Could the AI of our dreams ever become reality?

James L. McClelland
Stanford University

Fictional portrayals of humanoid robots, including Ava in *Ex Machina*, make us think about what mental abilities really are, whether we can capture them in artificial systems, and whether humanoids could someday surpass and control us. However, these fictional robots exhibit traits no current artificially intelligent systems have. I consider some of the distance remaining between current artificial systems and truly intelligent human behavior, captured in breakthroughs that have been achieved by humans who set their own long-term goals and build on invented formal systems and the previous insights of others. At the end I consider whether an artificial system that could pursue its own goals would necessarily turn on its creators.

Keywords: Artificial Intelligence; Human Intelligence; Neural Networks

Artificial Intelligence in fiction and reality

Ava, the humanoid robot in Alex Garland's *Ex Machina*, startles us with her beauty, her sexuality, her vulnerability, and her intelligence -- and ultimately with her willingness to deceive and to exploit others' weaknesses. She seems, and of course she is, all too human, even if, in a science fiction world, she has capabilities that exceed our own. We fear her because we are all too aware of our human frailties and limitations and imagine that someday, an artificially intelligent being with all of our abilities and none of our limitations will be created and, like our human conspecifics, will be all too liable to exploit our weaknesses, leaving us unable to control of the outcome.

Watching *Ex Machina,* I was struck by how different Eva seemed to me than the artificially intelligent computer systems that we have today. It is true that in one of its matches, a contemporary artificial intelligence, DeepMind's *AlphaGo*[1]*,* made a move that no human understood or anticipated – a move widely credited with giving it an advantage that let it go on to win its match again the Korean Grand Master Le Sedol. We can marvel at *AlphaGo* and its apparent intuition and insight, and perhaps this alone is enough to spark the fears that Ava instills. Yet, *AlphaGo* is ultimately only a computer program, an object that runs entirely under the control of the scientists and engineers who created it – and perhaps, more importantly, has no will of its own. *AlphaGo* and its successor *AlphaØ* ('AlphaZero') are ultimately entirely mechanical systems whose capabilities derive from the brilliance of the computational intelligence researchers who designed it and the hardware and software engineers who turned its design into a reality. This program, which takes a board position as input and produces a legal move on its output, can learn through massive experience, while playing against a series of ever improving previous versions of itself. But an instance of *AlphaØ* that can beat every human player in the world at Chess doesn't know anything about absolutely anything else, and the same instance of the

[1] Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser et al. "Mastering the game of Go with deep neural networks and tree search." *nature* 529, no. 7587 (2016): 484-489.

To appear in: Vernallis, C., Rogers, H., Leal, J. and Kara, S. (Eds.), *Cybermedia: Science, Sound and Vision.* New York, NY: Bloomsbury Academic.

Figure 1. A Dalmatian dog emerges from an assemblage of individually uninterpretable blotches. From James, R. C. (1965), Photo of a dalmatian dog. LIFE Magazine, 58(7), 120. Copyright © 1965 Ronald C. James, permission pending.

program cannot learn to play both games at the same time. Furthermore, you can't talk to it, it can't explain itself, and it cannot learn except through millions of games of play experience.

For me it is useful to contrast today's AI systems like *AlphaGo* with a PhD student in the emerging field of computational intelligence, at the interface between human cognition and artificial intelligence. Comparing these programs to Ava is more difficult because some aspects of her abilities are difficult for me to separate from her overt sexuality and the mixed-up motivations of her creator. Setting these more fraught issues aside, in what follows I will focus on the purely intellectual side of human likeness, and on the emergence of advanced intelligent functions in researchers who go on to be independent contributing scientists. I have been lucky enough to have had many excellent PhD students and post-docs in my own laboratory over the years, and many of them have gone on to be professors at outstanding universities – or, more recently, computational intelligence researchers in AI companies. Surely, we would call these young scientists intelligent. What do they have that today's AI systems lack?

To help us consider this, I'll introduce Dana, a fictional young PhD scientist. I use as pronouns 'e, 's, and 'em to emphasize Dana's humanness while avoiding designating a gender. I'll start with more basic properties I think all humans possess, and then go on to consider what it is that makes Dana and others capable of succeeding in what I will consider to be the hallmark of intelligence: identifying and successfully addressing novel and previously unsolved challenges.

