Invited Talk: Thinking With Diagrams Conference, Aberystwyth, 1998
To appear in *Diagrammatic Representation and Reasoning*
Eds. M. Anderson, B. Meyer & P. Olivier, Springer-Verlag

Diagrams in the Mind?

Aaron Sloman
School of Computer Science, University of Birmingham
Birmingham, B15 2TT, UK
http://www.cs.bham.ac.uk/˜axs

Abstract

Clearly we can solve problems by thinking about them. Sometimes we have the impression that in doing so we use words, at other times diagrams or images. Often we use both. What is going on when we use mental diagrams or images? This question is addressed in relation to the more general multi-pronged question: what are representations, what are they for, how many different types are they, in how many different ways can they be used, and what difference does it make whether they are in the mind or on paper? The question is related to deep problems about how vision and spatial manipulation work. It is suggested that we are far from understanding what is going on. In particular we need to explain how people understand spatial structure and motion, and how we can think about objects in terms of a basic topological structure with more or less additional metrical information. I shall try to explain why this is a problem with hidden depths, since our grasp of spatial structure is inherently a grasp of a complex range of possibilities and their implications. Two classes of examples discussed at length illustrate requirements for human visualisation capabilities. One is the problem of removing undergarments without removing outer garments. The other is thinking about infinite discrete mathematical structures, such as infinite ordinals. More questions are asked than answered.

1 We can think with diagrams

Consider the trick performed by Mr Bean (actually the actor Rowan Atkinson): removing his (stretchable) underpants without removing his trousers.[1] Is that really possible? Think about it

[1]The first draft of this paper located Mr Bean in a launderette. Toby Smith corrected me, pointing out that the shy Mr Bean was on the beach, and wished to remove his underpants then put on his swimming trunks, both without removing his trousers. On 29th July 1995 I posted Mr Bean's problem as a followup to a discussion of achievements of AI in several internet news groups (comp.ai, comp.ai.philosophy, sci.logic, sci.cognitive) and received a number of interesting and entertaining comments. Chris Malcolm pointed out the similarity with the bra and sweater problem, i.e. removing a bra without removing the sweater worn above it. Readers are invited to reinvent the jokes that were then posted, about which problem was easier for whom under which conditions. In particular, someone pointed out the distinction between difficulty due to unfamiliarity *vs* difficulty due to being distracted.

if you haven't previously done so.[2]

Is it possible to remove the underpants without removing the trousers, leaving the waistband of the trousers constantly around the person's waist, allowing only continuous changes of shape of the body and the underpants and trousers, e.g. stretching, bending, twisting, but with no separation of anything into disconnected parts, no creation of new holes, etc.? Does it matter whether the waistband of the trousers is tight or not?

Many people can answer this question by thinking about it and visualising the processes required, even if they have not seen Rowan Atkinson's performance. A harder question is: in how many significantly different ways can the underpants be removed?

2 Some comments on the underpants problem

It is easier to consider the underpants being distorted, ignoring who does it and how, than trying to work out all the contortions of posture Mr Bean would have to go through to produce the appropriate sequence of changes. If we abstract away from the problem of how the wearer makes the transformations happen we can suppose Mr Bean remains rigid and still and someone else pulls and stretches his underpants, perhaps using long thin tongs where necessary. (Is it obvious that this change makes no difference to the main problem? Why?)

Even with this abstraction there are several different ways of thinking about the underpants problem. Some use only topological relationships preserved under all continuous transformations, including those which change size, shape and distances. Some also use metrical relationships involving shape and size. We can also use topological relationships with structural features of under-specified metrical relationships.

Thinking purely topologically is quite hard to do, since it involves finding the most general way to characterise the relationship between Mr. Bean and his garments in the initial and final states. From that point of view the start and end states are equivalent and there is no problem for Mr Bean to solve. So it cannot be the right way to think about the problem of how to do it. Most people do not think like that. They conceptualise the problem in a largely qualitative but partly metrical fashion, including various ways the underpants might stretch and fold. We shall see that it is useful to combine different abstractions.

2.1 How many distinct solutions are there?

Most people at first see only two symmetrically related solutions to the problem. One involves stretching the left side of the underpants down through the left trouser leg, over the foot and back up the left leg, leaving only the right leg through its hole. The underpants can then be slid down the right leg and out. A similar solution starts on the right side, with the underpants emerging through the left trouser leg.

If the waist band of the trousers is loose there are several more pairs of symmetrically related solutions, e.g. sliding one side of the underpants up over the head and down the other side and out through the leg, or sliding the central (leg-divider) part of the underpants down inside a leg then over the foot and up the same leg on the outside, then out past the waist band, over the head and down the other leg. It is easier to visualise than to describe! Another pair of solutions

[2] I have previously given audiences the task of finding out how many possible numbers of intersection (or tangent) points there can be between a triangle and a circle in the same plane. It is easier than Mr Bean's problem, but many people miss out some cases unless prompted.

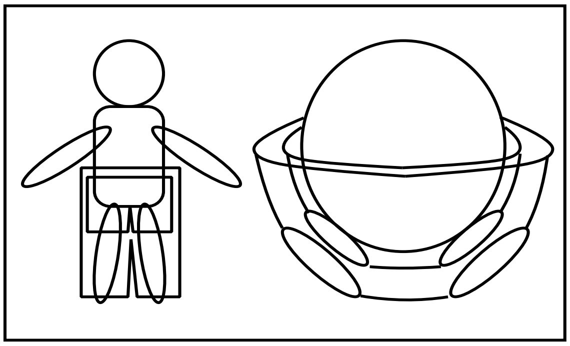

Figure 1: *Mr Bean still wearing his trousers and underpants, before and after being continuously transformed into a sphere.*

starts the same way, and ends with the underpants going off past the head. That is four pairs of solutions so far. But there is at least one still missing! (Or more, depending how solutions are counted.)

At first I saw only two solutions, and did not think of pulling the underpants over the head until a usenet poster mentioned the possibility. Then I looked for more solutions and noticed that the central part of the underpants could be moved first, leading to underpants around the waist. Eventually, after further abstraction, followed by some arithmetic, explained below, I found *nine* different solutions. Most people don't find them all.

