategic withdrawal*).

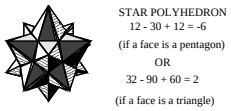

Figure 3. Different representations of the star polyhedron

Another alternative to barring a monster or counterexample is to bar a 'monstrous interpretation'. That is, reinterpret (or re-represent) a counterexample so that it is no longer a counterexample. For example the star-polyhedron in Figure 3 is only a counterexample if it is thought to have 12 star-pentagon faces, forming 12 vertices and 30 edges. It can also be interpreted as having 60 triangular faces, forming 32 vertices and 90 edges - in which case C holds. This is the method of *monster-adjusting*, and leads to analysis of the concept *face*.

3.2 Using the proof - a dynamic evaluation criteria

At this point the teacher complains that the students' concept and conjecture modifications will always be *ad hoc* and further counterexamples may potentially exist even if they cannot be found. Mathematical certainty, therefore, cannot be achieved using these methods. Additionally they have yet to refer to the proof initially offered. The objection is that the students have no way of evaluating their conjecture other than by reference to known examples (or counterexamples). The 'proof' is one way of evaluating a conjecture, where a convincing proof suggests a valuable conjecture (clearly there are other criteria such as non-triviality too). Lakatos calls the interplay between conjecture and proof "the intrinsic unity between the 'logic of discovery' and the 'logic of justification' " [17, p. 37]. A new conjecture gives rise to a new proof, or, in machine creativity terms, dynamic evaluation criteria. Similarly, new evaluation criteria suggest new results. The teacher distinguishes between global and local counterexamples, where a global counterexample is one which violates the conjecture (so the conjecture is flawed) and a local counterexample violates a step of the proof but not the conjecture (a criticism of the proof but not the conjecture). When a counterexample is found it should be tested to see which type it is (it may be both).

The method of *explicit lemma-incorporation* consists of finding which step of a faulty proof a global counterexample violates and then making that step a condition of the conjecture. For example all of the counterexamples found above violate the first step of the proof since they cannot be stretched flat onto a blackboard. Therefore C becomes 'for any polyhedra which, by removing one face can be stretched flat onto a blackboard, $V - E + F = 2$' thereby upholding the proof but reducing the domain of the main conjecture to the very domain of the guilty lemma (proof step). It may be the case that a counterexample to the conjecture appears not to violate any of the proof steps, for example a cylinder is a counterexample to C but does not obviously violate any lemmas. However it does violate the hidden assumption in lemma 1 that the result of stretching the polyhedron on the board will be a connected network. The method of *hidden lemma-incorporation* is where a global counterexample is used to identify hidden assumptions within a step which are violated, and then to make the assumption a condition. So this assumption is added to the proof - thus modifying the evaluation criteria - and made a condition, i.e. C becomes 'for any polyhedra which, by removing one face can be stretched flat onto a blackboard and the resulting network be connected, $V - E + F = 2$'.

The method of *proofs and refutations* provides a way of generating counterexamples by analysing proof steps, as well as improving a proof by use of counterexamples. Counterexamples which violate particular steps are sought, such as the picture frame in Figure 2 which the students discovered by looking for a polyhedron which after having a face removed could not be stretched flat onto a plane. When a counterexample is found it should be ascertained whether it is local, global or both. If it is local then the lemma should be modified to exclude the counterexample, and if it is global then either the method of *explicit lemma incorporation* (if it is also local), or *hidden lemma incorporation* (if it is not local) should be used.

3.3 How is Lakatos relevant to machine creativity?

In §2.4 we identified issues within machine creativity of key interest. We now show how Lakatos' work is relevant to these issues.

Identifying a putative concept space - Lakatos' work is in the mathematical domain which is a putative concept space since the boundaries of its concept spaces are relatively clearly defined by axioms, rules of inference, meta-mathematical beliefs and concept definitions.

