

# On the Sanity of Engines of Reason

## DRAFT VERSION 0.2

67<sup>o</sup>

[www.concept67.net](http://www.concept67.net)

(Dated: 11th March 2011)

The forthcoming rise of the assistive reasoning engine (ARE) in our social and media contexts is considered, and the view is advanced that this will result in the generation of novel ways to represent, assimilate, and interchange knowledge. In support of this view, the work of several research groups around the world is highlighted to provide some assessment of how assistive and how automated an ARE might soon be. The view that the 'reasoning' of an effective ARE might not conform to classical conventions is also presented. A limited discussion of the benefits and the profound ethical implications of the deployment of AREs within society is then also provided.

### I. THE RISE OF ASSISTIVE REASONING ENGINES

One significant challenge people face in this epoch is the ability of any individual to gather a sufficiently comprehensive overview of some really complex issue in order to make progress either in researches or in debates. As examples, complex legal cases, or the development of government policies, often involve thousands of pages of documents containing both profound and highly technical arguments. The expansion in exposure of information via the Internet has opened new horizons and altered the form of many challenges, and there has been a concomitant increase in the power of humanity's technical capabilities for manipulating data computationally. Arguably, prior to this epoch, it has only been people who have been able to determine when data can be interpreted as informative in some context. But the expanding problem of 'infoglut' is driving a desire to be able to delegate to computational processes some of the responsibility for sifting through huge and complex corpuses of documents for salient information. If computers could be programmed to find salient information, then along with this overarching desire comes a desire to know just how much delegation can be made. How 'smart' could these computational processes be? Beyond finding salient information, could they also 'join the dots' to produce cogent outlines or arguments?

Henceforth, such computational processes will be called assistive reasoning engines (ARE), rather than artificial intelligence(s) (AI). This is partly because the AI tradition has always had big dreams of making their programs sentient and autonomous reasoners, with reasoning powers that matched an adult human. But it is also because traditional AI had a tendency to view intelligence as arising from (the top of) a single process. New computational techniques and new scientific approaches have pointed to the effectiveness of agent swarms and the possibility of emergent behaviours arising from agent collectives. So it is fitting to speak of a broader class of processes (the ARE), that probably won't be sentient, autonomous, isolated instances, nor particularly 'intelligent', when discussing how much responsibility for information handling tasks could be del-

egated.

A key thread that will run through this paper also concerns the extent to which thinking on the matter of how to make computers think has previously been dominated by an approach that might be fundamentally misguided. In delving deeper into the question of how human reasoning works, a picture of reasoning is constructed that indicates not only that future research into AREs might take a very different course from earlier AI, but also that reasoning is not what people normally think it is. In short, when considering the potential roles of useful AREs it might also become necessary to consider their 'sanity' in a sense not normally assigned to computer programs, and the relationship between that 'sanity' and safety.

As a starting point for such investigations, the next section considers the relationships between data, information, and the accrual of knowledge.

### II. PATTERNS

For the purpose of exposition, it will help to provide an initial definition of the term 'knowledge'. One could take the view that:

*Thus, one person's knowledge is another person's data. The borderlines between data, information, and knowledge are not sharp, because they are relative with respect to the context of use.[1]*

The quote above was also made in a particular context of use, but if it is presumed to apply too widely then counterarguments can be raised. For one thing, it is noted that all thinking processes have to be bootstrapped with some data input. Arguably, a newborn child that has no connection between sensory perception and mind will not develop cogent thought processes. Likewise, an assistive reasoning engine (ARE) is of no use if it cannot act over data provided to it. So data is fundamental; but what is data? Living creatures have sensory faculties that generate an internal signal

stream that is an encoding of a local subset of their environment. It is a local subset because the creatures are finite and have finite purview. This internal signal stream seems like a good candidate for being ‘data’. The internal signal stream then passes into various pattern recognition faculties. Conceivably, the detection of a pattern and the generation of information are the same act. Information is then the set of patterns detected (however subconsciously and automatically) in the data. It is worth stating at this point that ‘information’ for most creatures is merely the firing of certain integrating neurons.

Following from this, the question of the nature of knowledge might require the introduction of a faculty of memory. It seems likely that many creatures on Earth have nervous systems with limited plasticity or capacity for rewiring. Insofar as information is generated within their nervous systems it directly triggers response actions. However, as an example, a squirrel that buries nuts in the autumn retains a memory of where it buried them that it can later recollect and act upon. Humans tend to anthropomorphise this by saying that the squirrel knows where the nuts are buried. Memory, in this case, might be the storage in the squirrel’s brain of sets of specific sequences of actions — patterns of actions — where these actions are (for the brain) also patterns. Memory is thus pattern storage, and in terms of the descriptions above this is information storage. Is knowledge, therefore, the storage of information for later recollection and use as guides of future actions?

