(IJRST) 2011, Vol. No. 1, Issue No. III, July- ISSN: 2249-0604

ARTIFICIAL INTELLIGENCE: A MODERN APPROACH

Neha Rastogi, Rachna Kumari, Summi Nigam

MCA Students, ABES Engineering College, Ghaziabad

ABSTRACT

Artificial Intelligence comes about through a similar accretion of working algorithms, with the researchers having no deep understanding of how the combined system works. General intelligence is a between-species difference, a complex adaptation, and a human universal found in all known cultures. Artificial Intelligence is symmetrical around potential good impacts and potential bad impacts. That is why the title of this chapter is "Artificial Intelligence as a Positive and Negative Factor in Global Risk", not "Global Risks of Artificial Intelligence."

INTRODUCTION

By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it. Of course this problem is not limited to the field of AI. (Monod 1974.) The field of AI has a reputation for making huge promises and then failing to deliver on them. Most observers conclude that AI is hard; as indeed it is. But the *embarrassment* does not stem from the difficulty. It is difficult to build a star from hydrogen, but the field of stellar astronomy does not have a terrible reputation for promising to build stars and then failing. The critical inference is *not* that AI is hard, but that, for some reason, it is very easy for people to think they know far more about Artificial Intelligence than they actually do.

It is far more difficult to write about global risks of Artificial Intelligence than about cognitive biases.

ANTHROPOMORPHIC BIAS

When something is universal enough in our everyday lives, we take it for granted to the point of forgetting it exists. Imagine a complex biological adaptation with ten necessary parts. If each of ten genes is independently at 50% frequency in the gene pool - each gene possessed by only half the organisms in that species - then, on average, only 1 in 1024 organisms will possess the full, functioning adaptation. A fur coat is not a significant evolutionary advantage unless the environment reliably challenges organisms with cold. Similarly, if gene B depends

on gene A, then gene B has no significant advantage unless gene A forms a reliable part of the *genetic* environment.

Querying your own human brain works fine, as an adaptive instinct, if you need to predict other humans. If you deal with any other kind of optimization process - if, for example, you are the eighteenth-century theologian William Paley, looking at the complex order of life and wondering how it came to be - then anthropomorphism is flypaper for unwary scientists, a trap so sticky that it takes a Darwin to escape.

PREDICTION AND DESIGN

We cannot query our own brains for answers about nonhuman optimization processes - whether bug-eyed monsters, natural selection, or Artificial Intelligences.

AI, which stems not from the difficulty of AI as such, but from the mysterious ease of acquiring erroneous beliefs about what a given AI design accomplishes. Some early AI researchers believed that an artificial neural network of layered thresholding units, trained via backpropagation, would be "intelligent". The wishful thinking involved was probably more analogous to alchemy than civil engineering. Magic is on Donald Brown's list of human universals (Brown 1991); science is not. We don't *instinctively* see that alchemy won't work. We don't *instinctively* distinguish between rigorous understanding and good storytelling. We don't *instinctively* notice an expectation of positive results which rests on air.

The human species came into existence through natural selection, which operates through the nonchance retention of chance mutations. One path leading to global catastrophe – to someone pressing the button with a mistaken idea of what the button does - is that Artificial Intelligence comes about through a similar accretion of working algorithms, with the researchers having no deep understanding of how the combined system works.

Nonetheless they believe the AI will be friendly, with no strong visualization of the exact processes involved in producing friendly behavior, or any detailed understanding of what they mean by friendliness. Much as early AI researchers had strong mistaken vague expectations for their programs' intelligence, we imagine that these AI researchers succeed in constructing an intelligent program, but have strong mistaken vague expectations for their program's friendliness.

UNDERESTIMATING THE POWER OF INTELLIGENCE

We tend to see individual differences instead of human universals. Thus when someone says the word "intelligence", we think of Einstein, instead of humans.

General intelligence is a between-species difference, a complex adaptation, and a human universal found in all known cultures. There may as yet be no academic consensus on intelligence, but there is no doubt about the existence, or the power, of the thing-to-be-explained. There is *something* about humans that let us set our footprints on the Moon.

Intelligence is the foundation of human power, the strength that fuels our other arts.

The danger of confusing general intelligence with g-factor is that it leads to tremendously underestimating the potential impact of Artificial Intelligence. Even the phrase "transhuman AI" or "artificial superintelligence" may still evoke images of book-smarts-ina- box: an AI that's *really good* at cognitive tasks stereotypically associated with "intelligence", like chess or abstract mathematics. But not superhumanly persuasive; or far better than humans at predicting and manipulating human social situations; or inhumanly clever in formulating long-term strategies.