Exploration of this concept space - Using Lakatos' methods of *induction*, *surrender*, *monster-barring*, *piecemeal exclusion*, *strategic withdrawal*, *monster-adjusting* and *lemma incorporation* the students have built and explored a theory of polyhedra, containing:

• *conjectures* - 'for all polyhedra, $V - E + F = 2$'; 'for any polyhedra without cavities, tunnels or multiple structure, $V - E + F = 2$'; 'for any convex polyhedra, $V - E + F = 2$'; 'for any polyhedra which, by removing one face can be stretched flat onto a blackboard and the resulting network be connected, $V - E + F = 2$';
• *concepts* - regular polyhedra; a solid whose surface consists of polygonal faces; a surface consisting of a system of polygons; a system of polygons arranged in such a way that (1) exactly two polygons meet at every edge and (2) it is possible to get from the inside of any polygon to the inside of any other polygon by a route which never crosses any edge at a vertex; polygon; area; edge; polyhedra with

Method	Description
Induction	Generalise from particulars
Surrender	Look for counterexamples to a conjecture C and use them to refute it
Monster barring	Modify a concept definition so as to exclude an unwanted counterexample
Piecemeal exclusion	Find those properties which make a counterexample fail C and then modify C by excluding that type of counterexample
Strategic withdrawal	Consider the examples for which C does hold, generalise and limit C to that type of example
Monster adjusting	Reinterpret a counterexample so that it no longer violates C
Lemma incorporation	Given a *global* counterexample, find which step of the proof it violates and then modify C by making that step a condition. Given a *local* counterexample (which violates a proof step but not C), look for a hidden assumption in the proof step, then modify the proof and C by making the assumption an explicit condition.
Proofs and refutations	Use the proof steps to suggest counterexamples. For any counterexamples found, test whether they are local or global counterexamples and perform lemma incorporation

Table 1. Summary of Lakatos' methods

cavities; polyhedra with tunnels; polyhedra with multiple structure; convex polyhedra; face; and

• *examples/counterexamples* - the regular polyhedra (tetrahedron; cube; octahedron; icosahedron and dodecahedron); hollow cube; twin tetrahedra; picture frame; star polyhedron; and cylinder.

Transformation of this concept space - much of the dialogue in [17] is about how mathematics should be done, for instance what the role of proof is, how conjectures should be modified, the value of searching for counterexamples, which of the methods are preferable, etc. In §5.6 we discuss how the methods themselves could help determine which method is most appropriate.

Re-representing knowledge and using the new representation - *monster-adjusting* has suggested ways in which to re-represent items, motivations for doing so (in order to preserve a conjecture which is thought valuable), and ways to use the new representation to further explore the domain (by analysing subconcepts inherent in a more encompassing concept).

Developing an account of dynamic evaluation criteria - the interplay shown in [17] between generation and evaluation is one of the most exciting aspects of Lakatos' work. Conjectures (and concepts and examples/counterexamples) are generated using the methods described and evaluated by the 'proof'. That is, 'proofs' can be seen as ways of evaluating the conjectures, since if a proof - or strong argument - can be found then the conjecture is highly valued as it may be a theorem. Indirectly the concepts within the conjectures are also evaluated by a proof if concepts within theorems are considered useful. Although Lakatos does not describe how to initially generate a proof (leaving this to [23]), he *does* - in his *lemma incorporation* method describe how it can be modified, thus forming dynamic evaluation criteria.

4 LAKATOS'S METHODS CAN BE AUTOMATED

4.1 The HR program

To show that Lakatos-style reasoning can be automated, we implemented such methods in the HR automated theory formation program [6]. HR is given background information about a domain, including some objects of interest and some initial concepts, supplied with a definition and examples. For instance, in number theory, the objects of interest are integers, and the initial concepts include multiplication and addition. HR forms new concepts by using one of 10 general production rules to transform one (or two) old concepts into a new one. For example, to construct the concept of prime numbers (with exactly two divisors), HR passes the concept of integers a and b for which b divides a, through the size production rule, to produce the function: $f(a) = |\{b : b \text{ divides } a\}|$, i.e., $f(a)$ is the number of divisors of a. Following this, HR uses the split production to produce the concept of integers a for which $f(a) = 2$, i.e., integers which have exactly two divisors (prime numbers).