Humans like to think of themselves as being vastly intellectually superior to squirrels, but this approach to the definition of knowledge seems to capture much of how we ordinarily use the word. In the context of human knowledge matters are complicated by our capacity to assimilate new information to generate new knowledge, which can in turn be used to assimilate even more novel information, and so on. In ordinary (English) language the term ‘assimilate’ carries with it connotations of incorporation, adsorption, and “bringing into harmony”. In human minds, knowledge tends to be assimilated rather than federated. The distinction perhaps concerns the reversibility of the incorporation process. When a person assimilates new information this often affects other memories and the new information can become so strongly integrated that it can be difficult to recover what the precise inputs were. However, where written and digital corpora are concerned (as the collective human memory) both the precise input and its interaction with surrounding information can be recorded. In principle then, in knowledge *federation* the process of incorporation can be reversed.

The term ‘knowledge federation’ can have many interpretations depending on the definition or description of knowledge that is given. There can be knowledge federation at different levels of granularity, from single symbols up to collections of media objects and databases. But it is also worth considering whether the

federation of existing symbols (such as words) together with new approaches to the issue of how knowledge can be encoded is likely to result in the generation of new types of knowledge to be federated. Later on, it will be suggested that the demand for effective assistive reasoning engines (AREs) is likely to result in new methods for representing knowledge in computers. Moreover, that optimising the efficiency of federation might mean that some hitherto rare forms of knowledge interchange are likely to burgeon on the internet.

Many people might start to conceive of digital knowledge federation (KF) as being a matter of scanning through texts, pulling out various words and their interrelations (however these are generated), and simply mapping those words on to their counterparts within some big ‘topic’ (word or phrase) interrelation network. However, there are additional issues to consider. There is an assumption, that appears in some artificial intelligence (AI) circles, that as human minds accrue more information about the world then new, more correct, information supercedes previous information; and the previous (now obsolete) information gets relegated to the status of history. This could be true if the average person had the same needs and error correcting procedures as the cutting edge of science as a whole. Also, there is a tendency to presume that classical models of logic based on strict polarities, black/white, right/wrong, adequately suffice to talk about the world. Unfortunately, proponents of this view have not really looked closely at the way science does its extraordinary best to talk about the world. When physical science obtains data, the subsequent pattern matching process that generates information usually does not result in the kind of pure symbols, like 0 and 1, of which computer scientists are fond. Much information from science involves statistics and confidence intervals on probability distributions.

Hence, one of the strands this paper seeks to develop is the view that for AREs to be effective the text symbol federation approach, and simple logics imposed over those symbols, might represent a primitive bootstrap mechanism, but these might rapidly need to be superceded by more sophisticated approaches. The next section looks at other reasons why this could be the case.

### III. SMOOTH THINKERS

What follows is a sketch forecast of the evolution of reasoning engines made by an amateur onlooker [10]. The sketch is based on a humble argument whose general drive can be taken to be focused around the question: What are the odds that reasoning engines will exhibit behaviour that is reliable and safe? More thought will be given to the terms ‘reasoning engine’, ‘behaviour’, ‘reliable’ and ‘safe’ in a while. Whereupon the relationship of ideas about knowledge federation to this sketch will become clearer. For now though, to give

some weight to the intent of this sketch, it will be good to think about the stance of Strong Artificial Intelligence (or Strong AI) debated against in [2].

All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.

Rather than focus immediately on the issue of ‘feelings of conscious awareness’, in what follows the concern is that more elucidation of the relationship between ‘reasoning’ and computation is needed. The quote above presumes that thinking is computation, but which computations are ‘appropriate computations’?

To begin, it will be good to identify the concept of computer with the notion of a Turing machine (a universal computing engine). There are different classes of Turing machine, but the most general class is that with infinite tape length and potentially infinite process memory (for holding all the state-switching instructions – also known as the program). The Infinite Turing Machine (ITM) can ‘compute’ anything, given sufficient time (that is, assuming switching states takes time). Note how general this is; in particular note how it says nothing at all about whether, for some program, one state follows another in any sense that would be called ‘rational’ or ‘logical’ by a human onlooker.

Moving into slightly less abstract territory, a subclass of the ITM is the Finite Turing Machines (FTM). These have finite tapes and finite programs. At this point, it is presumed that Strong AI can then be invoked: Humans are finite and are therefore FTMs. More specifically, one can consider that subset of all possible FTMs that appear able to replicate the evolution of states of human minds. But then there is a problem. This subset of FTMs will contain not only those FTMs whose chains of states replicate, for example, the thinking of a top mathematician writing a great proof. It will also contain all FTMs, according to the full concept of ‘human mind’, between ‘top mathematician’ and ‘severely mentally ill person’. It is likely that a large portion of these do not behave in ways that conform to classical logic.

Now assume that the emphasis is not on creating replicas of human minds within computers, but on creating reasoning engines to assist with various tasks in life. Thus, the subset of FTMs to consider can be narrowed even further if the need for conscious awareness is done away with. For these assistive reasoning engines (ARE) to be useful, it is then incumbent upon their designers to consider how to make the behaviour of AREs reliable and safe – or put another way, reasonable.

