

# AGI needs the Humanities

Sam Freed

Dept. of Informatics, University of Sussex

Abstract

Central scholars in AI have argued for extending the search for new AI technology beyond the tried-and-tested biologically and mathematically-inspired algorithms. Following in their footsteps, areas in the humanities are introduced as possible inspirations for novel human-like AI. Topics discussed include play-acting, literature as the field researching both imagination and metaphors, linguistics, music, and hermeneutics. In our ambition to reach *general* intelligence, we cannot afford to ignore these avenues of research.

## 1 Introduction

AI as commonly practised generally no longer even aspires to Human level AI. The people who keep this dream from before 1956 alive have largely been confined to conferences about AGI – somehow the general AI has become a subfield. This has to do with how successful specific techniques in machine learning have become, and how embarrassingly stuck general AI seems: The opinion that AI has been at some level “*brain dead*” since at least the 1970s is voiced by pillars of the AI community such as Marvin Minsky (McHugh & Minsky, 2003), Geoffrey Hinton (LeVine & Hinton, 2017), and Rodney Brooks:

*... modern-day [AI] research is not doing well at all on either being general or supporting an independent entity with an ongoing existence. It mostly seems stuck on the same issues in reasoning and common sense that AI has had problems with for at least 50 years...* (Brooks, 2017)

AI so far has been heavily influenced by the rationalist tradition. It is characterised by approaching any and all problems in a series of steps:

1. *Characterise the situation in terms of identifiable objects with well-defined properties.*
2. *Find general rules that apply to situations in terms of those objects and properties.*
3. *Apply the rules logically to the situation of concern, drawing conclusions about what should be done.* (Winograd & Flores, 1986, pp. 14–26)

Note how Brooks complains about AI being incapable of “*supporting an independent entity with an ongoing existence*”. On the one side this has to do with mathematics’ infatuation with functions, that by their very definition return the same value for the same parameters regardless of the time of evaluation; On the other hand it has to do with science and technology’s aversion to all things subjective and human-like. This paper will march straight into this terrain – asking where in the Humanities would we find the best input for our effort to develop AGI.

Several arguments have been advanced as to where AI should go to find ideas for novel algorithms. Langley argues that AI should go back to its roots in the cognitive sciences (2006). That is hardly controversial, since cognitive science and AI evolved together since the 1950s. Some argue for extending our horizons: Boden, acknowledging that AI is an integral part of the cognitive sciences, laments the absence of any research in anthropology informing either cognitive science or AI

(Boden, 2008). Boden's promotion of anthropology can be seen as a first tentative step towards a more radical position, articulated by CP Snow (see below).

The most vociferous critic of AI from the humanities has been Hubert Dreyfus (Dreyfus, 1979, 2007). He argued for AI researchers to understand humans better (mainly by reading Heidegger and Merleau-Ponty). Mainstream AI research mostly either ignored him or trivialised his critiques. This work stands with mainstream AI in demanding programmable results (see (Freed, 2019)), and stands with Dreyfus in pointing out the shortcomings of AI research. This call for a more human-aware AI may sound radical methodologically, but is quite trivial personally and subjectively. Methodologically, science likes objectivity and abhors subjectivity. But in programming a mind like our own, can we afford to ban our own personal view of our own mind? Personally, there is nothing difficult in noticing our human, subjective side.

Especially in AGI, we need to be more daring than people who are pursuing merely the next incremental step in AI.

## 2 Approach

CP Snow pointed out, with some alarm, that a chasm had opened between two distinct intellectual cultures – What we would now call STEM (Science, Technology, Engineering, Mathematics) and the Humanities or the Arts. He lamented that even basic communications across this divide have become difficult. He argued that such a chasm would necessarily be detrimental to the development of society, and would specifically hinder the UK's ability to compete with the USA and Russia in those days of the cold war (Snow, 1964).