Mutual Simultaneous Constraint Satisfaction

Something very basic that humans possess and todays AI systems do not is the ability to exploit multiple simultaneous sources of information to settle into an overall interpretation of a situation and its parts or aspects and/or to formulate a plan of action that addresses many such constraints simultaneously. A beautiful visual illustration of this is provided in Figure 1. At first, we may experience this picture as an inchoate assemblage of splotches of ink, but at some point, we are likely to begin to see that the photograph depicts a dalmatian with its back toward the viewer sniffing at the ground. None of the blobs individually appear to signal the presence of a dog, but somehow, when all are considered together, the percept emerges. At the moment we see the dog, we also see the blobs differently. Some now help define the contours of its body or are seen as spots on the dog's coat, and other blobs now become scattered leaves on the ground or parts of a tree. We can even perceive the contour of the

dog's back where no actual contour is present in the image. Thus, the perception of the whole emerges from constraints provided by many aspects of its parts, and the perception of the parts depends in turn on the perception of the whole. This is what I mean by the idea of multiple, mutual constraint satisfaction.

Experiences like my seeing this photograph converged with findings in the psychology of perception and language understanding, inspiring me and others to think that it might be useful to view our perceptual systems as neural networks, because of several key properties that neural networks have that seemed to make them suitable to capture this kind of experience. The goal was not simply to simulate the brain, but to draw on the properties that might make the brain especially useful to solve this kind of constraint-satisfaction problem. The brain contains hundreds of millions of neurons, each capable of receiving inputs from up to one hundred thousand other neurons. Each neuron adjusts its activation depending on the inputs it receives from others, and in turn signals its activation to other neurons via its outgoing connections. Inspired by this idea, which we called *Parallel Distributed Processing,* David Rumelhart and I teamed up with others and drew on earlier work to develop neural network models that simulated this mutual constraint satisfaction process[2].

A key part of the inspiration for our work was the idea that the constraints influencing the outcome of perception or understanding can come from a wide range of sources. Our brains naturally and automatically integrate input from sight, sound, touch, posture, motion, smell and taste in interpreting the inputs we receive. Spoken and written language contribute to and participate in this process as well; the words and sounds we experience hearing depend on other sources of input that accompany them, and likewise the objects that we perceive through other senses are simultaneously constrained through language. Constraints affecting perception and thought can come from a wide range of mutually constraining sources.

Another potent source of constraint is input from memory. Consider this tiny story:

> John put some beer in a cooler and went out with his friends to play volleyball. Soon after he left, someone took the beer out of the cooler. Meanwhile, the volleyball match was very intense, and it seemed that John's team was going to lose. But after plenty of fierce competition, John's side was able to pull out a string of victories and won the final game when John served an amazing service winner that no one on the other team could even touch.
>
> John and his friends were thirsty after the game and went back to his place for some beers. When John opened the cooler, he discovered that the beer was ___.

In this situation, if you as a reader have been following the story, you will anticipate that the missing word is 'gone' and this will influence how likely you are to perceive it from a very brief or indistinct presentation of the word itself or a mis-spelled version of it. But if the text had said 'someone took the *ice* out of the cooler' you would instead be ready to perceive the word 'warm'. We as humans have the ability to exploit such constraints based on information we encountered in the indefinite past, not just the immediate current context.

---

[2] Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MA: MIT Press, 1986.

Finally, the considerations that may come into play are potentially unbounded and seemingly unrelated to a particular situation at hand. I believe I heard a version of the anecdote below from Jerry Fodor. Whether it really happened I don't know, but it seems to capture something real about how we think.

> Jeff, a good bridge player, has just bid six Hearts and is about to start play on the last bridge deal at the end of an evening at his bridge club. Another player, Al, from a table that has just finished its last deal, comes over, walks around the table to see the hands of all of the players, and lingers to observe the play. The player to Jeff's left makes the opening lead. As Jeff's partner lays down the dummy hand, Jeff surveys the situation. It looks like an easy contract. But Jeff notices that Al is still hanging around. This makes Jeff think: maybe the hand is not such an easy one after all. If it were, Al would surely have lost interest by now. He ponders: what could conceivably go wrong? Seeing only one possibility—one that would ordinarily seem remote—he devises a plan of play that would ordinarily fail but succeeds in this case, and triumphantly, he makes his contract. His opponents are outraged and complain to the director. But the director can do nothing, since Al never said or did anything that was against the rules in any way.