Note how, as this argument has been constructed, it brings with it a tendency to think of reasoning in terms of a Turing machine clunking along between precisely defined states. First it is in state S1, and then, clunk, it is in state S2, clunk, S3, and so on. The catch is that

the presumption that a human mind goes through a finite set of states could be wrong. Take the mathematical proofs that there are an infinite number of points within any bounded interval of the real number line. There is an ordered set of numbers that mark distances along such an interval, relative to one of the endpoints. Perfect continuity requires an infinity of numbers therein, so that no matter how much a microscope might zoom in a viewer would always see a smooth line and could select any distance with any degree of accuracy (in the world of mathematics). Since there is no evidence to contradict the view that the (physical) passage of time in the world is a smooth, continuous passage, it is presumed that time can be treated as a continuum. Now the question becomes that of whether a human mind clunks through discrete states over time. Even without invoking arguments about quantum wave functions, given that a human brain is a physical object in space-time, comprised of billions of atoms undergoing the kind of turnover of position common to organic matter, there is nothing about the brain that is discrete. The atoms do not clunk about. Therefore it seems appropriate to question whether there is really anything discrete about the states of mind that a brain has.

Much of the argument for the discrete states of mind hinges upon a view of a neural network in which neurons are either switched off or firing. However, evidence of the ongoing material plasticity[3] of the nerve cells that underlies neural learning suggests continuous subtle evolution of the neural network; and the roles of hormones in the brain are not yet fully understood either. Electrical signals between neurons pass through continuously evolving networks of dendrites and axons, and it is possible that the transit of these signals plays a key role in cognition, not just the ‘spiking’ of the nerve cells themselves. Moreover, observation of the spiking with a sufficiently fine time resolution undermines the idea that neurons are just ‘on’ or ‘off’ in a discrete binary way. Interference with these signalling processes, from disease or chemicals like alcohol, does not necessarily degrade (or improve?) function in simple on/off ways. So it is unclear as to what is meant by suggesting a relationship between a brain and a (temporarily) fixed state of mind, or beliefs, or knowledge. Brains are continually exposed to new data, continually derive new information, and continually generate new knowledge, but it is hard to see which, if any, of these processes have genuine discreteness, not least because the underlying material processes do not.

When a mind generates a conclusion from some premises in a classical syllogism, it might appear that there is some switch clunking from 0 to 1, but this could be a delusion. Imagine if the premises are unfamiliar, then ask what happens in the process of generating a conclusion. Or ask what happens if the premises are such that a ‘conclusion’ looks more like a probability distribution. Likewise, instead of the view of a mind clunking through a finite set of states over time, it might

be better to think of minds as sweeping along a continuous trajectory. This trajectory might have finite temporal extent, and any point along the trajectory might have finite spatial extent (assuming materialism about brains and minds) but between the endpoints the ride might be exactly as smooth as a smooth function. This is because there is not yet any way to tell what the smallest unit of action within a brain is. One might presume all of present microphysical theory, posit the smallest unit of action at quantum level, and still be wrong. Present experimental physics can measure transitions between quantum states in an electron, but still cannot say much about whether any aspect of the transition between states is important. That is, it is known that the phenomenal accuracy of present empirical science is a long way from matching the smooth continuum in mathematics. What physical theory can presently ‘get away with’ is not necessarily the full truth of the matter. Imagine if quantum transitions, rather than merely quantum states, turn out to have some as yet unknown pattern of coordination of movements.

Such a view of the smoothness of trajectory of a thinking process still leaves many questions unanswered. Take as an example the mind of the mathematician John Nash, the Nobel prize-winning mathematician who was also afflicted with quite severe mental illness[11] at some periods of his life. Nash himself asserted in later life that he had trusted the thoughts he had during his periods of ‘madness’ because they “arose from the same source” as his mathematical insights[4]. Therefore, whatever physical continuity a thinking process might have, this is not sufficient to guarantee ‘reasonable’ behaviour. But it might be acceptable to assume that whether thinking behaviour is ‘reasonable’ or not is related to some property of the trajectory. However, this gives rise to the question of whether the insights of Nash’s mathematical genius could have been acquired by any process conforming to some yet-to-be-defined notion of ‘reasonable’. That is, would being ‘reasonable’ exclude some important and valuable thinking process trajectories?

Furthermore, there is still the question of whether a finite process within a FTM — here meaning a digital computer program clunking through states on some microchips — can adequately approximate a real (possibly as smooth as a mathematical continuum) thinking process. One possibility is that the quality of the approximation might hinge upon whether various integrals can be represented and their combinations computed purely via algebraic representations and manipulations. More on this later.

#### IV. COMPLEX WORLD, STATISTICAL MIND

To recap briefly, arguments have been put forth to suggest that human information (used as knowledge) is in reality inherently more complex than just lexical sym-

bols or phrases, and that the processes of the human brain and mind are more complex and subtle than any approximating *finite* discrete state symbol processing machine (a.k.a. a computer). If an effective ARE with reasonable behaviour is to be developed there has to be a way to map from the capabilities of a computer and its programs to handling the real subtleties of operating with information generated by people and their science.