3 A spherical Bean

The solutions outlined above all used metrical notions including stretching and translation. We can de-emphasise metrical features (size, shape, distance, orientation, sizes of angles) and focus more on topology if we envisage the body shrinking to a sphere, or egg, as in Fig. 1, with the trousers and underpants following faithfully, so that each becomes a hemisphere with two holes, while their waistbands remain around the equator.

It is clear that if the revised problem starting from a spherical shape can be solved, the original problem can be also. It is *not* clear what kinds of cognitive mechanisms enable us to grasp that fact.

Considering the shrunken Bean makes it "obvious" (how?) that the underpants can slide out through one of the holes in the trousers. Since there are two holes there are essentially two symmetrically related solutions.

Loosening the waistband permits another type of solution in which the underpants slide out past the band, with the sphere passing through one of the leg holes. Since there are two leg holes we have another symmetrically related pair of solutions.

Another solution has the underpants sliding out past the waist band, without the sphere passing through the leg holes.

So with the trousers attached and impassable at the waist, there are two distinct solutions. Loosening the waistband enables several more distinct solutions. Have we found them all?

3.1 Holey spheres

We can think of a two-holed hemisphere as a sphere with three holes! Then we can envisage underpants and trousers each as three-holed spherical sheets, concentric with each other and with the spherical Bean. The two sheets have their holes aligned, but we can ignore that.

What kind of cognitive process allows you to grasp the three-holed sphere view? I saw it like that only after I attended to the task of looking for more general characterisations of the problem and then saw that talking about the loose waistband was a distraction: it is just another hole in the trousers. Similarly there was all along just another hole in the underpants, at the waist.

What are the cognitive mechanisms that enable us to perform that sort of re-conceptualisation? How does the mechanism relate visual and non-visual information, e.g. about the nature of holes and waistbands? Why is the mechanism sometimes not invoked? What triggers its invocation?

There are two distinct but related re-conceptualisations. One involves noticing the similarity in structure and function between the big hole at the top and the two small holes at the bottom. Ignoring differences in size and location, they are similar in *function*: something inside (the underpants or trousers) can come out only by going through one of the three holes. Alternatively one can visualise a simple continuous deformation, i.e. stretching the garments up over the sphere, turning them into spheres with three similar holes.

I.e. we can discern the more abstract characterisation *either* by noting common aspects of the functional roles (causal powers) of the holes despite their difference in size (seeing affordances), *or* by visualising a deformation which makes them indistinguishable anyway (visualising structural changes and relationships). Different cognitive mechanisms and skills would be needed for these two tasks. *How are these skills implemented? How do they develop? Which animals have them?* (Cf. Kohler 1927.)

Having noticed that Mr Bean with his lower garments is equivalent to a solid sphere surrounded by and concentric with two spherical rubber sheets each with three holes, we can also notice (*how?*) that removing the underpants involves two steps:
(1) getting the sphere out of the underpants through one of the three holes in the inner sheet.
(2) getting the underpants (the inner sheet) out of the trousers through one of the three holes in the outer sheet.

Suddenly it becomes clear that there are three ways of doing step (1) each consistent with three ways of doing step (2), so there must be $3 \times 3 = 9$ different solutions, covering all possible combinations at this level of abstraction, which ignores protrusions (e.g. legs) through holes. It is also possible to do step (2) before step (1), doubling the number of solutions!

It is worth noting that the type of abstraction identified here which enables us to reason about the combinations of steps does not require Mr Bean and the two garments to have any specific shape as long as the garments are approximately convex, or at least have a distinction between inside and outside and three communication ports between them. We can discuss the spheres and their changing relationships without assuming all the metrical properties of spheres, e.g. smoothness, constant curvature, fixed radius, etc. This is what I meant by "structural features of under-specified metrical relationships".

3.2 Yet more abstraction

Further abstractions are possible. The initial configuration is topologically equivalent (deformable by continuous changes) to one in which the three items are simply separated

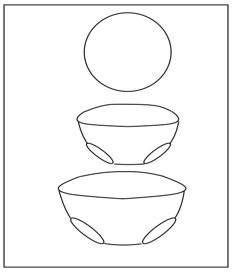

Figure 2: *'Exploded' abstract representation of Mr Bean and his garments. The hemispheres can be continuously 'flattened' into plates. Is it 'obvious' that there is no topological difference between the original and final state?*

vertically, by moving the sphere up and the trousers down (as in Fig. 2. This treats the relation of being inside and the relation of being outside a spherical surface with holes as equivalent. Moreover The two stretchable enclosing spheres can be continuously deformed into two flat sheets each with two holes. In that context nothing is inside or outside anything else, and there is therefore no difference between the initial and the final state. Either way, there is no problem to solve!

Only mathematicians react to the original problem that way, concluding that it is trivial. Unfortunately that doesn't help Mr Bean get his underpants off. Even when a mathematically satisfying solution to a problem has been found at a high level of abstraction, there is still work to be done if detailed actions have to be specified.

When moving between different abstractions we need to know where to stop. E.g. in analysing options for the removal process it is useful to go from the *fully metrical* initial specification, where the detailed shapes and sizes are relevant, to the *minimally metrical* nearly topological situation where only inside–outside relations are relevant (but still metrical because being "inside" an object with holes is a metrical property). Having enumerated possible strategies at the minimally metrical level (where each strategy involves use of one hole in the underpants and one in the trousers) we can then move to more detailed planning and evaluation in the fully metrical representation, where changes of shape and length are required, i.e. stretching of underpants over the head or down and under the foot. At that level there are far more options and the search space is much larger.

4 Coexisting search spaces

We found that there are nine different solutions when the problem is construed as involving three concentric spheres (or, to be more precise, three spheres totally ordered by an "encloses" relation). This discovery was not made by visualisation or simulation of the removal process, but by using the general information that for something to move from being inside a holed sphere to being outside it must go through one of the holes. (How does a child grasp *that* fact? Does a chimpanzee?)