The catastrophic scenario which stems from underestimating the power of intelligence is that someone builds a button, and doesn't care enough what the button does, because they don't think the button is powerful enough to hurt them. Or, since underestimating the power of intelligence implies a proportional underestimate of the potential impact of Artificial Intelligence, the (presently tiny) group of concerned researchers and grantmakers and individual philanthropists who handle existential risks on behalf of the human species, will not pay enough attention to Artificial Intelligence. Or the wider field of AI will not pay enough attention to risks of strong AI, and therefore good tools and firm foundations for friendliness will not be available when it becomes possible to build strong intelligences.

And one should not fail to mention - for it also impacts upon existential risk - that Artificial Intelligence could be the powerful solution to other existential risks, and by mistake we will ignore our best hope of survival. The point about underestimating the potential impact of Artificial Intelligence is symmetrical around potential good impacts and potential bad impacts. That is why the title of this chapter is "Artificial Intelligence as a Positive and Negative Factor in Global Risk", not "Global Risks of Artificial Intelligence." The prospect of AI interacts with global risk in more complex ways than that; if AI were a pure liability, matters would be simple.

CAPABILITY AND MOTIVE

There is a fallacy oft-committed in discussion of Artificial Intelligence, especially AI of superhuman capability. Someone says: "When technology advances far enough we'll be able to build minds far surpassing human intelligence. Now, it's obvious that how large a cheesecake you can make depends on your intelligence. A superintelligence could build *enormous* cheesecakes - cheesecakes the size of cities - by golly, the future will be full of giant cheesecakes!" The question is whether the superintelligence *wants* to build giant cheesecakes. The vision leaps directly from *capability* to *actuality*, without considering the necessary intermediate of *motive*.

FRIENDLY AI

It would be a very good thing if humanity knew how to choose into existence a powerful optimization process with a particular target. Or in more colloquial terms, it would be nice if we knew how to build a nice AI.

Proving a computer chip correct requires a synergy of human intelligence and computer algorithms, as *currently* neither suffices on its own. Perhaps a true AI could use a similar *combination of abilities* when modifying its own code - would have *both* the capability to *invent* large designs without being defeated by exponential explosion, and *also* the ability to *verify* its steps with extreme reliability. That is one way a true AI might remain knowably stable in its goals, even after carrying out a large number of self-modifications. It is disrespectful to human ingenuity to declare a challenge unsolvable without taking a close look and exercising creativity. It is an enormously strong statement to say that you *cannot* do a thing - that you *cannot* build a heavier-than-air flying machine, that you *cannot* get useful energy from nuclear reactions, that you *cannot* fly to the Moon.

Such statements are universal generalizations, quantified over every single approach that anyone ever has or ever will think up for solving the problem. It only takes a single counterexample to falsify a universal quantifier. The statement that Friendly (or friendly) AI is *theoretically impossible*, dares to quantify over *every possible* mind design and *every possible* optimization process - including human beings, who are also minds, some of whom are nice and wish they were nicer. At this point there are any number of vaguely plausible reasons why Friendly AI might be *humanly* impossible, and it is still more likely that the

problem is solvable but no one will get around to solving it in time. But one shouldnot so quickly write off the challenge, especially considering the stakes.

TECHNICAL FAILURE AND PHILOSOPHICAL FAILURE

Bostrom (2001) defines an existential catastrophe as one which permanently extinguishes Earth-originating intelligent life *or destroys a part of its potential.* We can divide potential failures of attempted Friendly AI into two informal fuzzy categories, *technical failure* and *philosophical failure.* Technical failure is when you try to build an AI and it doesn't work the way you think it does - you have failed to understand the true workings of your own code. Philosophical failure is trying to build the wrong thing, so that even if you succeeded you would still fail to help anyone or benefit humanity. Needless to say, the two failures are not mutually exclusive.

The border between these two cases is thin, since most philosophical failures are much easier to explain in the presence of technical knowledge. In theory you ought first to say what you *want*, then figure out *how* to get it. In practice it often takes a deep technical understanding to figure out what you want.