Using the objects of interest to provide data, HR also builds sets of examples for each concept and the examples are used to make conjectures empirically. In particular, if HR finds that the examples of one concept are all examples of another, it makes the implication conjecture that the definition of the first implies the definition of the second. In addition to forming concepts and conjectures, HR also uses the Otter theorem prover and others to prove theorems, and the MACE model generator to find counterexamples to non-theorems. HR's approach can be characterised as concept-driven, i.e., conjectures are made in response to the invention of new concepts. However, in addition to advocating the social nature of discovery, Lakatos also suggests a conjecture-driven approach, where concepts are introduced in response to the discovery of a conjecture. Hence we improved HR's model of theory formation to enable:

- Production of conjectures with known counterexamples. Previously, only conjectures true of *all* examples were made.
- Analysis of faulty conjectures to suggest theory formation steps which invent concepts that fix the conjecture statement.
- A multi-agent approach, with agents able to request and communicate concepts, conjectures, proofs and counterexamples.

We implemented two ways to adjust faulty conjectures, inspired by Lakatos's exception barring methods. Firstly, if HR makes a conjecture for which there are a small number of counterexamples (with the number specified by the user), it will invent a concept with a definition which excludes the counterexamples. For instance, given the numbers 1 to 10, HR makes the conjecture that all odd numbers are prime, with 1 and 9 as counterexamples. It then invents the concept of odd numbers except 1 and 9, in order to make the conjecture that odd numbers except 1 and 9 are prime. Of course, if HR was given the numbers 1 to 30, the counterexamples would be $1, 9, 15, 21, 25$ and 27, and it is likely that HR would reject the conjecture.

Secondly, if HR makes a conjecture which states, say, $P(X)$ implies $Q(X)$ and there are a sizeable number of counterexamples (again specified by the user), HR will try to find a concept stating a property $R(X)$ which is true of all the counterexamples and no others. If successful, it will invent the concept of objects X for which

$P(X)\&-Q(X)$, which will fix the conjecture. For example, when HR works with the numbers 1 to 30, it makes the conjecture that all integers have an even number of divisors, with $1, 4, 9, 16$ and 25 as counterexamples. It then identifies that square numbers have exactly these examples, and alters the conjecture to be: all integers except square numbers have an even number of divisors.

With the multi-agent version of HR set up such that each agent asks the others, rather than MACE to generate counterexamples, the following scenario occurs: the agents are given different sets of objects of interest, and one of them makes a conjecture for which it has no counterexample. It then communicates this to the other agents, who reply with suitable counterexamples from their theory. The first agent can then reject the conjecture or attempt to drive the theory formation towards concepts which fix the conjecture. This is analogous to Lakatos's polyhedra example discussed above, and we claim to have implemented Lakatos-style reasoning, demonstrated further by the following two number theory sessions with HR.

4.2 Case study: number theory sessions

Number theory is ideal for testing HR's new abilities because many well-known theorems appear to require Lakatos's methods. For instance, the fundamental theorem of arithmetic states that all integers *except 1* are uniquely expressible as a product of primes. Similarly, Fermat's Last Theorem states that the equation: $x^n + y^n = z^n$ has no integer solutions x, y and z for any n *except 1 and 2*.

In the first session with HR, we wished to demonstrate the multi-agent capabilities. Two agents were run in number theory, the first with the numbers 1 to 20 and the second with 20 to 40. They were allowed to request and communicate counterexamples to each other whenever they made an implication conjecture, and to use counterexample-barring to fix any conjecture with 1 or 2 counterexamples. As an example of their interaction, the second agent invented the concept of odd numbers, and, later, prime numbers. It then made the conjecture that prime numbers are odd, and could find no counterexamples between 20 and 40. This conjecture was passed to the first agent as a set of theory formation steps to carry out in order to invent the concepts of primes and odds, followed by the conjecture that the prime implies odd. The first agent realised that the concepts it required were already in its theory, and found only the number 2 as a counterexample to the conjecture. On receiving this, the first agent added three theory formation steps to the agenda, to invent the concepts of: (a) numbers which equal two (b) numbers which don't equal two and (c) primes which don't equal two. On inventing this concept, it made the conjecture that primes except two are odd, which fixed the conjecture, making it true. In the same session, they made many conjectures which aren't possible without Lakatos's techniques, such as: all refactorable numbers except 1 and 9 are even (refactorable numbers are such that the number of divisors is itself a divisor). This conjecture is false: the next odd refactorable number is 225.