Why was the full range of solutions not obvious with the original configuration? Not everyone spots the solution where both leg holes of the underpants are moved round to the top of Mr Bean's head, so that the underpants are upside down, and then pulled off upwards (i.e. Mr Bean exits the underpants through their waist hole while the underpants exit the trousers through the outer waist hole). There are different ways of doing this which are equivalent at a high level of abstraction, though they involve different contortions of Mr Bean and different locations where the underpants risk being torn.

At a fully metrical level the search space is far more complex: there are more detailed options, with more explosive combinatorics. At that level it is hard to see patterns among the routes, because the simpler structure got by grouping (almost) topologically equivalent options is not visible.

This is an illustration of the general fact that finding an abstract spatial representation and combining that with some abstract non-spatial (arithmetic or logical) reasoning can give a deeper insight into a problem than simply using very concrete spatial visualisation capabilities. Information about solutions at the abstract level can be transformed to lower-level solutions (e.g. with metrical information) by adding details, though generally there will not be a unique extension.

Having different views of a diagram or 3-D scenario involving different types of abstraction often helps in the process of solving a problem, e.g. planning a detailed sequence of actions. This is used by multi-level planners, which form meta-plans in one or more abstraction spaces (e.g. ABSTRIPS, NOAH) to control the search more effectively than a "flat" single-level planner can (e.g. STRIPS).

A related theme in the history of mathematics is the constant development of new forms of abstraction and techniques for relating and combining different abstractions. A similar theme can be found in work on in child development, e.g. by Karmiloff-Smith.

In principle, given enough time to explore visually all the possible metrical transformations we could eventually discover instantiations of all nine possibilities described above, though we might not notice the partitioning into nine cases.

However, even given enough time, most people would not get around to considering all of the options because the more complex search space involves a more complex book-keeping task if the search is to be exhaustive. Human architectures do not cope well with deep stacks or long queues, though these are easy to implement on computers.

Our limitations may arise in part because different arrays of possibilities compete in parallel for attention. When considering any spatial structure there are indefinitely many changes of size, shape, orientation, colour, etc. that we can envisage if we think of them (Sloman 1996a). AI models of visual or spatial reasoning do not yet match this, though perhaps they will in the distant future.

Part of the price of such human flexibility is unmanageable combinatorics when searching for a sequence of changes to solve a problem. This can be alleviated by using more abstract patterns to control the search, though not everyone can do this equally well. Could this also explain the different achievements of the chimpanzees in Kohler's famous experiments?

Explaining how capabilities at different levels of abstraction are used and combined to control the search for a solution to a complex problem requires not only a specification for the representations and mechanisms used, but also the architecture which combines them and allows different processes to interact fruitfully (Sloman and Logan, 2000).

4.1 What makes us fail?

Why do people sometimes fail to visualise an action or change, or fail to draw an inference?

It may be due to (i) use of poor representations, (ii) use of inadequate mechanisms or algorithms for manipulating the representations, (iii) inadequate architecture for combining and integrating different sorts of representations and mechanisms (e.g. ability to construct only simple structures, limited possibilities for modifying structures, limited possibilities for analysing structures, limited short-term memory for storing sequences of modifications), (iv) wrong or incomplete stored information (e.g. about changes possible in a physical system, about consequences of changes), (v) inadequate mechanisms for monitoring effects of changes in order to infer consequences, (vi) lack of meta-level know-how and architectural support required for systematically exploring all the available information and all the available transformations, (vii) not using available know-how e.g. because of an attention problem or a motivational problem or some kind of "fixation" on a different inadequate strategy.

The above points illustrate some of the requirements for a system able to explain or model human abilities. Some failures may involve transient dysfunctions, such as distracted attention, or forgetfulness. There may be others produced by brain damage, genetic brain malformations, drugs, chemical disorders, etc.[3] Some tasks may come too early for a developing architecture, in childhood.

5 External and internal diagrams

Our discussion shows that a diagram on paper is not necessarily a good model for what is grasped when someone visualises a spatial structure.

One person looking at the diagram may see only the more detailed, metrically specific configuration whereas another can see ("grasp"? "comprehend"?) in the same diagram a more abstract structure in which metrical relationships play a reduced role. The two views support different ways of seeing possible changes. So even if both perceivers had an internally inspectable 2-D diagram they might still view it quite differently. Simply having internal spatial structures cannot explain what it is to grasp or visualise a spatial aspect of a scene or problem. (Otherwise simply having a brain would suffice.)

Asking whether people can build internal diagrams is less important than asking how diagrams can be viewed, analysed, interpreted, and used, no matter whether they are internal or external. Introspective reports should be treated as highly ambiguous and incomplete descriptions, and certainly not as explanations.

5.1 Representations and transformations

All of the different ways of thinking about Mr Bean's problem require not only some way of representing the original configuration, but also a grasp of the possible *transformations* of that configuration, a capability discussed more fully in a discussion of 'actual possibilities' in (Sloman, 1996a).

We have seen that different transformations are possible at different levels of abstraction. At one level there are many detailed changes of shape as Mr Bean pulls part of the underpants down his trouser leg, over the foot and then back up again. At the highest level of abstraction that is a non-operation: the sphere is still in the underpants, as if a protrusion from the sphere (the leg) has been squashed in, leaving the underpants free to rotate around the sphere.

So visual experiences of looking at the diagram at various levels of abstraction differ in (among other things) the *possibilities for change* that are seen. Mental visualisation without

[3] I conjectured in (Sloman, 1989) that some autistics lack the perceptual ability to move up levels of abstraction in perception, also described in more recent papers (Sloman and Logan, 2000; Sloman, 2000b)

an external diagram must also involve assembling possibilities for change in thinking about a solution to the problem. Practice somehow develops fluency in doing this: *How?* I learnt a great deal by playing with Meccano sets, as a child. Different visualisation skills are developed by mathematical or other sorts of training. What changes during such learning?