But criticism of AI's limited view of the mind was not only external, but came also from the very centre, from MIT's AI labs:

*We are to thinking as Victorians were to sex. We all know we have these horrible moments of confusion when we begin a new project, that nothing looks clear and everything looks awful, that we work our way out using all sorts of odd little rules of thumb, by going down blind alleys and coming back again, and so on, but since everyone else seems to be thinking logically, or at least they claim they do, then we figure we must be the only ones in the world with such murky thought processes. We disclaim them, and make believe that we think in logical, orderly ways, all the time knowing very well that we don't. And the worst offenders here are teachers, who present crisp, clean batches of knowledge to their students, and look as if they themselves had learned that knowledge in a crisp, clean way. It didn't happen that way, but the teachers don't admit it, and the students groan inwardly, feeling so hopelessly dumb. (McCorduck, 2004, p. 339)*

The author has argued elsewhere for the rehabilitation of *introspection* as a source of ideas in AI, after it been frowned upon since the behaviourist revolution in psychology (Freed, 2017, 2019). Here we will examine other areas that were historically neglected, that seem to have salience for the insights required for AGI.

## 3 Play-Acting

As argued elsewhere, One can see the process of programming as consisting of:

1. Understanding the requirement (say adding up items in an invoice and adding some sales tax to form a total);
2. Projecting ones mind into an imagined world where the environment, instead of consisting in chairs and desks, consists of (say) the Python interpreter (and associated libraries);

3. Imagining how one one could solve the problem if one were acting using the tools available in the Python environment (loops, variables, input/output functions); and
4. Logging these actions (or the equivalent “instructions”) in a text file, henceforth called the “program” (Freed, 2018).

So it would seem that the role of a programmer is a *role*, taken on willingly by the skilled programmer, a bit like a character-role taken on by a theatrical performer. Note that this observation is not alien to our field, in that Herbert Simon wrote (in his writing on administrative behaviour):

*Administration is not unlike play-acting. The task of the good actor is to know and play his role... The effectiveness of the performance will depend on the effectiveness of the play and the effectiveness in which it is played. The effectiveness of the administrative process will vary with the effectiveness of the organisation and the effectiveness with which its members play their parts.* (Simon, 1976, p. 252, 1996, p. xii)

If acting is central to much of our behaviour, or at least to our effective behaviour (known as work) then the study of theatre looks promising for advancing any effective behaviour also in machines – at least machines that we hope to endow with decision-making abilities.

## 4 Imagination, action, and the limits thereof

When we do some thing X, or recall doing the same X, or imagine doing the same X – our brain functions in a very similar manner (Hesslow, 2012). This fact alone should spark a degree of interest in imagination research for AGI. The AI community indeed has given imagination some attention, see (Mahadevan, 2018).

Imagination is of interest in at least two ways. It seems to be a locus of much human creativity, and creativity seems to be a “holy grail” yet to be achieved in AI or explained by cognitive science (Boden, 2010). Most research (in the context of AI) has been into imagination in the sense of some sort of a “Cartesian space” - like a canvas inside our mind, where we form and develop ideas, a bit like white-board.

Here is a different and perhaps more interesting angle of research into imagination: What can be imagined seems to be a limitation of what humans can do and think. In other words, the space of human endeavour is restricted to what is imaginable. The study of what is imaginable, of what is humanly credible – goes on in the fields of literature, theatre & cinema.

Note that beyond statements of fact being true or false in the real world, there can be imaginary worlds where statements can be equally true or false: Mary had a little lamb, not a pangolin, and Snow White had 7 dwarves – no more and no less.

## 5 Linguistics and Music

Linguistics have been central to the cognitive sciences. Many would even date the beginning of the cognitive revolution a paper by Chomsky (1959) – which argues that human capabilities in syntax cannot be explained by behaviourism. However, there is a further point that may be of interest – when we hear an idea, we often ask ourselves whether it “sounds right” - in more senses than one.

- Are the sentences grammatical?
- Do the ideas “make sense”? Do they fit in some established and accepted pattern like a syllogism?