In the second session, we wished to demonstrate that Lakatos-style reasoning can be used by HR to find important conjectures. We started a single version of HR with the numbers 1 to 10 and the initial concepts of divisors and addition. We specified that HR should attempt to fix implication conjectures by excepting counterexamples if there was only one. The size and split rules were used to turn the concept of divisors into the concept of prime numbers, as described above, and the split rule re-invented the concept of even numbers: integers divisible by 2. The compose rule combines the definitions of two concepts and this was used twice with the concepts of addition and prime numbers. In this way, HR invented the concept of integers which can be written as a sum of two prime numbers (which is true of the numbers 4 to 10). When it invented this concept, HR made the conjecture that even numbers have this property, but found the number 2 to be a counterexample. Thus HR invented the concept of even integers except 2, which fixed the conjecture. That is, HR made the conjecture that all even numbers except 2 are expressible as the sum of two primes. This is Goldbach's famous conjecture which, although discovered in 1742, is still unproved. Without Lakatos' techniques HR would not have been able to generate it. To our knowledge, this is the first time Lakatos's techniques have been implemented and used to re-discover an important mathematical conjecture.

4.3 Other Approaches

The agency described above is one approach to modelling Lakatos's methods. There are other approaches, both practical and theoretical. We do not argue that our approach is better than those described in this section, rather that the other approaches in AI help to support our argument that the methods may be implemented.

Hayes-Roth[9] describes 5 heuristics for repairing flawed beliefs, which are based on Lakatos' methods and have been partially implemented. He considers the card game Hearts (like Whist), in which the pack is divided amongst players. One player plays a card and the others must all put down a card in the same suit as the first if they have one and otherwise play any card. The person who played the highest card in the specified suit wins that trick and starts the next. One point is awarded for each heart won in a trick, and 13 for the queen of spades (QS). The aim of the game is to get either as few points as possible ("go low") or all the points ("shoot the moon"). A strategy which beginners sometimes employ is to win a trick to take the lead and then play a spade, in order to flush out the QS and avoid the 13 points. Hayes-Roth represents this as shown (p.230):

Plan:	Flush the QS
Effects:	(1) I will force the player who has the QS to play that card.
	(2) I will avoid taking 13 points
Conditions:	(1) I do not hold the QS
	(2) The QS has not yet been played
Actions:	First I win a trick to take the lead and whenever I lead I play a spade.

The plan (analogous to a faulty conjecture) may backfire if the beginner starts with the king of spades (KS) and then wins the trick and hence the unwanted points (this situation is a counterexample to the plan). The heuristics then provide various ways of revising the plan:

1) **Retraction** (like surrender) - retract the part of the plan which fails, in this case effect (2).

2) **Exclusion** (like monster-barring) - bar the theory from applying to the current situation, by excluding the situation. Add the condition *I do not play KS*.

3) **Avoidance** (like piecemeal withdrawal) - rule out situations which can be predicted to fail the plan, by adding conditions to exclude them. For example by assessing why the plan failed add the condition *I do not win the trick in which the queen of spades is played.* A system can further improve its plan by negating the new condition - *I win the trick in which the queen of spades is played*, using this and its knowledge of the game to infer that it must play the highest card in the specified suit, and then negating the inference to get *I must not play the highest card in the specified suit*. This is then incorporated into the action which becomes *First I win a trick to take the lead and whenever I lead, I play a spade which is not the highest spade.*

4) **Assurance** (like strategic withdrawal) - change the plan so that it only applies to situations which it reliably predicts. In this case the faulty prediction is effect (2), and so the system looks for conditions which guarantee it. It does this by negating it, inferring consequents and then negating one of these and incorporating it into the action.. For example negating effect (2) gives *I do take 13 points*, the game rules state that *the winner of the trick takes the points in the trick* so we can infer that *I win the trick*, then use this and the rule that *the person who plays the highest card in the suit led wins the trick* to infer that *i play the highest card in the suit led*. Given that *player X plays the QS* we can now infer that *I play a spade higher than the QS* and negate it to get *I play a spade lower than the QS.*
5) **Inclusion** (also like strategic withdrawal) - this differs from assurance in that the situations for which the plan is known to hold are listed rather than a new concept being devised. Therefore instead of adding *I play a spade lower than the QS* to the action, we add *I play a spade in* $\{2S, ..., 10, JackS\}$.