Experienced software engineers gain facility in grasping very abstract configurations of data-structures along with procedures which transform them. Likewise, being a composer, painter, mechanical engineer, dressmaker, etc., involves acquiring specialised abilities to grasp structures along with classes of possible transformations of those structures and their consequences.

Different structures in the same general class can support very different numbers and types of transformations. A drawing with a few lines supports far fewer "immediately available" transformations than more complex line drawings with far more lines, junctions, regions etc. Thus as you visualise a structure changing, the requirements for grasping which further changes are possible may also be constantly changing. Often a change is made intentionally in order to allow new possibilities, e.g. visualising a mechanical link being shortened in order to allow it to rotate further before being stopped. *How do we grasp these second-order possibilities for change?*

6 Thinking with qualia

All this is related to disputes about the nature of consciousness (Chalmers, 1996). E.g. are qualia simply unanalysable 'givens' or are they best understood as crucial parts of the functioning of an information processing system (as I have argued in a long, incomplete, still expanding paper, available at

www.cs.bham.ac.uk/research/cogaff/Sloman.consciousness.evolution.ps

and in (Sloman, 2000a))

Our discussion shows that visual qualia (e.g. an experienced red patch) have rich "internal" differences depending on what sorts of possibilities for change the experiencer is capable of handling. Changes could include changes of shape, size, orientation, location, splitting into two or more patches, and many ways of acquiring new coloured sub-regions (e.g a blue patch in the middle or a green line traversing the red patch, and so on.)

Wittgenstein wrote: "The substratum of this experience is the mastery of a technique" (Wittgenstein, 1953, p208). A full account of visualisation (and thinking with diagrams or other spatial structures) would require us to analyse the huge variety of techniques implicit in even the simplest human experiences, thereby uncovering requirements for mechanisms able to support apparently simple qualia.

Other animals may have much simpler qualia, especially *precocial* species born or hatched with genetically formed visual mechanisms ready for use, e.g. chickens, deer, horses. *Altricial* species, e.g. birds of prey, hunting or tree-climbing animals and humans, start off more helpless and grow their brains while interacting with the environment. Perhaps this 'bootstrapping' produces a much richer grasp of structure and motion than can easily be encoded in genes. (Contrast this with the popular opinion that humans are born so immature because their skulls would otherwise be too big to pass through a human pelvis. Elephants manage, so that can't be all there is to it.)

7 Visualising infinite structures

Some visualisation goes beyond what can be experienced in perception. How do we visualise infinite structures? The answer will depend on the type of infinite structure. When we visualise continuous objects or continuous changes this involves the possibility of "zooming in" to smaller and smaller portions of the object or motion, without limit. That is part of what is implied by being continuous. It also underlies some of Zeno's paradoxes.

Mr. Bean's problem involves continuous change (stretching, bending, moving), but solving that problem does not deploy most of what we know about continuous motion. The difference between continuous change and a finite succession of discrete states would not make any difference to our previous discussion. In fact a useful way to tame a problem involving continuous change is to identify a small number of key states, and ignore intermediate states. That is how we found 9 or 18 distinct solutions.

We can also think about infinite discrete structures, like the set of integers or the set of proofs in some formalism. Clearly we cannot create something infinite inside our heads. So visualisation in this case (and probably in all the other cases too!) does not involve actual creation and inspection of the structure visualised. Something far more subtle happens: when you visualise a spatial structure or process there need not be any actual spatial structure or process that is inspected, nor anything isomorphic with the structure or process.

There might be only a *representation* of inspecting the structure or process. If done well, that could fool us into thinking we are doing something that we aren't. But being fooled doesn't matter as long as the process which produces the illusion is exactly what is needed to implement a powerful reasoner or problem solver: i.e. it is a good biological solution, like being fooled into thinking tables are smooth, solid, continuous and rigid, because they *look* and *feel* as if they are.

7.1 Infinite "images" involving numbers

Let us consider some examples of infinite structures, such as the sequence **N** of natural numbers, 0, 1, 2, ... etc. This is easily visualised, going off into the distance away from us, or from left to right, for instance. **N** satisfies Peano's axioms for arithmetic. (i) There is an initial element. (ii) Every element has a unique successor. (iii) The initial element has no predecessor. (iv) Every non-initial element has a unique predecessor. (v) The axiom of induction: properties which are possessed by the initial element, and possessed by the successor of any possessor, are possessed by all the elements.

Any sequence satisfying those axioms, e.g. an infinite row of dots, or an infinite sequence of repeated actions is a Peano structure. It is clear that there are many visualisable subsets of **N** which are Peano structures, e.g. the even numbers, 2, 4, 6, ..., or the numbers starting from 999 and continuing indefinitely: 999, 1000, 1001, ... It is also clear that Peano structures all have certain properties, some of which are easier to grasp than others.

Grasping the relationship between the axiomatic characterisation and the visualised structure is non-trivial. For hundreds (thousands?) of years before Peano came up with his axioms, people thought about and used numbers and were able to visualise the infinite sequence of numbers. Kant discussed some of the issues in 1781.

What cognitive mechanisms enabled Peano to find the axioms? Consider the different roles of the axioms in characterising the required set. Axioms (i) and (ii) guarantee that the set is not empty and that you can go on along the sequence forever, with no choice points (because of the word "unique"). Axiom (iii) prevents you going backwards beyond the initial element. Axiom

(iv) implies that you can go back from any non-initial element, and again the word "unique" rules out choice points, thereby preventing the sequence doubling back and rejoining itself, as this one does: 0,1,2,3,4,5,6,3,4,5,6,3,4,5,6...I.e. axiom (iv) prevents 3 having both 2 and 6 as predecessors. Axiom (v) is more subtle, and prevents sequences which go on forever, and then have more items beyond that, like **S1** defined below.

We can easily infer some properties of a visualised Peano structure. E.g. given any two distinct elements in the structure, there must be a finite chain of successor elements starting with one of them and ending with the other. So the elements comprise a total ordering. Compare proving this from the axioms using logic. We can also see that every initial sequence of a Peano structure is finite, and every alternate initial sequence can be arranged as a rectangular 2 by N block of items, where N is some number, and the intervening ones cannot.