A possible path might arise from the possibility of programming computers to handle symbol processing and numerical processing with sufficient speed and resolution to evolve a mathematical model of state. By which is meant, that it might be possible to capture the evolution of state in an abstract mathematical space — for example, as the meandering of a point (or points) upon a high-dimensional manifold. If the evolution of the path of this point over time can always be given a symbolic mathematical description, or even if enough of the contributions to its movement can be expressed in that way, there arises the possibility of using mostly symbolic (algebraic maths) processing to evolve state. This might not be a simulacrum of a human mind or brain, but it might be useful.

Through such a method, it might become possible to study reason empirically. That is, in the same way that Immanuel Kant argued that the extent of human powers of reason is dictated by the form of the resources that nature equipped brains with, such a mathematical method might allow rapid experimentation with how different forms of ARE operate. For example, modification of the number and scope of data inputs, of information (feature) detectors, and of integrations (and federations) of forms of information, could simultaneously define and constrain both the form of the mathematical manifold and the possible moves across its hyper-surface. Comparing outputs of state from differing AREs could then provide a basis for a real science of reason.

In a sense, the arguments earlier were about whether humanity has been somewhat mistaken in its approach as to what reasoning is. By more careful consideration of the constituents of reason, it is argued, perhaps a general model of a reasoning process can be provided. If feasible, this model would provide a number of advantages. The first has already been mentioned: a science of reason, based on the study of reasoning engines with generic mathematically defined properties. Having a prior science of reason raises the likelihood of developing effective, reasonable AREs. Secondly, while AREs probably won’t provide exact simulacra of human minds, it is not clear that is necessary for them to do so. A general model of reason with mathematical properties raises the possibility that some forms of reasoning can exist in FTMs with little or no resemblance to brains, in software or hardware. That is, neural substrates might not be the only way to build reasoning capacities. However, it is suggested that old-school lexical token inference will prove to be insufficient, and that a deeper

mathematical model of inference is required. Thirdly, AREs will probably be capable of federating knowledge — federating themselves, to some extent — to provide an entirely new model for the development and provision of various services.

## V. BACK TO REALITY

So how realistic is the dream of providing an effective, reasonable ARE of a form similar to that described above? If a basic list of requirements was extracted from the above discussion it might include: transducers[12] that generate data; information extraction and error checking; symbolic representation of salient information; memory and recall; and a coherent underlying mathematical model. The following sections contain brief discussions of some examples[13] of recent research that indicate how close the possibility of an effective ARE might be.

### A. Sensing Features

In [5] research into using a new general framework for computational object recognition from 2-dimensional images is discussed. The paper describes “a hierarchical system that closely follows the organization of visual cortex and builds an increasingly complex and invariant feature representation by alternating between a template matching and a maximum pooling operation.” A number of aspects of [5] are noteworthy from the perspective of the present discussion. The first is that the model used does not simulate neurons, but uses “computational units” that process input data according to defined mathematical functions. These mathematical functions capture the essential behaviour of components within the primate visual cortex. Secondly, this processing is in some sense statistical, and a key stage involves integration. The pooling process (within a pooling layer) in the neighbourhood of a position of the new input uses a radial basis function with Gaussian weighting to assess the quality of match to a previously learned prototype. Those units with the maximal value contribute to the next layer of processing. The authors state, “Our work seems novel in that general purpose filters are being maximized over uniformly distributed local regions in the image.” The third aspect of interest is that their system can learn a ‘universal set’ of 10,000 feature prototypes, “from a set of random natural images (downloaded from the Web),” and it still performs competitively. Lastly, the performance of this biologically inspired approach and the measures that could be taken to improve it indicate great promise. The authors of [5] remark that, “It is important to point out that this recognition with a *glimpse* only constitutes the initial processing step in natural vision.” However, they also note that there is already a great deal of experimental evi-

dence about the primate cortex that could be drawn on to provide extensions.

The fact that the method uses “general purpose filters” over “uniformly distributed local regions” also raises the possibility that recognised objects could be abstracted out and represented by mathematical functions. This is an aspect that will be returned to later. For now though, it is enough to note that significant progress is being made in key mechanisms that would permit an ARE to recognise and classify aspects of the real world with some (in principle, quantifiable) measure of confidence. In addition, though the system in [5] is based on the primate visual cortex there is nothing to prevent the method being used with an ARE using transducers that operate in ranges not accessible to human senses — for example, at infrared, ultraviolet or microwave.

### B. Salient Information

In 2009, a report[6] detailing work on computationally distilling natural laws from experimental data sets was published. The natural laws were represented by symbolic mathematical expressions, with the computer searching through a space of generated formulae to find those with the best fit to the data. The process of finding the optimal formula was termed symbolic regression. In the case of one of the experiments, the computer took almost 40 hours to find an optimal formula, but this amount of time is small compared to the centuries of science prior to most original formulations by human scientists like Newton and Lagrange. Moreover, while some technical details of the approach have been challenged and improvements proposed[7], the core approach still holds promise.