But note that the question of “sounding right” insinuates also some musical quality, some balance or form that is aesthetically correct. Again, the other side of Snow’s divide beckons us (Miranda, 2013).

## 6 Metaphor

Often we hear naive people say things such as that “the computers knows” some fact or skill. The better informed would comment that computers do not “know” anything, and have no mental states – they are hulks of metal silicon and plastic that process electrical signals in a sophisticated way that we call “information processing” (Smith, 2005). The idea that the bank’s computer “knows” my address arises out of the fact that in the correct configuration, when queried with a string of characters that represents (by social convention) my name or account number, the system is capable of emitting a string of characters that would represent (again by social convention) my address. But there is no *knowing* there at all. We, humans, know how to operate the computer system in order to obtain what *for us* is useful *information*. For the computer, it is all electrons going hither and thither. Saying that the computer “knows” anything is metaphorical. And where does this metaphor reside? In the minds of the humans designing and using the system. The computer has no capability for any mental state – not for knowing, and definitely not for metaphorical thinking.

However, we can still learn something profound from this metaphorical ascription of knowledge to the electronic device we call “a computer”. What we see here clearly, is that *humans* think metaphorically. We as *humans* have this capacity to see “knowledge” where there is none, and to see “information” when all that physically exists are lit dots on a screen.

Further evidence of how metaphorical our thinking is was provided by Bolter (1984). He surveys how our culture described the mind in different eras, and argues that it was always through the metaphor of the latest technology: In ancient (Greek) times, the human was considered as “a clay vessel with a divine spark”. With the introduction of clock towers in late medieval times, the human and his mind were considered in terms of mechanical automata – to this day we use expressions like “cogs turning in our head”<sup>1</sup>. In the late 19<sup>th</sup> century, with the arrival of pneumatic and hydraulic technologies, the metaphor used (for example) by Freud was of pressures, repressions, and eruptions of emotions. Today we think of the mind as a computer, as in the title of Boden’s history of Cognitive science - “mind as machine” - there is little doubt which machine the mind is being likened to (Boden, 2008).

So, it would seem, that if we want to program human-level, general AI – we need to develop systems that can do at least *some* metaphorical thinking. This is a tall order – and some research is already underway into metaphor as analogy (e.g. Barnden, 2008). However, metaphorical thinking is far more complex than mere analogy. The topic of metaphor is already studied in detail, but in departments of literature, Not computer science of cognition. Perhaps we should start with reading about poetry – where scholars have studied metaphors for generations.

## 7 Hermeneutics<sup>2</sup>

**Hermeneutics** (the theory of interpretation) was founded as the theory of how to correctly understand ancient religious texts. Arguably hermeneutics is at least as old as the Pauline epistles in the new testament, however it is with **Martin Luther's** (1783-1546) protestant injunction, that the bible should be interpreted only on its own terms (without any reference to Catholic tradition) that

---

1 <https://www.youtube.com/watch?v=WEhS9Y9HYjU>

2 Much of his section is based on previously published work (Freed, 2017, 2019)

we see the first explicit statement of a *policy* or *principle* by which interpretation of a text should be carried out (Ramberg & Gjesdal, 2014).

Speaking against Cartesian notions of understanding (“clear and distinct”), **Giambattista Vico** (1668-1744) “argues that thinking is always rooted in a given cultural context. This context is historically developed, and, moreover, intrinsically related to ordinary language” (*Ibid.*).

Later **Friedrich Schleiermacher** (1768-1834) discussed the alien nature of old or foreign texts, and called for particular attention to our prejudices, so we can understand texts under their own alien context. He did not guarantee that such strict awareness of prejudice and openness will lead to a correct understanding of a text (that may be impossible). However such openness is *necessary* for understanding, and is required not only for foreign texts but for any type of communications (*Ibid.*).