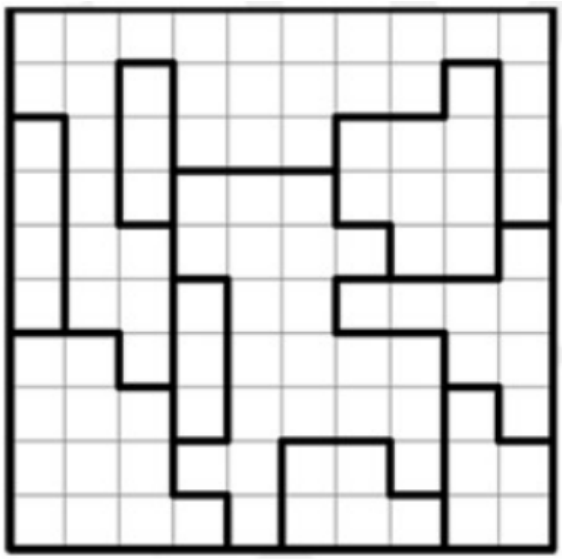
Here Jeff is using information from outside of the domain of the game itself to reason about what to do within the game. It was Fodor's point, and one that I agree with, that there is no limit on the constraints that we can ultimately bring to bear when we think and reason. In other words, the constraints that can enter into our mental constraint satisfaction process are completely open-ended.

I have described here what to me are extremely basic aspects of human intelligence, ones that we all possess. In spite of the fact that most of the recent breakthroughs in AI are based on artificial neural networks similar to the ones Rumelhart and I used in our early work to capture the mutual constraint satisfaction process, today's networks are generally far narrower in the constraints they consider than we as humans are. AI researchers at DeepMind and elsewhere are aware of these limitations, and progress has recently been made in creating language systems that can begin to bring a very wide range of information from context to bear in language comprehension. Researchers are actively exploring how to combine many input modalities and how to exploit relevant information presented only once at an arbitrary past time now out of mind. Much still remains to be accomplished here, however. Furthermore, the prospect of being able to bring completely open-ended considerations to bear as in our bridge game example remains an important future challenge.

Metacognition, Explanation, and Discourse

Another area where Dana, and humans in general, far exceed our current AI systems is in the ability to think about and exchange ideas with others about our own thought processes or to describe the reasons for the decisions, actions and predictions that we make. It seems fair to say that *AlphaGo* and most other contemporary AI systems are completely devoid of these abilities. Returning to the surprising move that *AlphaGo* made against Lee Sedol, the computer had no ability to explain why it chose the move it did. In contrast, during the match, human commentators provided a running commentary, describing the pros and cons of each move made by both the computer and the human player, and speculating on whether or not *AlphaGo's* move was a brilliant stroke of genius or a wild stab in the dark.

I do not mean to say that we as humans have perfect access to the basis of our own perceptions, feelings, and choices of actions. Those interested in human thought have been aware since the late 19th century that introspection is often uninformative or completely misleading. Yet we can and do share



Place 2 x's in each row, column, andbounded region. No two x's can be adjacent, not even diagonally.

Figure 2. A two not touch puzzle, with instructions as they appear with puzzles published in the New York Times. From Bumgardner, J. (2020). Two Not Touch Puzzles, https://krazydad.com/twonottouch/, accessed January 3, 2021. Copyright © 2020 www.crazydad.com, (permission pending)

information with each other that we can use to immediately alter our behavior – something that is not possible for machine systems like *AlphaGo* that simply learn to get better through massive experience.

As one simple example of this, consider the puzzle shown in Figure 2. You are given a grid and an instruction specifying a task goal, and without any further experience, most people I've shown this puzzle to can begin to perform the task of placing x's in the grid, and I have solved many such puzzles without any further instruction on how to solve them. Many contemporary AI systems could learn how to play this game very well, but they would *either* require the programmer to build quite a lot of the solution into the program *or* they would require a vast amount of experience, or a combination of both. Furthermore, I can point out things to you that you can use to help you play the fame. First, I can tell you that it is useful to try to determine where x's *cannot* go in the grid, marking these cells, say, with a small dot. Then I can point out that if all of the cells that remain possible places where you can put an x within a given enclosed region are within the same row or column, you can be sure there can't be an x in any of the other cells in that row or column. With this information, you can then find more cells that cannot contain an x, and we are well on your way to an overall solution.

What is more than this is that we as humans can make these observations and share them with each other. The above paragraph is evidence of this. No one told me the points I made above, but as I practiced solving these puzzles, I started making these observations to myself. I emphasize that I do not consider myself to be especially gifted in these ways, though I do believe my past experiences have helped set the stage for me to do this, at least in part.