7.2 More complex infinite structures

We can also visualise structures violating Peano's axioms. For example, imagine the even and odd numbers separated out, into two sequences, 0, 2, 4, ... and 1, 3, 5, ... We can visualise these concatenated in a structure **S1** with all the even numbers going from left to right, followed by all the odd numbers going from left to right.

Then **S1** has a successor relation just as **N** did, but it is "obvious" that Peano's axioms are no longer satisfied in **S1**. First, not every non-initial number has a predecessor in the new configuration. (There is one exception.) Secondly the axiom of induction no longer holds: properties which are possessed by the initial number, and possessed by the successor of any possessor are no longer possessed by all the integers in this new organisation. An example is *being even*.

We can visualise a different infinite series **S2** by reversing the odd numbers and adding them all *before* the even numbers. That produces a structure like the set of positive and negative integers which is infinite in both directions. There is no longer any item without a predecessor. **S2** has symmetry lacking in Peano structures.

Moreover, if we start from the fact that there are infinitely many prime numbers (which is provable algebraically, though not so easily proved visually), we can form infinitely many Peano structures and concatenate them. Starting from any prime number we can form a Peano structure consisting of all its powers, e.g. $2^1, 2^2, 2^3, ...3^1, 3^2, 3^3, ...5^1, 5^2, 5^3, ...$ It is then not hard to visualise *all* of these sequences concatenated to form **S3**, a totally ordered set of numbers, which has infinitely many elements violating axiom (iv) because they have no predecessor. This can either be proved formally from a logical specification of the construction of **S3**, or intuitively by visualising the process of construction and seeing that each time a new set of powers is added its first element has no predecessor.

7.3 Well-ordered structures

The original sequence **N** can be seen to be "well-ordered", i.e. every subset of **N** contains a "least" element, one which has no predecessor in the subset and which precedes all the others in the subset. This is connected with the fact that **N** is inherently asymmetric. It is built by starting with an initial element and going on indefinitely adding elements, one at a time, on one side only. Proving logically that every Peano structure is well-ordered is harder than *seeing* that it is.

Experienced mathematicians can also see that the structure **S3** got by concatenating infinitely many Peano structures, is well-ordered.

This would not be true if we reversed some of the sub-sequences, e.g. if all the powers of 13 were included in reverse order. That would violate well-ordering since there would be a subset with no first element.

7.4 Justifying Peano's axioms

Having noted that it is easy to visualise structures, like **S1**, **S2**, **S3**, which violate the axioms in different ways, we can see that one way to "justify" Peano's axioms is using them to rule out those structures. I have no idea if this is how Peano arrived at his axioms.

Whether those axioms suffice to determine uniquely the "intended" intuitive model is a controversial topic discussed more fully in my review ((Sloman, 1992)) of Penrose.

A Peano structure whether specified axiomatically or visually is asymmetric. Moving along it in one direction always leads to the least element, whereas the other direction goes on forever, which we often represent by "..." Being "well-ordered" is another type of asymmetry: every subset has a first element, though not necessarily a last one.

7.5 How do we grasp an infinite ordered sequence?

It may be that part of what makes the visualised infinite natural number sequence what it is rather than a non-Peano structure is an information-processing implementation of the asymmetry along with something closely related to the axiom of induction. I do not know how to make this precise.

Two aspects of such an implementation could be (1) a mechanism for expanding an incomplete sequence "on the right" as often as required, and (2) a reasoning mechanism that implicitly assumes that properties propagated to successors are propagated to *everything* further along. This sort of mechanism is not inherently connected with numbers.

Anyone who can visualise an infinite row of vertical dominoes going off to the right, and then visualise the wave of activation that occurs when the first domino falls over causing the second one to fall over, etc. and who finds it "obvious" that they will all (eventually) end up knocked over, is using the equivalent of the axiom of induction. *How is the ability to do this implemented in human brains?* It is probably part of a large suite of operations for manipulating finite and infinite discrete structures, which will be different in detail from those for continuous structures, but may have some overlap, e.g. the ability to concatenate structures, or to "move" something along a structure.

What makes something a visualisation of a Peano structure, rather than a different sort of structure such as **S1**, **S2**, or **S3**, depends on the applicability everywhere of this local property-transmitter. The infinite detail need never be constructed, as long as it is available when needed (as in lazily evaluated data-structures). This is partly analogous to whatever makes it possible indefinitely to zoom in to continuous structures. For Peano structures we use something like an ability indefinitely to "zoom to the right".

When and how do young children develop this ability? How did it evolve? Was it a side-effect of other abilities?

7.6 Visualising proofs and refutations

It is easy to visualise counter-examples to the claim that all ordered structures are Peano structures, or that all ordered structures are all well-ordered. It is not so easy to use visualisation to prove generalisations, such as that *any* concatenation of a well-ordered set of well-ordered structures will also be well-ordered. For some people, and perhaps for all, that is much easier to prove by reasoning logically from definitions than to demonstrate by somehow visualising all possible concatenations of well-ordered sets. How would one do that?

In general it is easier to visualise a case that refutes a generalisation than to visualise all possible instances of a generalisation in a reliable way. Sometimes that can be done by visualising a sort of pattern or template which covers all the possibilities. Mateja Jamnik's work (Jamnik et al., 1999) on verifying diagrammatic proofs, includes the use of diagrams to reason over an infinite set of finite structures, e.g. in proving that for every N the sum of the first N odd numbers is N^2. This depends on a common pattern shared by all the structures, so that they can be visualised in a uniform way.

A much harder visualisation of an infinite structure (or process) is required to prove the Cantor-Bernstein theorem, which says that if there are two sets A and B each of which is in one-to-one correspondence with a subset of the other, then there is a one-to-one mapping between A and B. The proof involves constructing the new mapping from the two given ones, and it is helpful when thinking about this to visualise something like a pair of mirrors facing each other with rays bouncing back and forth indefinitely.

