

# Limitations on artificial intelligence\*

Graeme Taylor

February 29, 2004

This writeup comprises in part a response I had been formulating to Gartogg's writeup in *The Failure of Artificial Intelligence*. Whilst I would agree with his assessment that we need to be able to recognise what constitutes intelligent behaviour in order to implement it within a machine, I take issue with the stance that this need be a precise algorithmic description of how to carry out that behaviour if we are to implement it. Rather, I hope to show that an appreciation of what constitutes intelligent output can suffice for its recreation within a machine: that we can model processes that we do not understand so long as we know what we expect them to do. Ultimately it is the inability to even identify such a fitness for purpose criteria that holds AI back- but the gains that have been made through more flexible approaches to computing mean that AI has not really failed. Given this more optimistic stance, this node seems a more appropriate home for my response and as Everything is not a BBS I place it here rather than append criticism to the end of the writeup I intend primarily to discuss.

To start, I would like to draw a distinction between AI the field and AI the product; that is to point out that AI research is not solely or even primarily focussed on the creation of human-like intelligences in a computer-based medium. Such a development of strong AI would indeed be a triumph for the AI field, but there are reasons why it might not be a desirable or feasible direction to work in which I shall attempt to outline later. When AI is popularly discussed, it tends to be in relation to this sci-fi vision of anthropomorphic intelligent machines- and I readily accept that such a level of achievement hasn't occurred and this could be considered a failure. What I do not see it as is as a failure of the field as a whole. Thus whilst arguing against Gartogg's critique of Artificial Intelligence, I do so in the context of AI research as a whole. This may be seen as dodging the 'real' issue (strong AI), but I would see this as changing of definitions to suit a conclusion- defining intelligence as behaviour we have been unable to replicate in machines will obviously lead to a conclusion that artificial intelligence has failed. In fact, this is a common problem for AI projects- once a behaviour has been successfully recreated in a machine, the perception of its value as a measure of intelligence is diminished. Gartogg observes that

---

\*First appeared on Everything2, at [http://www.everything2.com/index.pl?node\\_id=1522987](http://www.everything2.com/index.pl?node_id=1522987)

“many previously “intelligent” actions are now routinely performed mechanically by computers: pattern recognition, mathematical proofs, even playing chess.”

and later argues that

“the real failure of Artificial Intelligence is twofold; it is merely an application of previously understood ideas in general algorithmic computer science, and has done nothing truly new. Secondly, in a very real sense, it is completely goal-less, and therefore unable to succeed at defining the phantoms it chases.”

These to me seem at odds. If an action that is described as intelligent in a human is performed in a computer, why should it cease to be an intelligent action? Philosophical arguments can be made regarding the difference between performing an action and having a consciousness of it, or even an intrinsic comprehension; see for example Searle’s Chinese room. However, to try and avoid these concerns, I will use the following definition of intelligence (from “*Artificial Intelligence*” by Blay Whitby [1]):

“Artificial Intelligence (AI) is the study of intelligent behaviour (in humans, animals and machines) and the attempt to find ways in which such behaviour could be engineered in any type of artifact”

Obviously not all AI researchers would agree upon such a definition (Blay is a lecturer in Cognitive Science and AI; a robotics lab would probably have a very different vision) but it gives a suitably broad basis to work from. Certainly it is unfair to brand AI goal-less any more than, say, mathematics: it is impossible to obtain all mathematics knowledge (in fact, there are things that can be proven to be undecidable in given mathematical systems) yet within many subfields advances can be made against specific questions either for practical purpose or simply for intellectual satisfaction. It would seem strange to suggest that this inability to state or achieve a true goal for mathematics renders the subject a failure; yet this is essentially the complaint being made about AI. Personally, I would argue that AI as a field has the opposite problem- there is a candidate for an ultimate goal, the ‘phantom’ of strong AI; yet this goal should be recognised as being potentially as unobtainable as a complete grasp of mathematics and instead efforts tend these days to be concentrated on smaller (but no less worthwhile) projects.

In the paper *Intelligence Without Representation* [2], Rodney Brooks (Director of the MIT AI lab) gives an illustration of how trying to emulate human levels of intelligence at this early stage may be foolhardy. He suggests considering a group of scientists from 1890 who are trying to create artificial flight. If they are granted (by way of a time machine) the opportunity to take a flight on a commercial 747 then they will be inspired by the discovery that flight is indeed possible- but the lessons they learn from within the passenger cabin will teach them more about seats and cupholders than the underlying aerodynamics. Indeed, seeing that something as massive as a 747 can get off

the ground could have a seriously negative effect on designs they then formulate back in their own time, oblivious to advances such as aluminium or plastics and instead assuming that any weight can be lofted into the air. Even if they got a good look under the hood, a turbofan engine would be essentially incomprehensible. So it is with the human mind- whilst an inspiration that intelligence is indeed obtainable, direct emulation of so advanced a system would be counterproductive, a case of trying to fly before we can walk.

The second 'failure' then, I would discount- but what of the first criticism: that AI has done nothing truly new beyond the application of existing algorithmic computer science? Hopefully it should be clear from Blay's formulation of what constitutes AI that if algorithmic computer science yields intelligent behaviour in a computer, then it is AI, not proof of it's failure. So even if the criticism of unoriginality held, it wouldn't imply a failure of the field. Despite that, I believe it to be untrue in general- the development of intelligent behaviour through at least two methods - neural nets and genetic algorithms- is at odds to the algorithmic approach and has generated results not just in the commercial software arena but in other sections of science.