RATES OF INTELLIGENCE INCREASE

From the standpoint of existential risk, one of the most critical points about Artificial Intelligence is that an Artificial Intelligence might increase in intelligence *extremely fast.*

The obvious reason to suspect this possibility is recursive self-improvement. The AI becomes smarter, including becoming smarter at the task of writing the internal cognitive functions of an AI, so the AI can rewrite its existing cognitive functions to work even better, which makes the AI still smarter, including smarter at the task of rewriting itself, so that it makes yet more improvements.

Human beings do not recursively self-improve in a *strong* sense. To a *limited* extent, we improve ourselves: we learn, we practice, we hone our skills and knowledge. To a *limited*extent, these self-improvements improve our ability to improve. New discoveries can increase our ability to make further discoveries - in that sense, knowledge feeds on itself. But there is still an underlying level we haven't yet touched. We haven't rewritten the human brain. The brain is, ultimately, the source of discovery, and our brains today are much the same as they were ten thousand years ago.

In a similar sense, natural selection improves organisms, but the process of natural selection does not itself improve - not in a strong sense. Adaptation can open up the way for additional adaptations. In this sense, adaptation feeds on itself. But even as the gene pool boils, there's still an underlying heater, the process of mutation and recombination and selection, which is not itself re-architected. A few rare innovations increased the rate of evolution itself, such as the invention of sexual recombination. But even sex did not change the essential nature of evolution: its lack of abstract intelligence, its reliance on random mutations, its blindness and incrementalism, its focus on allele frequencies.

An Artificial Intelligence could rewrite its code from scratch - it could change the underlying dynamics of optimization. Such an optimization process would wrap around *much more strongly* than either evolution accumulating adaptations, or humans accumulating knowledge. The key implication for our purposes is that an AI might make a *huge* jump in intelligence after reaching some threshold of criticality.

THREATS AND PROMISES

It is a risky intellectual endeavor to predict *specifically* how a benevolent AI would help humanity, or an unfriendly AI harm it. There is the risk of *conjunction fallacy*: added detail necessarily reduces the joint probability of the entire story, but subjects often assign higher probabilities to stories which include strictly added details. There is the risk - virtually the certainty - of failure of imagination; and the risk of Giant Cheesecake Fallacy that leaps from capability to motive.

AI VERSUS HUMAN INTELLIGENCE ENHANCEMENT

By hypothesis, the computer runs a detailed simulation of a biological human brain, executed in sufficient fidelity to avoid any detectable high-level effects from systematic low-level errors. Any accident of biology that affects information-processing *in any way*, we must faithfully simulate to sufficient precision that the overall flow of processing remains isomorphic. To *simulate* the messy biological computer that is a human brain, we need far more *useful* computing power than is embodied in the messy human brain itself.

The most probable way we would develop the ability to scan a human brain neuron by neuron - in sufficient detail to capture *every* cognitively relevant aspect of neural structure - would be the invention of sophisticated molecular nanotechnology. Molecular nanotechnology could

probably produce a desktop computer with total processing power exceeding the aggregate brainpower of the entire current human population.

Furthermore, if technology permits us to scan a brain in sufficient fidelity to *execute the scan as code*, it follows that for some years previously, the technology has been available to obtain *extremely detailed* pictures of processing in neural circuitry, and presumably researchers have been doing their best to understand it.

Furthermore, to *upgrade* the upload - transform the brain scan so as to increase the intelligence of the mind within - we must necessarily understand the *high-level* functions of the brain, and how they contribute usefully to intelligence, in excellent detail.

Furthermore, humans are not designed to be improved, either by outside neuroscientists, or by recursive self-improvement internally. Natural selection did not build the human brain to be humanly hackable. All complex machinery in the brain has adapted to operate within narrow parameters of brain design. Suppose you can make the human smarter, let alone superintelligent; does the human remain *sane*? The human brain is very easy to perturb;

just changing the balance of neurotransmitters can trigger schizophrenia, or other disorders. Deacon (1997) has an excellent discussion of the evolution of the human brain, how delicately the brain's elements may be balanced, and how this is reflected in modern brain dysfunctions. The human brain is not end-user-modifiable.

INTERACTIONS OF AI WITH OTHER TECHNOLOGIES

Speeding up a desirable technology is a local strategy, while *slowing down* a dangerous technology is a difficult majoritarian strategy. *Halting* or *relinquishing* an undesirable technology tends to require an impossible unanimous strategy. I would suggest that we think, not in terms of developing or not-developing technologies, but in terms of our *pragmatically available latitude* to *accelerate* or *slow down* technologies; and ask, *within the realistic bounds of this latitude*, which technologies we might prefer to see developed *before* or *after* one another.