Hayes-Roth argues that these heuristics can be implemented using existing techniques (although adding that this may take considerable effort). The primary capabilities, he claims, are symbolic deduction and heuristic search. He suggests ways in which one heuristic may be preferred over another, in order to avoid a combinatorial explosion. These include preferring general to specific theories,seeking canonical representations and experimentally evaluating alternative fixes to determine the most fruitful.

Rissland[24] has carried out much work on the role of examples in understanding a domain, in particular within mathematics. She has implemented a system partly based on Lakatos' ideas on examples and counterexamples, ExGen, which can generate examples which meet specified properties.

Finally Bundy[6] has suggested other ways, such as using neural networks or version spaces[10], trained on positive and negative examples, to model monster-barring and exception-barring. These would contain a grey area in which an object would not be categorised (thus giving the informality needed). Bundy also suggests that proof analysis[4], in which a failed proof is analysed with a counterexample to see at which point it fails, is of use for lemma-incorporation. It could also be used to generate new counterexamples (the method of proofs and refutations).

5 LAKATOS' METHODS APPLY TO OTHER DOMAINS

If Lakatos's methods are to be of general use in machine creativity, we need to show that they apply to domains other than three-dimensional geometry. Theoretical applicability of the methods to other domains would mean that programs which model the methods may also work in these domains.

Multi-domain applicability is desirable as programs then avoid the accusation of fine-tuning (where a program uses procedures to generate specific desired output, but the procedures do not produce anything else of value). The degree of fine-tuning in a program modelling creativity was formally defined in [8] and argued to be relevant to the judgement of creativity. Writing the program for one domain and then applying it to another gives greater validity to any valuable output. We have already shown (§4) how the methods apply to number theory; in this section we describe how they may be used to retrieve concepts, conjectures and counterexamples in other domains, in particular within philosophy. We do not claim that this was how such results came about, nor that they occurred in that order - but if the methods can be used to retrieve a significant number of important results in a field then we can say that they may be usefully applied to that field. We have not yet implemented any of these ideas, but describe where we intend to do so and briefly detail how.

[6] Alan Bundy – personal communication, BBN's 49, 1182, 1212.

5.1 Game plans

In §4 we describe how Hayes-Roth[9] applied the methods to game plans. The plan *if I do not hold the QS and it has not yet been played and I win a trick and then play a spade, then the outcome will be to force the player with the QS to play it and I will avoid taking 13 points* is shown to be flawed by the situation occurring in which I play the KS but end up with the 13 points. Various ways of revising the plan were shown, and their relation to Lakatos' methods.

5.2 Two-Dimensional Geometry

A theoretical case study in [21] describes how we could implement the methods in two-dimensional geometry using HR. An agent uses *induction* on examples of squares and triangles to conjecture that for all shapes, the number of vertices (V) equals the number of edges (E). A second agent sends $S3$ below, which has 5 vertices and 6 edges. The first uses *monster-adjustment*, re-representing it as having 4 vertices (excluding the middle one) and 4 edges. The second then generates $S4$, which has 5 vertices and 6 edges. In an effort to save the conjecture they generate the concept polygon - a shape in which all vertices touch exactly two edges. This includes all examples and excludes all counterexamples (including the different interpretations of $S3$). *Exception-barring* is then used to modify the conjecture to 'for all polygons, $V = E$'.

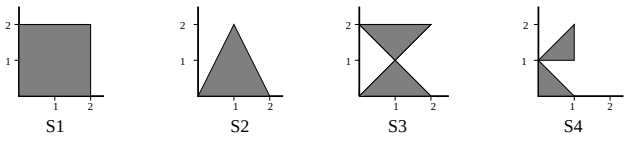

Figure 4. Example shapes we hope to generate in 2D geometry

5.3 Moral Philosophy

Lakatos' methods can be used to generate concepts, conjectures and examples from undergraduate philosophy textbooks. Utilitarianism, the principle that an action is right insofar as it tends to maximise happiness [20], is a standard position in normative ethics. Our universe of interest is now actions and the two initial concepts those actions which *maximise happiness* and those which are *good.*