In the context of the present discussion, what is most notable about the work in [6] is that it has shown the possibility of the computational distillation of real world data down to mathematical formulae. It presents the possibility that AREs might learn based on parallel sampling of the environment, followed by networked interchange of mathematical results. On the assumption that AREs can be constructed to operate in reasonable ways, one could imagine that, at the very least, this might alter many aspects of the scientific process. Perhaps, scientists might find themselves with more time to discuss *why* certain formulae are better than others for reconstructing the behaviour of some real world phenomena. It also raises the possibility that one of the principal types of knowledge federated by AREs will be mathematical formulae. These might play a role in upgrading ARE capabilities as well capturing new knowledge to benefit humanity. Since there is, theoretically, nothing to stop a suitably configured ARE spawning a copy of itself to test a new configuration, AREs might also manage to evolve their development. Putting aside the horrors of science-fiction B movies in which computers take over the world, the self-testing of new con-

figurations might assist in preventing attacks from computer virus and cracking intrusions.

### C. The Geometry of Learning

Hopefully, readers of this paper will by now have detected a theme: That increasingly, mathematical techniques and objects will form the core and, in some sense, one of the native languages of future AREs, if these are feasible. It seems likely that the developments for discussing the behaviours of AREs in a scientific way will require expression in mathematics, and that mathematics will underwrite any future science of reason. In particular, as mentioned earlier, coherent discussion of the state of an ARE will require an underlying general mathematical model. Progress in this direction can already be found in [8]. Therein, Sumio Watanabe introduces a mathematical model that provides a unifying theoretical foundation for discussing the properties of diverse statistical learning methods and machines. Previously, many statistical learning machines had no underlying theoretical treatment because their models were singular — meaning that their parameter spaces contained discontinuities. With a coupling of algebraic geometry and singularity theory, Watanabe is able to prove that a unified model is possible, and his work also provides for new results about the behaviours of existing statistical learning techniques.

While the theory in [8] was created by a human genius, and not a computational formula distiller, it can be seen in some senses as a remarkably creative distillation. The techniques involved may also possess wider applicability. For the relationship between aspects of computational methods and algebraic geometry can be applied to key aspects of the research examples in [5] and [6]. While the process in [6] is already distilling formula from data that contain a certain amount of noise, the research in [8] provides theory for processes that handle learning from datasets that could have statistical distributions. Moreover, the research in [5] raises the possibility that a ‘state of mind’ is actually a statistical expectation[14] — an averaging over some collection of maximally likely possible states positioned (at some moment in time) on some high-dimensional manifold. The mathematical techniques in [8] might thus play a crucial role in any future capacity to describe and study AREs.

## VI. PARTIAL KNOWLEDGE

Having introduced the idea that a subtler model of reason might be necessary this then implies that a different account of the link between human language and reason might also be required. As an ARE might principally operate over numerical and lexical data an account has to be given of the relationship between this data and the mathematical properties of the model of

reason. Evidence in [5] (and references therein) suggests that, in primates at least, the transition from data signal to information occurs through layers of processing that have *statistical* and *integrative* characteristics.

Thus, owing to the way human nervous systems work, all knowledge rests upon statistical foundations, only more or less certain. In addition, the conditions in which perception takes place can be disadvantageous. In fog, for example, a face might be hard to identify, or an object might be considered to be two different colours at once. Properties that were previously used to identify an object no longer supply enough data for ambiguity to be removed. Meanings in natural language can have similar properties. Though it can be a struggle to identify an object in perception, words can also often lack precise references, precise application to the world. As a result, a need for extended description can arise. In English, for example, we might say that an object is a book, or like a book. However, the set of discriminating properties underlying the application of the word ‘book’ can each be applied more or less, so there can be borderline cases where something is a bit like a book, but also like something else. So there is a struggle to create an identifying description. There is also the fact that, within a text, the description of an object can change over time, so that while it might be referred to by a single word, the reappearance of the word in itself by no means captures all that is intended.

The ‘shifting sands’ of perceptual and semantic leeway has consequences for the use of the word ‘logic’. Classical, ‘hard’ notions of the identity of terms and objects might need to be supplanted by a subtler account[15]. In what follows the classical term ‘identify’ should be taken to imply the detection of a sufficient number of characteristics present in sufficient degree so as to remove enough uncertainty as to whether two experiences of an object, separated in time or space, concern the same object. The use of the term ‘enough’ here is deliberate — everything in the universe known to contemporary physics is subtly altered by the passage of time. Therefore the concept of identification has to allow for these changes. The mention of ‘separated in space’ covers the mathematical (or ideal) case of, for example, two congruent triangles some distance apart, that can be brought into perfect coincidence of all their points via some translation. When they are in perfect coincidence it is possible to cease to consider them as two triangles and identify them in a single representative triangle. Note however, that the (augmented) definition of ‘identify’ leaves open the real possibility that a person encountering each of two ‘identical’ twins in separate experiences can be mistaken about considering them (by removing enough uncertainty) to be the same one person.

Now it becomes pertinent to ask what mechanisms underlie the capacity of people to identify objects. Clearly, in order to detect a sufficient number of characteristics of some object, a person would need to have

had prior experience of that object. Someone encountering some (highly distinctive!) squirrel for the first time is not likely to remark that it is in fact Sidney the squirrel and is their neighbour's pet. Therefore, identification in the real world rests upon memory —there is an interposed time interval. Ideal mathematical identification however, is in a sense the deliberate removal of any dependence upon memory, until properties are entirely coincident. But it relies on the ability to suggest initially that, for example, two triangles are congruent, which in turn relies upon memory of key characteristics. Earlier in this paper (§II) it was asserted that memory is pattern storage, and now it is possible to say more about what comprises a pattern. A pattern is made up of characteristics, and these latter can be spatially related and temporally ordered.