In general, to solve a problem with an algorithm requires an encapsulation of the problem at an atomic level: we can solve problem X by working through steps Y. However, there are many problems that we haven't devised algorithms for, or which do not lend themselves to such a formulation- how can we codify intuition? What exactly are the defining features of an a compared to an o when carrying out handwriting recognition? (in my case, they're virtually interchangeable!) Despite this, we know the answer when we see it: if a machine can consistently make the same diagnosis as a doctor, even if that doctor has no idea how they ultimately made it, then we can just as much faith in its diagnostic skills. We might like to know how it does it, and perhaps even more so how the doctor does it, but if the behaviour is appropriate then we have a successful implementation of AI. The handwriting example is even simpler- if the machine can output the correct ASCII text for a given scribble, it's achieved the goal of recognition.

How to go about creating such a 'black box'? In conventional CS, you'd need to devise an algorithm, but once created this need merely be implemented in the language/device of your choice. Understanding of the algorithm is understanding of the problem. However, with neural nets, the implementation is the solution. We can expose a net to a given input, and through feedback processes adjust its response until the desired output matches the input. Then we can hope for graceful degradation- namely that the net gives a reasonable output when presented with non-typical output (rather than simply returning an error as an algorithm may do so). This output may not be right- after all, a human could misread my writing and we wouldn't doubt their intelligence as a result- but it should be close (reading a as o is ok, reading it as z seems dubious). Further tuning of the system can hence be used to refine the quality of its output, even without an understanding of just how it gets there- as with the human brain, removing particular nodes doesn't correspond to identifiable failures (e.g a constant inability to read an i) but rather a degradation

of performance (more errors per string). Contrast this with the effect of pulling a line of code out of an algorithm, which is likely to be disastrous.

Continuing this theme are genetic algorithms. Here we are again concerned with results, not methods. Given a fitness criterion, solutions can be ranked in their ability to solve a problem. Appropriate mingling of the most successful solutions should yield a new set of solutions, some better than any we previously had, others worse. Repeated iteration gives us a suitable solution without ever figuring out what makes the problem tick- this is akin to gaining rules of thumb through practical experience rather than formal study. Often genetic algorithms can find solutions entirely different to the type that a conventional algorithmic approach offers- which is unsurprising, as mathematically rigorous solutions that appeal to a mathematically-minded designer aren't necessarily the only intelligent solution. Quite often, they find ways to cheat that could be considered ingenuity, or more likely abuse of factors we haven't taken into account which cannot always be depended on. For example, given a set of photos of men and women, we might seek a system that can tell one gender from the other. If however all the pictures of the men were taken in a different room to those of the women, a most efficient solution would be to recognise the decor rather than facial features. As soon as you supply pictures from the outside world, you'd run into problems. But the system hasn't necessarily failed in the task of distinguishing the pictures, it just uses different reasons to those desired.

But if we have these methods for getting intelligent behaviour, why can't we just keep adding to the net or running the algorithm until we evolve a strong AI system? After all, we believe that evolution has produced at least one intelligent 'machine'- us. However, it seems that the biggest problem is just that: scalability. We have seen that AI methods have allowed us to tackle problems which, despite a lack of an algorithmic understanding, we can declare to be solved either successfully or not. It is this fitness of purpose criterion which is both AI's strength and weakness- we can use it to generate more forms of intelligent behaviour than algorithmic computer science alone does; but it does not offer us the complete set. The problem is that there is no single fitness criterion for intelligence- when presented with a passage from Shakespeare, what should our system do- count the words, compare the spelling to today's, examine the meter, or write an essay about its relevance to modern life? All are varying forms of intelligence (and all might get thrown at you during an english lesson) yet quantifying how intelligent such an action is, and hence refining different solutions against that benchmark, seems impossible. We might be able to implement all of them given time- but getting them to play together nicely in a system will probably need another level of insight akin to the leap from algorithms to less predictable but more flexible methods such as neural nets. Ultimately, if we want to recreate human intelligence, we may well need to understand ourselves first. This too is a goal of AI research, and for many is the end goal- not to recreate ourselves, but to know just what is that makes our intelligence special in the first place.

In conclusion then, I wouldn't argue that AI has failed either to advance our ability to produce intelligent behaviour in devices other than ourselves, nor to build upon the foundations of standard computer science. There have been many remarkable solutions to problems and the methods used to solve them are of interest in themselves. Often they have turned out to give greater insight into other fields (such as the use of neural nets for modelling the activity of human brains) or to highlight questions that we need to ask about ourselves, tying together ideas about science, mathematics and philosophy. These solutions have thus far been limited both in their scalability and their interaction with each other- which isn't helped by deep divisions within the AI community as to which methods are best; divisions which are ultimately pointless if it turns out that all these ideas need to be applied together to create a superior whole- but this should not diminish the results that have already been seen.

## References

- [1] "*Oneworld Beginner's Guides- Artificial Intelligence*": Blay Whitby, ISBN 1-85168-322-4
- [2] "*Intelligence without representation*": Rodney A. Brooks, referenced in and also available at <http://www.ai.mit.edu/people/brooks/papers/representation.pdf>