Suppose that in an agent's data base all actions which maximise happiness are categorised as good. Therefore by *induction*, it arrives at the conjecture C that *all actions which maximise happiness are good* [20], and sends this to the other agents. They look for a counterexample, i.e. an action which maximises happiness but is not good. The act of breaking promises, for example if X owes Y ten pounds but gives it to Z because Z needs it more (x_1)[20], is found and the method of *surrender* used to claim that $\neg C$ [27]. Monster-barring might be used to redefine the concept *actions which maximise happiness* as *actions which, if practised generally, would maximise happiness* [20]. Since (it can be argued) the traditional practice of keeping

promises is one which maximises happiness, the new concept excludes x_1. C' is now *All actions which, if practised generally, would maximise happiness, are good.* This is rule-based utilitarianism [20]. Again agents search for examples of actions which satisfy the first concept but not the second. McCloskey[18] considers a situation in which the sheriff of a small town can prevent mass riots, in which hundreds of people will be killed, only by framing and executing an innocent man. If the action to be practised generally is minimising human suffering, then the sheriff should do so, thus satisfying the first concept but possibly failing on the second as it can be seen as wrong. *Monster-adjusting* might be used to claim that if seen if a different way the action is good. If it is accepted that there are only those two options specified, then either option is morally reprehensible and the sheriff's action of framing and executing the innocent man is the lesser of two evils [26]. Alternatively the concept of *unjust actions* (discussed in [20]) might be formed by doing *exception barring (piecemeal exclusion)* on the counterexamples. The conjecture *All actions (except unjust actions) which maximise happiness, are good* might then be formed (not a stance taken in the philosophical literature as it weakens the conjecture too much).

We might use Mill's 'proof' of the utilitarian principle to form more concepts and conjectures, and to find more counterexamples: 1) happiness is our sole ultimate end, 2) promotion of human ends is the test by which we judge human conduct, therefore 3) all actions which maximise happiness are good.
The search for a counterexample which would violate the first lemma (*proofs and refutations*) might lead to the example of George who is a pacifist but is offered a job researching chemical warfare. He desperately needs a job, and could slow down research if he took it, but would have to go against his principles (x_4)[27]. Williams claims that it would be wrong for George to take the job since this would lack personal integrity. By this he means that the link between commitments and actions would be broken. Personal integrity is, he claims, an end in itself. Whether this refutes the conjecture as well as the first claim (i.e. is a global or local counterexample) is controversial (doesn't personal integrity in general maximise happiness?). Since the claim is explicitly stated in the lemma, *lemma incorporation* might be used on x_4 to form the concept *actions in which happiness is the only end concerned,* and conjecture *all actions in which happiness is the only end concerned, which maximise happiness, are good.* Finally an agent might find the example of capital punishment for murderers which may maximise happiness (the relatives of the victims might feel happier) but not be considered good. *Hidden lemma incorporation* could be used to find the hidden assumption that all reasons behind happiness are morally equivalent, i.e. someone's revengeful glee is morally equivalent to the satisfaction Mother Teresa gets from helping the needy (the distinction between legitimate and illegitimate desires is made in [16]). This could be made explicit in the argument, i.e. the first premise becomes: 1) happiness is our sole ultimate end, and all reasons for happiness are morally equivalent.

We intend to work in this domain using a datafile *Moral Reasoning* which is currently employed by machine learning programs. It contains information about 200 people, attributes including whether or not they caused harm, whether or not they had a plan, if they were careful and whether they were found guilty or not guilty. ML programs are supposed to learn a rule or rules which will enable them to predict whether a person is guilty or not. When we ran HR (without Lakatos' techniques) on this data it constructed a large set of implication conjectures such as $caused\_harm(A) \wedge had\_plan(A) \rightarrow guilty(A)$.

5.4 Philosophy of Mind

Horn [12] has developed a graphical approach to argumentation analysis. Using the technique of mapping issues and their connection to each other onto large posters, he is mapping the 'great debates'. His first debate to be published presents the arguments which surround the question "Can Computers Think?", and while not comprehensive, the map is the result of a major project which consists of seven large wall maps in which more than 800 arguments are represented. It is interesting to see how many of the arguments can be retrieved using Lakatos' methods (all arguments in this section are on [12] and are not our own). For instance one of the main conjectures is that 'computer programs cannot be creative' - found by *induction* on the many non-creative (and lack of creative) computer programs. Counterexamples - computer programs which are creative - include Johnson-Laird's jazz generator, Cohen's AARON and Klein's book generator. Some of these are then *monster-barred*, for instance it is argued that the book generator is not really creative. The concept of creativity is then explored and tightened up, in the same way as as barring the hollow cube lead to a tighter definition of the concept polyhedron.