8 How do we do it?

What is going on when we visualise these infinite structures? We obviously don't construct infinite physical structures since our brains are finite. However, it may be accurate to say that infinite structures are constructed in some sort of virtual machine, like the familiar virtual machines that support sparse arrays or infinite lazily evaluated lists, constructable in some programming languages. It is not hard to create in a computer a sparse array with more locations than there are electrons in the universe, as long as we leave most locations containing the default value. Perhaps brains (or the virtual machines we call minds) use similar tricks for representing extremely large, or even infinite, structures.

It might be tempting to think that what we do when we visualise an infinite structure is construct a very large set and use that as an approximation to the infinite set, since after all a very very large visualised collection of dots, like a starry sky, might as well be infinite if we cannot take in the whole lot and see how many there.

But that won't do. If you visualise the structure **S1**, with ALL the even numbers followed by ALL the odd numbers, then no very large finite subset of the even numbers will do as an approximation to ALL of them. For example, the structure **S1**violates Peano's axioms, as explained above, whereas if there are only finitely many even numbers preceding the odd numbers then the axiom that every number has a unique predecessor will no longer be violated, for the first odd number will now have a predecessor, the last even number. Moreover the axiom of induction will again hold. I.e. if we replace the infinite sequence of even numbers with a finite subset this will transform **S1**into a Peano structure. So a large finite row of even numbers cannot model the required infinite row in this context.

Something deep goes on when we visualise the two infinite sets as being concatenated. Perhaps the important point is that what we experience as pure visualisation is actually a combination of visualisation and unconscious but explicit specification of rules for indefinite expansion and rules for inference? (I think that sort of idea goes back to Immanuel Kant (1781).) E.g. we may have something like the previously mentioned mechanism for "continuing to the right" waiting in the wings to prevent any interpretation of the set of evens as a finite set, however large. This is like the 'lazy evaluation' of an infinite list structure in a computer: the list has a 'generator' procedure and looking beyond the already expanded portion of the list causes the generator procedure to be run, to produce previously unavailable list elements.

Using lazy evaluation is a fairly abstract and sophisticated kind of visualisation, on a par with the domino/induction mechanism that was previously waiting in the wings to propagate properties along all the natural number sequence.

How many other sorts of visualisations involve such a mixture of implicit rules or axioms or mechanisms along with something like a spatial structure? One of the requirements for a mechanism of the sort discussed here is that whether the visualised spatial structure is finite or infinite, discrete or continuous, the visualisation is possible only insofar as it implicitly involves the availability of a large number of *possible* changes in the structure, as previously discussed. What exactly is visualised depends on exactly which transformations are available.

9 Visualising is not like seeing

From the discussion so far, it is clear that whatever visualisation of a structure is, it *cannot* be something very similar to seeing even if it *feels* similar. That is because the kind of grasping of a spatial structure involved in visualising is *part* of what happens in seeing the structure. Hence if visualising involved seeing then visualisation would be part of visualising and we'd have an infinite regress.

Also we cannot *see* an infinite (discrete) structure but we can *visualise* one. And it is arguable that when we visualise the kind of abstract topological structure that we previously discussed, that cannot be like seeing because seeing always involves *specific* metrical or topological structures and relationships which are missing in the *abstract* visualisations.

We need a new way of thinking about the problem, other than proposing that the brain creates 2-D or 3-D arrays and then "looks at" or "inspects" them, for if the looking at or inspection involves understanding the spatial structure we are going round in circles chasing a non-existent homunculus. There must be a way of understanding spatial structure (or more generally) a way of understanding, which is not to be explained in terms of understanding another structure!

It must, however, be something like a type of information-rich control state, i.e. a state which affects what the system can or will do next. Elsewhere I have argued that we need to view minds as control systems and representations as control substates with syntax, pragmatics and in some cases semantics, e.g. (Sloman, 1993a; Sloman, 1993b; Sloman, 1996b).

What sort of control state? How does grasping some structure affect what you can do? Note that "what you can do" does not refer only to external behaviour. It includes the sorts of *internal* processing which become available when we grasp some structure. We need a theory of an architecture that can accommodate all these processes.

10 Other problems involving visualisation

Mr Bean's task is just one of many problems which people seem to be able to solve by *visualising* transformations of a structure.

Some are much easier: e.g. if a penny with the "head" on top is turned over three times will the head or the tail be on top? That one is easy to do *either* by visualising the process (simulating it mentally) *or* by reasoning about it. If we modify the problem to one in which the penny is turned over three thousand and five times, it is much easier (and far more reliable) to reason about than to visualise ((Sloman, 1971)).

Here, the more sophisticated process, using meta-level knowledge about the nature of the less sophisticated process, is easier and faster to do than the less sophisticated process which blindly goes through the steps to get from the start state to the end state. Being able to discover new ways of solving old problems and being able to select between alternative approaches requires "meta-level" knowledge, i.e. the ability to reflect on and reason about knowledge and problem solving. One of the earliest interesting examples of this was Sussman's Hacker (1975), which debugged itself by watching itself at work, though it dealt only with a tiny fragment of the problem, like most AI models so far.

Being able to understand the possibility of looking for and using "easy" short cuts requires a more sophisticated processing architecture than a typical problem solver or planner. It requires an architecture which supports mechanisms for observing, analysing, evaluating, and noticing patterns in internal processes ((Sloman, 2000a)).

However, having an architecture supporting such meta-level abilities does not guarantee general meta-level competence. It seems that humans have to learn to be reflective in different domains. E.g. someone who is good at noticing opportunities for improving his software designs may fail to notice opportunities for improving communication and relationships with other people.

Much mathematical ability seems to depend on grasping patterns and structures in one's own thinking and reasoning processes, like noticing that the outcome of a counting process does not depend on the order in which items are counted, or noticing that a repetitive process can continue indefinitely. I suspect that our ability to visualise infinite structures is related to the ability to grasp and reflect on properties of repetitive processes, and our ability to manipulate them by performing operations like concatenation or reasoning about subsets depends on noticing analogies between infinite structures and finite structures.

Children don't seem to start off with these abilities, but, unless damaged by teachers (or parents?), they somehow manage to bootstrap the more sophisticated architecture and to apply it in different domains. (For some speculations about this in connection with learning about numbers, see Sloman (1978, Chp 8). and compare with Karmiloff-Smith (1996).)