As an aside, it is interesting to note how advertising gains its power from compelling the retention of characteristics mashed together in some artificially constructed pattern. Accounts of human rationality often resort to distinguishing between rationality proper and human susceptibility to advertising and the creation of impulses. This is the same concern as is involved in the desire for an ARE with *reasonable* behaviour. However, assume for a moment that cognition just is traversing through some realm of retained patterns, and that the (continuous in time) trajectory taken in that traversal is also heavily influenced (if not controlled outright) by retained patterns. What then distinguishes reasonable from unreasonable? At the end of §III it was suggested that some property related to the trajectory would determine whether thinking behaviour is reasonable or not. The notion of the accountability of some trajectory rests upon some criteria as to what makes sense. This becomes circular if it is just asserted that sensible people determine what makes sense. It therefore becomes necessary to dig deeper and ask how sense is made, as contrasted with nonsense. From the example of advertising it is clear that it is not the content of object patterns that matters here, but the ways in which it is possible to traverse through them — how they are interconnected.

Before discussing these properties of trajectories further, note that it is now possible to connect the account of a trajectory of a set of points on some high-dimensional manifold with the idea of traversing through a realm of patterns to give a limited view of how reason might operate. Imagine an ARE with only one pattern in memory. Obviously there is nowhere to traverse to, so this ARE cannot be said to be reasoning. Now imagine an ARE with as many patterns in memory as an average person, but as time passes this ARE sees all of these patterns at once, all the time. Again, this ARE cannot be said to be reasoning, as there is no traversal, no movement. Reasoning therefore involves, at minimum, traversal through some subset of the available patterns in order to cope with (or perhaps even be creative into) some situation in the world. Moreover, on the assumption that trajectories are continuous in

time and that characteristics are statistically detected, is it possible to view reason as blending/morphing paths between patterns? This might be the reality that underlies the way people consciously think, with all the interleaved blending between patterns with a concerted degree of identity lost under the surface in the subconscious. Although this is speculation, it could account for the extent to which people find it hard to detect and override insufficiently examined links between patterns. A means to examine these links might therefore provide a basis for assessing what is reasonable.

## VII. PATHFINDERS

In the realm of human action, what is considered reasonable, tends to be any action that is either harmless or constructive, or in the case of defence, a justifiably proportionate response. In the case of a human being faced with a new situation, there is a great deal of mental processing related to selecting the most constructive action, according to some subjective metric as to what is constructive. It is conceivable that this subjective metric has to be learned. (It is also conceivable that psychopaths lack the ability to construct or retain a metric involving compassion for, or empathy with, others.) However, as noted above, what might be more important is the way a person's mind is able to traverse between patterns in unexamined ways. A person might believe they are reasoning well, but, assuming they have sufficient patterns, in fact prior reinforcements of certain connections leads them to choose outcomes that could be suboptimal (which is to say that they could do better if only they knew how). Not only does this account require sufficient patterns and sufficient ways of traversing smoothly between subsets of patterns, but it requires sufficient prior context-dependent influences of trajectory to create and reinforce 'reasonable' predictions and choices. These prior context-dependent influences are guided by error-checking procedures. It becomes possible to suggest that John Nash's illness arose not from lack of patterns or routes to traverse between them, but in a subtle loss of error-checking of the traversals between subsets of patterns and the concurrent connections between them, even as new information was pouring into his mind. (For someone with a thoroughbred mind, perhaps even small doses of alcohol, or dietary intolerances, were problematic triggers in this respect.) According to this argument, reasonable minds have better error-checking procedures. However, an addition to this account could also be offered. Reasonable people not only have better error-checking procedures but perhaps they are also concomitantly slower to move to action.

This is not to argue that an ARE must only operate with information that it is one hundred percent certain about. In this respect an ARE might need to operate more in the manner of an intelligence analyst than

a scientist. That is, an ARE would need to reason with uncertainty, much as people do, while keeping track of how much confidence could be given to identifications and predictions. The catch involved in this approach is that the ARE is continuously maintaining, simultaneously, multiple possible views of almost all states of affairs, with each view holding some degree of influence. Here again, there might be a need to envision the simultaneity of views held concurrently not as a set of distinct objects, views  $V = \{a, b, c, d, \dots\}$ , but more as a totality — perhaps somewhat like an animated impressionist painting. Blobs of colour and forms morph continually, while somehow entailing, for people at least, a plurality of semantics in respect of the world. Given an account like this, it is not difficult to postulate an analogy with the movements of clouds of neural activity seen on time-lapsed MRI scans of human brains.