5.5 Political argument

It seems clear that Lakatos' methods apply to subjects such as philosophy, law and politics, in which persuasive reasoning is all important and definitions, claims and arguments are modified according to the proponents' goals. A recent example of political argument is the controversy over the American government's treatment of the Taliban prisoners. The Geneva Convention states that "all prisoners of war should be treated humanely". This is the initial claim (or conjecture). The counterexample then arose of prisoners of war who were not being treated humanely. Since the government did not wish to violate the Geneva Convention, when challenged with the counterexample they claimed that the Taliban prisoners were *not* prisoners of war. (This is similar to claiming that the hollow-cube is *not* a polyhedron, i.e. *monster-barring*.) Therefore the government's treatment was not a valid counterexample, and the Geneva Convention was not violated. To aid the argument a new concept - *battlefield detainees* - was invented to describe the prisoners[7].

5.6 Meta-level reasoning

The domain of meta-level reasoning is clearly important if the methods are to be used to transform a concept space. Some of the methods may be used to determine which method is appropriate in a given situation. For example a mathematical agent may use *induction* to form the conjecture 'exception barring always produces good results'. Another agent may find an example of something produced by this method but not consider it a good result (the disagreement may arise if the first agent does not know of the counterexample, or it may know it but evaluate it differently). For instance if the second agent worked in philosophy it might produce the counterexample above, in which the conjecture 'actions which maximise happiness, are good' is modified using exception barring to 'All actions (except unjust actions) which maximise happiness, are good' which may be evaluated poorly since it too weak to be useful. This might result in examining

[7] The definition of *humane treatment* was also disputed - in particular whether it could ever include interrogation, as the American government felt it important to interrogate the prisoners while not wanting to be open to the charge of inhumane treatment.

the concept 'good', or using *exception-barring* to modify the conjecture to 'exception barring always produces good results, except in philosophy'.

With respect to HR the methods are heuristics which suggest theory formation steps such as which concept or conjecture should be developed. This is not meta-level reasoning but could be used to enable it to transcend usual heuristic boundaries. One of the successes of HR is its ability to reason at the meta-level without major changes being made (shown in [7]), and so the combination of HR and Lakatos's methods looks promising for meta-level reasoning.

6 CONCLUSION

Lakatos's *Proofs and Refutations* provides a rich account of the history and development of a mathematical field, and contains much of interest to AI researchers. Our argument is that it is especially relevant to machine creativity. In describing methods which explore, transform, re-represent and evaluate a domain it seems tailor made to creativity research. We have argued in this paper that Lakatos's work can be used to address key issues in machine creativity, helping to:
(i) identify a putative concept space - mathematics;
(ii) explore this concept space - by building a theory containing conjectures, concepts and examples/counterexamples;
(iii) transform the concept space - using the methods themselves to suggest theory formation;
(iv) suggest when and how to re-represent knowledge, and what to do with the new representation - by the method of monster-adjusting which is used to preserve a conjecture thought valuable, and results in exploration of subconcepts inherent in the conjecture; and
(v) develop an account of dynamic evaluation criteria - by the method of lemma incorporation, which suggests ways of modifying a proof or argument - see as evaluation criteria for a conjecture.

We have also described various approaches to implementing these methods - including our own approach using HR, and our preliminary results which we consider very promising. Finally we have argued that Lakatos' methods are not limited to mathematics but apply to many domains.

We now intend to implement further methods, in number theory and then in 2-dimensional geometry and moral reasoning. We will then attempt to show that this is important to machine creativity, by evaluating our resulting system with respect to its creativity (and if appropriate, compare it to previous versions of HR). We will refer to recent creativity measures such as those in [22].

Despite the obvious fact that almost all areas of knowledge have grown through collaboration of some sort (in which we include competition, disagreement, etc), there has been little attempt within machine creativity to model social interaction ([25] is one exception). Lakatos's work on the history and philosophy of mathematics is invaluable to researchers in this field.