**Wilhelm Dilthey** (1833-1911) distinguished “*living experience*” which is how each of us experience ourselves, from “*understanding*” which is how we more systematically understand the world outside us and others. He claimed that true self-awareness can only be achieved when one understands oneself on the same terms one understands others. In understanding history and historical texts one should combine (what we would now call) empathy, i.e. a “*living experience*” identification with the historical characters, with “*understanding*”, which is a more rigorous “from the outside” observation. The “*living experience*” component allows the historian to form hypotheses about history, while the “*understanding*” part allows one to critique such thoughts, and see how well they stand to reason (*Ibid.*).

For modern thinkers such as Heidegger (and Dreyfus, the premier philosophical critic of AI (Dreyfus, 1979)) interpretation is not only a matter of understanding texts, but of our entire mode of being, which is continuously involved with comprehending the world and acting in it – hence hermeneutics becomes one and the same project as phenomenology. Heidegger was concerned with many issues in phenomenology, and viewed the specifics of hermeneutics *as such* as a sub-field, the detailed exploration of which he later entrusted to a large degree to Gadamer (Malpas, 2013, Chapter 4).

**Hans-Georg Gadamer** (1900-2002) viewed hermeneutics not only as the theory of understanding ancient texts and art in general but also, and perhaps mainly, as the act of continuously understanding/interpreting all situations. In this sense, interpretation is an unceasing human activity, during at least most waking hours (Gadamer, 2004, pt. 1).

Here is an example (my own) of what is meant by interpretation in this context. Consider the following:

- הכלב מכוער
- Ha-kelev meh'oar
- Il cane é brutto
- The canine is brutish
- The dog is ugly

At this point you may be perplexed by this strange list, as one would be with any other strange sequence that is presented with little warning. In a sense I just caused you to be “thrown” onto this unusual list, and to the urgency of making sense of the situation. The lines above all convey the same meaning (in different alphabets, languages and dialects). Note how much easier it is to interpret (for an English monoglot) these examples the further down one goes. Note also that as an English-speaker you may be further interpreting the situation and objecting that “brutish” does not mean the same as “ugly”, but you also may be aware that in the Italian “brutto” does actually mean

ugly, and may further be aware of how such words change meanings over the centuries and the geographic distances involved. All these thoughts are interpretative – they are attempts to make sense of a situation, at this instance the situation at hand is the bizarre list above. *This* sort of interpretative effort is what is the mental activity that hermeneutics study, and I argue is necessary for AGI.

Interpretation (in the sense that interests us here) is the ability to “follow along”, to “make sense” of the “inputs”. In following along with (say) a song, this is easier with a familiar tune than it is with foreign music. The crux (here) of the knowledge or skill accumulated as we become more familiar with a situation does *not* consist of beliefs - we have no position on the ugliness or beauty of a dog we have never seen. What is being formed is an *interpretation*, an understanding, a grasp – before (and not requiring) any judgement.

Gadamer being a student of Heidegger’s, following Gadamer to explore AGI is in line with Dreyfus’s (2007) call for a more Heideggerian AI. Gadamer was first mentioned as a possible source for AI research by Winograd and Flores (1986), and an algorithm that aspired to follow this path is proposed in (Freed, 2017, 2019).

## 8 Final Notes

As we have seen, beyond the great divide between the STEM subjects and the humanities several promising fields offer tantalising prospects for the adventurous AI researcher. In bringing this survey to a close, it is worth noting that some 20<sup>th</sup> century thinkers that would be considered more conventional would agree with the directions outlined above.

Wittgenstein described our perception as “seeing as” - we see the duck-rabbit picture either as a rabbit or as a duck (Wittgenstein, 2001). This process is interpretative – as was outlined above.

Developmental psychologies such as Piaget (1989) offer schemas of how cognition develops in children. Regardless of the veracity of any one such theory, any theory that seems programmable may be used as a model for an AI system (Freed, 2019; Matthews & Mullin, 2018).