11 Some questions

The examples discussed above raise a host of interesting questions, relevant both to understanding how human minds work and how to give intelligent machines the ability to reason spatially.

1. What sort of knowledge enables people to work out the answer? (This subsumes the deep question: what sort of knowledge enables them to understand the problem?)

2. How is that knowledge represented in their brains – both physically in chemical and neural structures and within the information-processing *virtual* machines implemented in brains? How many different forms of representation do we have available for such knowledge? (Sloman, 1971; Sloman, 1985; Glasgow et al., 1995; Peterson, 1996; Sloman, 1996b)

3. Can the information used be expressed in predicate calculus? In first-order predicate calculus? In some other mathematical or logical notation?

4. What would the knowledge actually look like if expressed in some form of predicate calculus, or other logical system? (I.e. which predicates, functions, etc. would be used? Which axioms? How would the initial state and desired end state be described? Would modal operators be needed, e.g. to express which transformations are *possible*? Would temporal operators be needed to express the notion of a *process* and the constraints on the process? How would the requirement that the waistband not be moved be expressed?)

5. What sorts of logic engines would be able to find the solution? What sort of search space is involved? How can such a search be controlled?

6. What alternatives are there to logical representations and manipulations? What are their advantages and disadvantages?

7. What sorts of reasoning mechanisms do people actually use for this sort of problem? Can they use logic? Do they ever use logic? What alternatives are available, for humans or intelligent machines?

8. Can some or all of the human competence be replicated on computer-based machines using a very different physical implementation?

9. Which of these abilities are shared by which other animals, e.g. a magpie building a nest in a treetop out of twigs of many shapes and sizes, a squirrel working out a route to the bag of nuts hung up for birds, a female orang-utang in a tree clutching her infant with one hand and using the other to weave a nest for the night, out of branches and leaves?

11.1 Has AI made much progress on these questions?

Like many others, I have been thinking (and writing) about such questions, and about how human and animal vision works, for many years (see the References) and have seen various ideas about this re-invented many times. But I remain deeply puzzled since nothing I have come across in AI, or in psychology or brain science, seems to come close to explaining human (and animal) visual and spatial reasoning abilities.

Often an implementation appears to be doing something like human visualisation, but on closer examination lacks the generality and power: give it a slightly different problem and it cannot cope. There are now many wonderful systems for generating stunningly realistic static or moving images on computer displays, yet such programs cannot perceive and understand such images. Programs which can reason by manipulating diagrams containing a few discrete structures cannot cope with continuous structures or continuous change. In general, programs which reason about images using 2-D arrays or networks do not have a grasp of space or time as continuous. Work by Hayes (1985) and others related to the idea of 'naive physics' helps to define some aspects of the problem of characterising our grasp of spatial structure, but does not as far as I know specify mechanisms that can solve the problem.

Psychological and neural theories do not answer the questions either. Neural theories tend to identify locations where low-level visual processes occur, but say little or nothing about higher-level capabilities or how visualisation mechanisms are used in problem solving. When attempts are made to formulate theories about how brains do visual reasoning I usually find that they do not describe anything that I can interpret as a workable design with explanatory power. E.g. talking about mechanisms which "manipulate images" by rotating, or stretching or translating them explains nothing. It merely re-formulates what needs to be explained.

In order to add more detail to the specification of what needs to be explained, I have tried to show that visual reasoning covers a variety of different things, using two examples of what we can visualise: one a finite but deformable structure and one a discrete but infinite type of structure.

12 Spatial vs logical: what's the difference?

Introspectively, many people are convinced that there is a deep difference between solving problems by reasoning logically (or verbally) and solving them by visualising and transforming spatial structures. Whether such introspections are reliable is a matter of dispute.[4] However, it is not so commonly noticed that both sorts have much in common, and what they have in common is probably more important and harder to account for than the differences.[5]

Whenever we reason, whether with pictures, words, imagined movements, or anything else, processes occur in which structures are created and manipulated, usually in virtual machines. If you reason logically or algebraically using pencil and paper, you'll normally create a *sequence* of spatial structures, where the transition from one element of the sequence to the next corresponds to a step in the reasoning. (This is why visualisation of sequences plays such an important role in a lot of meta-mathematical reasoning.)

Problems in Euclidean geometry can often be solved without a spatial sequence: instead we modify a diagram *in situ*. (See Nelson (1993).) Modern interactive graphics technology supports this and also allows direct transformation of a single logical or algebraic structure presented on the screen without having to produce a sequence of spatially separate structures, as happens when we reason with sentences, equations, logical formulae. Perhaps brains got there first?

The collection of structure-manipulations possible in a class of structures defines a generalised notion of "syntax" for such structures. The kinds of parts that can be replaced and the kinds of features and relations that can be changed define the structural properties of the information medium, its syntax. We can also generalise a notion of "pragmatics" from linguistics, to refer to the functional roles of information structures in larger systems. In some cases there will also be "semantics" insofar as the structures are used to describe, summarise or plan, other internal or external structures, actions or goals.

We need a better grasp of the types of structure-manipulation mechanisms there are and the many ways in which different possibilities for further manipulation are actively made available by the current contents of a particular structure. This may enable us to come up with better theories of how brains or minds do all this. That would require, yet again, re-inventing ideas discovered long ago by evolution, and in the course of doing so we'll probably have to discard many of our cherished distinctions.

[4]Some of the differences between "Fregean" (applicative) and "analogical" representations were analysed in Sloman (1971). The differences are often misdescribed.

[5]I have previously argued that there are not only two categories, but a wide range of significantly different types of representation, e.g. in (Sloman, 1971; Sloman, 1975; Sloman, 1996b). Similar strictures apply to other alleged dichotomies, e.g. between implicit and explicit, computational and non-computational mechanisms, or procedural and declarative representations, etc.