In nanotechnology, the goal usually presented is to develop defensive shields before offensive technologies. Offense has outweighed defense during most of civilized history. Guns were developed centuries before bulletproof vests. Smallpox was used as a tool of war before the development of smallpox vaccines. Today there is still no shield that can deflect a nuclear explosion; nations are protected not by defenses that cancel offenses, but by a balance

of offensive terror. The nanotechnologists have set themselves an intrinsically difficult problem.

CONCLUSION

For decades the U.S. and the U.S.S.R. avoided nuclear war, but not *perfectly;* there were close calls, such as the Cuban Missile Crisis in 1962. If we postulate that future minds exhibit the same mixture of foolishness and wisdom, the same mixture of heroism and selfishness, as the minds we read about in history books - then the game of existential risk is already over; it was lost from the beginning. We might survive for another decade, even another century, but not another million years.

But the human mind is not the limit of the possible. *Homo sapiens* represents the *first* general intelligence. We were born into the uttermost beginning of things, the dawn of mind. With luck, future historians will look back and describe the present world as an awkward in-between stage of adolescence, when humankind was smart enough to create tremendous problems for itself, but not quite smart enough to solve them.

Artificial Intelligence is one road into that challenge; and I think it is the road we will end up taking. Stars were once mysteries, and chemistry, and biology. Generations of investigators tried and failed to understand those mysteries, and they acquired the reputation of being impossible to mere science. No one knew how living matter reproduced itself, or why our hands obeyed our mental orders.

All scientific ignorance is hallowed by ancientness. Each and every absence of knowledge dates back to the dawn of human curiosity; and the hole lasts through the ages, seemingly eternal, right up until someone fills it. Intelligence must cease to be any kind of mystery whatever, sacred or not. We must execute the creation of Artificial Intelligence as the exact application of an exact art. And maybe then we can win.

REFERENCES

[1] Asimov, I. 1942. Runaround. *Astounding Science Fiction,* March 1942.

[2] Barrett, J. L. and Keil, F. 1996. Conceptualizing a non-natural entity: Anthropomorphism in God concepts. *Cognitive Psychology*, **31**: 219-247.

[3] Bostrom, N. 1998. How long before superintelligence? *Int. Jour. of Future Studies,* **2**.

[4] Brown, D.E. 1991. *Human universals.* New York: McGraw-Hill.

[5] Deacon, T. 1997. *The symbolic species: The co-evolution of language and the brain.* New York: Norton.

[6] Drexler, K. E. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation.* New York: Wiley-Interscience.

[7] Ekman, P. and Keltner, D. 1997. Universal facial expressions of emotion: an old controversy and new findings. In *Nonverbal communication: where nature meets culture,* eds. U. Segerstrale and P. Molnar.

[8] Hayes, J. R. 1981. *The complete problem solver.* Philadelphia: Franklin Institute Press.

[9] Hibbard, B. 2004. Reinforcement learning as a Context for Integrating AI Research. Presented at the *2004 AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research.*

[10] Hofstadter, D. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid.* New York: Random House

[11] Jaynes, E.T. and Bretthorst, G. L. 2003. *Probability Theory: The Logic of Science.* Cambridge: Cambridge University Press.

[12] Jensen, A. R. 1999. The G Factor: the Science of Mental Ability. *Psycoloquy*, **10**(23).

[13] MacFie, R. C. 1912. *Heredity, Evolution, and Vitalism: Some of the discoveries of modern research into these matters – their trend and significance.* New York: William Wood and Company.

[14] Minsky, M. L. 1986. *The Society of Mind.* New York: Simon and Schuster.

[15] Monod, J. L. 1974. *On the Molecular Theory of Evolution*. New York: Oxford.

[16] Rice, H. G. 1953. Classes of Recursively Enumerable Sets and Their Decision Problems. *Trans. Amer. Math. Soc.*, **74**: 358-366.

[17] Sandberg, A. 1999. The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains. *Journal of Evolution and Technology*, **5**.

[18] Vinge, V. 1993. The Coming Technological Singularity. Presented at the VISION-21 Symposium, sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute. March, 1993.

[19] Wachowski, A. and Wachowski, L. 1999. *The Matrix*, USA, Warner Bros, 135 min.

[20] Weisburg, R. 1986. *Creativity, genius and other myths*. New York: W.H Freeman.