## 9 Bibliography

- Barnden, J. A. (2008). Metaphor and artificial intelligence: Why they matter to each other. *The Cambridge Handbook of Metaphor and Thought*, 311–338.
- Boden, M. A. (2008). *Mind as Machine: A History of Cognitive Science*. OUP Oxford.
- Boden, M. A. (2010). *Creativity and Art: Three Roads to Surprise*. Oxford University Press.
- Bolter, J. D. (1984). *Turing’s man: Western culture in the computer age*. Duckworth.
- Brooks, R. A. (2017, October 6). *Robotics pioneer Rodney Brooks debunks AI hype seven ways*. MIT Technology Review. <https://www.technologyreview.com/s/609048/the-seven-deadly-sins-of-ai-predictions/>
- Chomsky, N. (1959). A review of BF Skinner’s Verbal Behavior. *Language*, 35(1), 26–58.
- Dreyfus, H. L. (1979). *What computers can’t do / The limits of artificial intelligence* (Revised). Harper & Row.
- Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, 171(18), 1137–1160. <https://doi.org/10.1016/j.artint.2007.10.012>
- Freed, S. (2017). *A role for introspection in AI research* [University of Sussex]. <http://sro.sussex.ac.uk/66141/>

- Freed, S. (2019). *AI and Human Thought and Emotion*. CRC Press.
- Freed, S. (2018). Is Programming Done by Projection and Introspection? In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2017* (pp. 187–189). Springer International Publishing.
- Gadamer, H.-G. (2004). *Truth and method* (2nd, rev. ed ed.). Continuum.
- Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Research*, 1428, 71–79. <https://doi.org/10.1016/j.brainres.2011.06.026>
- Langley, P. (2006). *Intelligent behavior in humans and machines*. Technical Report). Computational Learning Laboratory, CSLI, Stanford University. <http://lyonesse.stanford.edu/~langley/papers/ai50.dart.pdf>
- LeVine, S., & Hinton, G. (2017, September 15). *Artificial intelligence pioneer says we need to start over*. Axios. <https://www.axios.com/ai-pioneer-advocates-starting-over-2485537027.html>
- Mahadevan, S. (2018, April 29). Imagination Machines: A New Challenge for Artificial Intelligence. *Thirty-Second AAI Conference on Artificial Intelligence*. Thirty-Second AAI Conference on Artificial Intelligence. <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16147>
- Malpas, J. (2013). Hans-Georg Gadamer. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013). <http://plato.stanford.edu/archives/sum2013/entries/gadamer/>
- Matthews, G., & Mullin, A. (2018). The Philosophy of Childhood. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2018/entries/childhood/>
- McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence* (25th anniversary update). A.K. Peters.
- McHugh, J., & Minsky, M. (2003, August 1). *Why A.I. Is Brain-Dead*. WIRED. <http://www.wired.com/2003/08/why-a-i-is-brain-dead/>
- Miranda, E. R. (2013). *Readings in Music and Artificial Intelligence*. Routledge.
- Piaget, J. (1989). *The Child's Conception of the World*. Rowman & Littlefield.
- Ramberg, B., & Gjesdal, K. (2014). Hermeneutics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2014). <http://plato.stanford.edu/archives/win2014/entries/hermeneutics/>
- Simon, H. A. (1976). *Administrative behavior: A study of decision-making processes in administrative organization* (3rd ed.). Collier Macmillan. <http://capitadiscovery.co.uk/sussex-ac/items/38710>
- Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). MIT Press. <http://capitadiscovery.co.uk/sussex-ac/items/546838>
- Smith, B. C. (2005, January 31). *Digital Future: Meaning of Digital*. C-SPAN Video Library. <http://c-spanvideo.org/program/FutureM>
- Snow, C. P. (1964). *The two Cultures: And a Second look* (2 ed.). C.U.P. <http://capitadiscovery.co.uk/sussex-ac/items/22469>
- Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Ablex. <http://prism.talis.com/sussex-ac/items/272586>
- Wittgenstein, L. (2001). *Philosophical Investigations: The German Text with a Revised English Translation* (3rd Edition). Wiley-Blackwell.