13 Conclusion

This paper draws attention to a collection of unexplained features of our frequently noted ability to think and to visualise. All such cases (whether diagrammatic or not) seem to involve the ability to create structures – not necessarily the structures we think we are visualising, and not necessarily physical structures, since they can be structures in virtual machines (the "physical symbol system hypothesis" taken literally is a huge red herring). They also involve the ability to have readily available a collection of mechanisms for manipulating those structures which somehow implement our grasp of the possibilities for change inherent in a structure. The possibilities for change determine how the structure is grasped or understood, and provide the basis for its pragmatic and semantic functions.

What constitutes a grasp of something spatial as opposed to algebraic, or continuous as opposed to discrete, or finite as opposed to infinite, or linear as opposed to tree structured, or planar as opposed to three dimensional, etc. will depend in part on the collection of types of transformations and inferences available and ready to be applied to the structure.

In some cases the same structure may be viewed or understood in different ways by making different classes of transformations or inferences available, as in the difference between a metrical and a topological understanding of a spatial configuration.

Using such a grasp in solving a problem or making a plan involves somehow being able to orchestrate the collection of possible changes in such a way as to find collections of changes which satisfy some condition. When the situation represented is continuous, continuous changes can be visualised. Whether we can actually produce such changes or only convincing representations of them is not clear.

Being intelligent often involves simultaneously viewing something in two or more ways and relating the sets of possible changes in the different views. What does and does not work has to be learnt separately in the context of different classes of structures, different classes of manipulations and different classes of problems, which is why there is no such thing as totally general intelligence.

How all this can be implemented in brains or computers remains an open problem. If we study lots more special cases we may eventually understand what sorts of structures and mechanisms can implement such capabilities, and what sorts of general architecture can accommodate them all, along with closely related capabilities such as vision and motor control. I don't think this will be easy to do, not least because we still don't understand what the problem is.

Acknowledgements and Apologies

I have learnt much from reading papers by others who have written on these topics and from conversations I have since forgotten. My perspective was strongly influenced by reading Kant's views on the nature of mathematical knowledge (1781), with which most philosophers and logicians disagree, wrongly in my view. I apologise for not providing a literature review. Useful sources can be found in Brachman & Levesque (1985), Glasgow *et al.* (1995) and Peterson (1996). For inspiration see the examples in Nelson (1993).

References

Brachman, R. and Levesque, H., editors (1985). *Readings in knowledge representation*. Morgan Kaufmann, Los Altos, California.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York, Oxford.

Glasgow, J., Narayanan, H., and Chandrasekaran, B., editors (1995). *Diagrammatic Reasoning: Computational and Cognitive Perspectives*. MIT Press, Cambridge, Massachusetts.

Hayes, P. (1985). The second naive physics manifesto. pages 1–36. Ablex, Norwood, NJ. Also in (Brachman and Levesque, 1985), pp. 468–485.

Jamnik, M., Bundy, A., and Green, I. (1999). On automating diagrammatic proofs of arithmetic arguments. *Journal of Logic, Language and Information*, 8(3):297–321.

Karmiloff-Smith, A. (1996). Internal representations and external notations: a developmental perspective, in Peterson (1996), pages 141–151.

Nelsen, R. B. (1993). *Proofs without words: Exercises in Visual Thinking*. Mathematical Association of America, Washingon DC.

Peterson, D., editor (1996). *Forms of representation: an interdisciplinary theme for cognitive science*. Intellect Books, Exeter, U.K.

Sloman, A. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd IJCAI*, London. Reprinted in *Artificial Intelligence*, pp 209-225, 1971, and in J.M. Nicholas, ed. *Images, Perception, and Knowledge*. Dordrecht-Holland: Reidel. 1977.

Sloman, A. (1975). Afterthoughts on analogical representation. In Schank, R. and Nash-Webber, B., editors, *Theoretical Issues in Natural Language Processing (TINLAP)*, pages 431–439, MIT. Reprinted in (Brachman and Levesque, 1985).

Sloman, A. (1978). *The Computer Revolution in Philosophy*. Harvester Press (and Humanities Press), Hassocks, Sussex. Online at http://www.cs.bham.ac.uk/research/cogaff/crp.

Sloman, A. (1985). Why we need many knowledge representation formalisms. In Bramer, M., editor, *Research and Development in Expert Systems*, pages 163–183. Cambridge University Press.

Sloman, A. (1989). On designing a visual system (Towards a Gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, 1(4):289–337.

Sloman, A. (1992). The emperor's real mind. *Artificial Intelligence*, 56:355–396. Review of Roger Penrose's *The Emperor's new Mind: Concerning Computers Minds and the Laws of Physics*.

Sloman, A. (1993a). The mind as a control system. In Hookway, C. and Peterson, D., editors, *Philosophy and the Cognitive Sciences*, pages 69–110. Cambridge University Press, Cambridge, UK.

Sloman, A. (1993b). Varieties of formalisms for knowledge representation. *Computational Intelligence*, 9(4):413–423. (Special issue on Computational Imagery).

Sloman, A. (1996a). Actual possibilities. In Aiello, L. and Shapiro, S., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifth International Conference (KR '96)*, pages 627–638, Boston, MA. Morgan Kaufmann Publishers.

Sloman, A. (1996b). Towards a general theory of representations, in Peterson (1996). In D.M.Peterson, editor, *Forms of representation: an interdisciplinary theme for cognitive science*, pages 118–140. Intellect Books, Exeter, U.K.

Sloman, A. (2000a). Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In Dautenhahn, K., editor, *Human Cognition And Social Agent Technology*, Advances in Consciousness Research, pages 163–195. John Benjamins, Amsterdam.

Sloman, A. (2000b). Models of models of mind, in *Proceedings Symposium on How to Design a Functioning Mind* AISB'00, Birmingham, April 2000. pages 1–9.

Sloman, A. and Logan, B. (2000). Evolvable architectures for human-like minds. In Hatano, G., Okada, N., and Tanabe, H., editors, *Affective Minds*, pages 169–181. Elsevier, Amsterdam.

Sussman, G. (1975). *A Computational Model of Skill Acquisition*. American Elsevier.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell, Oxford. (2nd edition 1958).