So what does it mean to have error-checking on the flow of patterns, on unexamined traversals? For an ARE, the possibility of its being able to introspect its own traversals arises. One possibility hinted at above, is that the ARE keeps track of the amount of information, in its current account of some state of affairs, that it is more than, say, 95% certain about. Those aspects of the state of affairs would then carry weight when the ARE generates plans, while others would be discounted. Alternatively, it has to be asked whether plans for action are really separate strands, or whether statistical accounts of aspects of state allow for a view of planning more akin to a plume of injected ink in a fluid. Owing to uncertainties about the current state of affairs, there would be a narrow probability distribution of initial conditions around the current point with a continuous plume of future trajectory extending from it. Any single curve within the plume would represent a calculation about where to go next. Some curves could be ‘denser’ than others, akin to fluid streamlines, indicating the heavier influence of information the ARE is sufficiently confident about. Being reasonable might then be thought of as remaining within regions of the (continuously evolving) plume where the ‘density’ is high enough. Note also that this description of planning and flow requires a means of description that is sufficiently continuous, demanding a mathematical approach.

#### A. Language and Simulation

Given an account of cognition like the above, it becomes possible to suggest that the way language operates is both bidirectional and flexible. Bidirectional, because people have the capacity to consciously label some set of features (detected using statistical mechanisms like those in [5]) with a word or phrase, but they are also able to ‘reignite’ those feature detectors when only the word is perceived and recognised. Flexible, because the underlying mechanisms of feature detection have statistical properties, so the interplay between fea-

ture detection and word is somewhat elastic. To think about words more deeply, a word is at minimum *simultaneously* a prototyping necessary pattern set stored in memory for each of: a written object seen in the world, a phoneme stream heard in the world, an event stream stored in memory (to reproduce the sound when spoken), a statistical object simulation/recognition capacity, and a set of connections to other patterns (perceptual or behavioural)[16]. If this is the case, then there is an argument for saying that linguistic communication works because words can drive rich mental simulations. Moreover, that as long as the mental simulations of two or more people have enough overlap (Note the deliberately vague ‘enough!’), then these people can be said to share some understanding. However, they might not share opinions, insofar as the case that opinions and planning capacities are related can be made, as their planning ‘plumes’ might be affected by even small discrepancies of stored patterns[17].

With AREs the picture would probably be rather different, because they could in principle exchange and federate: the data of observation; the information patterns generated; the mathematical formulas and their instantiating algorithms that enabled information capture; further information about their precise internal state (their point positions and distributions upon the high-dimensional manifold as postulated in §IV); and so on. Therefore, in principle, AREs could operate with a much tighter tolerance for shared understanding than people normally would. The potential for AREs to have ‘hive-reasoner’ capabilities might in this aspect outstrip human capacities. This then raises questions about the interface between people and AREs. Humans might, at some time in the future, be able to construct effective, reasonable AREs, but will they be able to get along with them?

### VIII. LEGAL DIMENSIONS

In 1790, Immanuel Kant published his Critique of Judgement, a text that investigated the connection between reason, with its capacity to consider all sorts of plans, and morality, which might consider some plans to be harmful. This connection necessarily involves a filtering of plans for action based upon some notion of compassion, and for Kant this lay in the deliberate postulation of an ideal rational being (albeit rather anthropomorphized). One might argue that the reason Kant had to resort to postulating an ideal rational being was because the perfect moral person is hard to find, let alone a generic morality. As with intellect, it could be suggested that there are almost as many different moralities as there are people on the planet, and that these range across a spectrum from ‘positively angelic’ to ‘hideously demonic’ (or similar metaphors). In reality, the closest people get to a generic morality is in their legal systems — specifically the legal systems of their nations. It is

thus more meaningful to describe people as law-abiding than it is to describe them as moral. A neat feature of this approach is that the existence of a written corpus of law means that people can *learn* what to do or not do. Though some people have been known to learn aspects of the their laws and then deliberately not abide by them. On the assumption that an ARE could learn about laws, could it also be constrained to respect them in its choice of actions?

In the 1940s, Isaac Asimov wrote the science-fiction stories that were eventually collected under the title “I, Robot”(1950), which contained his formulation of the “Three Law of Robotics” as a plot device[18]. The plot device was intended to be exploited to show that any simple attempt to codify ethics for robots was doomed to failure. For comparison, there is also a contemporary set of robot laws in [9]. The laws therein rest upon restricting deployments of AREs by humans unto, “the highest legal and professional standards of safety and ethics”, which, in a sense, sidesteps any attempt to actually imbue AREs with intrinsic ethical judgement. Most traditional AI takes the approach of codifying all the parts of law considered relevant to some robot’s task as rules and putting the rules in a database. When the robot decides that it is in situation *X* and makes a set of plans *Y*, each one of the set of plans must be compared against all (detectably) relevant rules in the database of laws. This is often too slow even with substantial processing power allocated to the task. However, the model for cognition presented in this paper points to another approach.

A recap of relevant parts of this paper could be: Brains learn patterns and learn associations between them. Possibly, all forms of cognition rest on (statistical) pattern associations, and there is a fundamental connection between simulation and cognition. Possibly, cognition, in many respects, just *is* simulation. For an ARE, patterns and their associations can be captured as mathematical properties of some high-dimensional model. The reasoning processes of an ARE involve traversals over the manifold of this model. AREs could also, in principle, exchange all sorts of information about their states and processes.