Acknowledgements

We would like to thank the anonymous reviewers in the ECAI conference as well as the workshop for some very helpful comments on an earlier draft. This work was supported by EPSRC grants GR/M45030 and and GR/M98012. The second author is also affiliated with the Department of Computer Science, University of York.

REFERENCES

[1] M. A. Boden, 'Creativity: a framework for research', *Behavioural and Brain Sciences*, **17**(3), 558–570, (1994).
[2] M. A. Boden, 'Computer models of creativity', *The Psychologist*, **13**(2), 72–76, (2000).
[3] M.A. Boden, *The Creative Mind: Myths and Mechanisms*, Weidenfield and Nicholson, London, 1990.
[4] A. Bundy, 'Proof analysis : a technique for concept formation', Technical report, Dept Artificial Intelligence, University of Edinburgh, (1983).
[5] A. Bundy, 'What is the difference between real creativity and mere novelty?', *Behavioural and Brain Sciences*, **17**(3), 533 – 534, (1994). Open peer commentary on [3].
[6] S. Colton, *Automated Theory Formation in Pure Mathematics*, Ph.D. dissertation, Dept. of Artificial Intelligence, University of Edinburgh, 2001.
[7] S. Colton, 'Experiments in meta-theory formation', in *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science*, ed., G. Wiggins, (2001).
[8] S. Colton, A. Pease, and G. Ritchie, 'The effect of input knowledge on creativity', in *Technical Reports of the Navy Center for Applied Research in Artificial Intelligence*, (2001).
[9] Hayes-Roth, 'Using proofs and refutations to learn from experience', in *Machine Learning: An Artificial Intelligence Approach*, eds., R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, 221–240, Tioga Publishing Company, Palo Alto, CA, (1983).
[10] H. Hirsh, *Incremental version-space merging : a general framework for concept learning*, Ph.D. dissertation, Stanford University, 1989.
[11] D. Hofstadter, *Fluid Concepts and Creative Analogies*, HarperCollins, NY,USA, 1994.
[12] R. Horn, *Map 1 - Mapping Great Debates: Can Computers Think?*, http://www.macrovu.com/CCTMap%201.html.
[13] W. James, *The Principles of Psychology*, Henry Holt, NY, 1890.
[14] A Karmiloff-Smith, 'Is creativity domain-specific or domain-general? clues from normal and abnormal development', *AISB Quarterly*, (85), 26 – 29, (1993).
[15] T. Kuhn, *The Structure of Scientific Revolutions*, The University of Chicago Press, Chicago, USA, 1970.
[16] W. Kymlicka, *Contemporary Political Philosophy*, A Bradford Book, Clarendon Press, Oxford, 1995.
[17] I. Lakatos, *Proofs and Refutations*, CUP, 1976.
[18] McCloskey, 'A note on utilitarian punishment', *Mind*, **72**, (Dec 1963).
[19] G. McGraw and D. Hofstadter, 'Perception and creation of diverse alphabetic styles', *AISBQ*, (85), 42 – 49, (1993).
[20] J.S. Mill, *Utilitarianism*, Parker, Son and Bourn, London, 1867.
[21] A. Pease, 'Ph.D. research proposal: A computational model of mathematical creativity via interaction', Technical report, Dept Artificial Intelligence, University of Edinburgh, (2001).
[22] A. Pease, D. Winterstein, and S. Colton, 'Evaluating machine creativity', in *Technical Reports of the Navy Center for Applied Research in Artificial Intelligence*, (2001).
[23] G. Polya, *Mathematical Discovery*, John Wiley and Sons, New York, 1962.
[24] E. Rissland, 'Exgen : a constraint satisfying example generator', Technical report, Department of Computer and Information Science, University of Massachusetts, (1988).
[25] R. Saunders and J. S. Gero, 'The digital clockwork muse: A computational model of aesthetic evolution', in *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science*, ed., G. Wiggins, (2001).
[26] J. Smart, 'An outline of a system of utilarian ethics', in *Utilitarianism: For and Against*, eds., J. Smart and B. Williams, 1 – 74, CUP, (1973).
[27] B. Williams, 'A critique of utilitariansim', in *Utilitarianism: For and Against*, eds., J. Smart and B. Williams, 75 – 150, CUP, (1973).