Where our ability to engage in meta-cognition comes from is an open scientific question. One could hold that it is something that evolution endowed us with, or one could hold that evolution and culture gave us language, and with language we developed the ability to understand and give explanations, and once these abilities developed, we became able to use language to make observations for ourselves. The recent AI language system GPT-3[3] may have some abilities along these lines. This system was trained on a vast corpus of language including quite a lot of transcribed human discourse. Since such discourse contains examples of explanation, it is possible that the system would, if assessed, be able to give some form of self-explanation. Suppose we gave it the passage about John and his beer. Because this passage will fit in GPT-3's buffer, it may be able to predict that the missing word should be 'gone'. Suppose we continued the story, 'The beer was missing because …' and then let GPT-3 complete this sentence. Perhaps it would go on to say 'someone had taken the beer out of the cooler'. It is conceivable that GPT-3 would even come up with this explanation, given the right kind of relevant experience, even if the previous sentence about the removal of the beer had not been included. This is an area where we still have a lot to learn about what we need to build into our AI systems for them to begin to exhibit abilities we take for granted as humans.

The Role of Culturally Invented Modes and Tools for Thinking

The abilities I have described above are abilities all humans rely on every day. Multiple constraint satisfaction is always in play as we identify spoken words and recognize objects or make everyday motor planning and action decisions. Whenever we discuss the events of our day, the behavior of others, politics, the weather, or anything else we are always engaged in explanations and teaching each other through discourse and discussion. A human graduate student like Dana engages in this kind of discussion as well. For example, Dana may explain to me a plan to analyze the data collected in an experiment, and we might discuss alternative approaches before we settle on a particular plan. Dana will then go off and execute what we have discussed, based on material learned in a statistics course, which in turn involves a lot of direct instruction. I certainly believe Dana and other graduate students learn gradually from experience as well, and that expertise ultimately does depend on a great deal of experience; but I think today's AI is missing out on the tremendous leverage that instruction and explanation can provide.

However, to be truly successful as an advanced practitioner of a discipline such as computational intelligence, Dana also needs additional skills that, I believe, depend on acquiring specialized mental modes of thought and tools for thinking that aid and support the efforts of skilled experts. This is just as true I believe in the arts and humanities as it is in the sciences, and so in this section I will draw my examples from both domains, but with the primary focus on science, since that's the domain I know more about.

Doing science requires mastery of an extensive set of tools for thought, in conjunction, perhaps, with a kind of meta-level tool for thinking that I will call *formal thinking ability.* Some examples of specific tools are the ability to develop sound logical arguments, to solve problems that require the use of mathematics, to prove mathematical theorems, and to write computer programs that accord with the

[3] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

conventions of complex and highly structured programming languages. To make a contribution in science today, one must rely heavily on many of these tools.

One example of such a tool is propositional logic. For much of the 20th century, logic played a central role in widely-held conceptions of these abilities. Bertrand Russell said 'All of mathematics is symbolic logic', and so central was logic to mid-century conceptions of intelligence that Herb Simon, a leading early figure in AI research, was able to say in 1953 'Over the Christmas holidays, Allan Newell and I programmed a computer to think'[4]. Their computer program proved simple logic theorems, and he was using a system that owed its very essence to the traditions of logic that were instantiated in the architecture of the digital computer. Decades later, Fodor and Pylyshyn[5], argued that thought is, essentially, the manipulation of structured assemblies of symbols according to structure sensitive rules, and used the logical syllogism called *modes ponens* as their central example. It goes like this. If you know that some proposition *p is true*, and you know that *if p is true, then* some other proposition *q is true*, then you can conclude that *q is true.* So if you know (*p*): *John is strong*, and you know (if *p* then *q*): *If John is strong, then John will beat Bill at armwrestling,* you can conclude (q) *John will beat Bill at armwrestling.* You can do this, they argued,without regard to the actual content of the propositions. This is the kind of thing that Newell and Simon relied on in their computer program.

I find myself in partial agreement with these views. This may be surprising, because I believe that formal thinking is not the natural mode of human thought, and that it can get in the way of mutual constraint satisfaction. In fact, humans don't actually succeed with arbitrary propositional content, (which is why I used an example that appeals to prior knowledge). More importantly, today's neural network-based AI models can be seen as refutations of Fodor and Pylyshyn's arguments, since their successes seem to come in part from the fact that they expressly eschew commitment to foundational principles of formal thinking. For example, today's AI language translation systems are neural network-based systems that do not rely on the systems of rules that Fodor and Pylyshyn argued were central to human thought and language processing. Yet, there's no doubt that formal systems have played a huge role in supporting our ability to understand our universe well enough to create and control nuclear reactions, to create computers, and to create technologies that have allowed humans to direct spaceships that will intersect with the orbits of tiny objects in the vast space at the edges of our solar system.