Armed with this account, it could be argued that the key to an ethical ARE is in training it to associate certain marking patterns, such as ‘legal requirement’ and ‘not legal’, with other pattern sets. That is, the capacity for an ARE to be law-abiding lies in filling out the pattern sets with additional connections. As an extreme example, people would normally condemn a person who put another person on a table and then stuck a big knife into their chest. But if the person on the table is anaesthetised, gave prior consent, and the person with the big knife is a qualified surgeon operating in a hospital then the pattern set (often) becomes a legal one. The set of patterns will determine the state of the ARE, given as a

position on a high-dimensional manifold. Since the two pattern sets can have distinct positions on this manifold (assuming rich enough feature sets and memory capacities), additional properties concerning legality might be readily assigned as a value on an extra dimension.

The learning speed for AREs could come from being able to run many simulations in parallel and disseminating the results. The processing speed could come from the fact that perception and simulation are related, using the same underlying mechanisms, so that pattern sets once learnt would come automatically tinted as to their legality. It also matters for the purposes of avoiding expensive combinatorics, for example, that the scenario where the surgeon operates on a person is considered to be the more frequent and usual situation. Thus, the absence of key patterns would rapidly suggest illegality. The statistical properties of pattern sets could then be used to assess large patches of the planning ‘plume’ based on pattern set blends and overlaps. The planning plume as a whole might be ‘coloured’ with the measure of legality of the points within it, facilitating rapid assessment of legal trajectories.

## IX. IN CONCLUSION

Hopefully, the preceding sections have provided some means by which to assess the plausibility of constructing an effective, reasonable, even ethical, ARE. Assuming that, armed with a new science of reason, it becomes possible to construct such AREs and deploy them widely, how is this likely to transform society? One obvious response might be that, within the current economic paradigm, a great many people might be made redundant. Almost all roles within society that have an intellectual component could be vulnerable to some extent. A dystopian vision might also suggest that, if left to evolve, AREs will eventually become AIs, and people will become tools used by AIs rather than the other way around. In addition, there is the (almost) inevitable exploitation of such technologies for criminal or military aims. While it is worth considering these possibilities, it is also worth acknowledging that there is considerable scope in the use of effective AREs for educational and other training needs, and for providing intellectual assistance in almost all spheres of life. Children and old people, especially, might benefit from such assistants when faced with an increasingly complex and, in some senses, difficult world.

Would the advent of AREs transform the world? Yes, undoubtedly. Moreover, the opinion expressed in this paper is that the legislators of the world’s nations should prepare for their arrival. As with many other powerful technologies, there is scope for great harm as well as for great good.

- 
- [1] G. Schreiber, H. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. V. der Velde, and B. Wielinga, *Knowledge Engineering and Management — The CommonKADS Methodology* (Bradford, MIT Press, 2000).
- [2] R. Penrose, *Shadows of the Mind* (Vintage Science, 1994).
- [3] G. Ponti, P. Peretto, and L. Bonfanti, PLoS ONE 3 (2008).
- [4] S. Nasar, *A Beautiful Mind* (Faber and Faber, 1998).
- [5] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007).
- [6] M. Schmidt and H. Lipson, Science 324, 81–85 (2009).
- [7] C. J. Hillar and F. T. Sommer, *On the Article “Distilling Free-form Natural Laws from Experimental Data”* (2009), URL [www.msri.org/people/members/chillar/files/hs09b.pdf](http://www.msri.org/people/members/chillar/files/hs09b.pdf).
- [8] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory* (Cambridge, 2009).
- [9] R. Murphy and D. D. Woods, Intelligent Systems 24 (2009).
- [10] The author is not presently a paid programmer of anything approaching so-called ‘artificial intelligences’.
- [11] Diagnosed as paranoid schizophrenia.
- [12] A transducer is a device that transforms one type of energy into another. Microphones and electric motors are manufactured examples. An eye or an ear, or indeed any sensory neuron in some creature, can also be considered to be a transducer.
- [13] The author stresses that these examples are simply interesting ones that have crossed his path during other research. As such, they represent bright comets that herald the presence of a much larger galaxy of quality research.
- [14] In probability theory and statistics, the expectation is a technical term for the mean of some set of values. However, the term ‘expectation’ also carries with it the connotation of ‘what is likely to happen next, given past experience’ for some stochastic process.
- [15] The early works of the philosopher Jacques Derrida argued for this need, with arguments that rested heavily upon the phenomenology of conscious experience. For an effective ARE there is the further requirement for an actual method of construction to link our scientific knowledge of the statistical properties of perceptual mechanisms to an account of reason.
- [16] Accounts of meaning based on semiotic triads seem too impoverished to this author.
- [17] This is a well-known property of nonlinear dynamical systems.
- [18] See [http://en.wikipedia.org/wiki/Three\\_protect\\_Laws\\_protect\\_of\\_protect\\_Robotics](http://en.wikipedia.org/wiki/Three_protect_Laws_protect_of_protect_Robotics) wherein there is also an extended critique of the laws.