One approach some cognitive scientists advocate is that we must build systematic, symbolic reasoning into our artificially intelligent systems. This is the approach advocated and exploited by Josh Tenenbaum at MIT and many of his collaborators and associates. Acknowledging the usefulness of neural networks, this group has recently explored what they call the "neurosymbolic" approach to capturing intelligence[6], which relies on computational systems that use neural networks for processing inputs and controlling outputs, but rely on more symbolic approaches to capture the part of the process that Herbert Simon thought of as thinking. Their systems also exploit sophisticated advances in

---

[4] Simon, Herbert A. *Models of my life*. MIT press, 1996.

[5] Fodor, Jerry A., and Zenon W. Pylyshyn. "Connectionism and cognitive architecture: A critical analysis." *Cognition* 28, no. 1-2 (1988): 3-71.

[6] Mao, Jiayuan, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision." *arXiv preprint arXiv:1904.12584* (2019).

probabilistic reasoning, which makes them more powerful than the systems Russel, Noam Chomsky, and later Fodor and Pylyshyn relied on.

For my part, I am pursuing an approach in which systematic mental processes arise from the structuring of our minds that occurs through exposure to and mastery of the tools of thought I mentioned previously[7]. On this view, these tools are human inventions that began to emerge as humans started to develop technologies and civilizations. Gradually institutions arose within these civilizations, creating notation systems and artifacts that supported the further development of these systems, and that then structure the minds of those who immerse themselves in them, giving them the ability to build on the ideas of those who went before them to exploit and extend these systems. Number systems are good examples of these kinds of formal systems. Primitive cultures may lack number systems, having only words for very few, some, and many, as was documented by Peter Gordon in an important paper in 2008[8]. Many cultures have invented or adopted such systems from other cultures, but even throughout most of the first millennium of the current era, the number systems used in the west were cumbersome and unsystematic. The base-10 place value system used world-wide today is the product of cultural innovations and makes possible the creation of tools such as the abacus and mechanical calculators that vastly enhance the power and efficiency of human reasoning about number. Like our number system, geometry, trigonometry, calculus, logic, probability theory, and computer programming are all examples of culturally constructed systems and tools that vastly increase the power of human reasoning.

While I am certainly more of a scientist than an artist or musician, my exposure to art and music history during my undergraduate years taught me that the same points apply in these domains as well. The towering achievements represented by the painting, sculpture and architecture of Michelangelo or the musical compositions of Beethoven depended crucially on the developments introduced by their predecessors and were achieved after decades of immersion in the study of these prior developments, many of which have strong formal elements. For example, in music, the twelve tone scale, the various modes within this scale, the notational systems invented to allow explicit representation of values and durations of musical notes, and the further structures built on top of them such as the sixteen bar frame of most songs and the basic structure of sonata form, etc, are all cultural inventions, as are the actual instruments musicians use to render the resulting patterns acoustically, from simple drums and flutes to the well-tempered clavier. These conventions and tools underlay the achievements of Bach, Beethoven, and others, and subsequent extensions including new tonalities, rhythms, and tools such as synthesizers further extend these resources, allowing further developments in the nearly 200 years since Beethoven's last compositions.

To summarize my point in this section, I turn to the views of Henri Poincaré and Albert Einstein. Clearly, these are individuals who anyone would have to describe as intelligent. Poincaré, the 19th century French mathematician, physicist, and engineer, wrote 'It is by logic that we prove, but it is by intuition

[7] McClelland, James L. Are humans still smarter than machines? Manuscript in preparation, Department of Psychology, Stanford University, February, 2021, based on the Graham Lecture at the University of Torono recorded October 20, 2020, on YouTube at https://www.youtube.com/watch?v=9ysH58hQ2n0&feature=youtu.be

[8] Gordon, Peter. "Numerical cognition without words: Evidence from Amazonia." *Science* 306, no. 5695 (2004): 496-499.

that we discover'[9], and Einstein is said to have viewed the intuitive mind as a sacred gift and the rational mind as its faithful servant[10]. Both Poincaré and Einstein seem to identify the essence of insight and discovery with intuition rather than logic and rational thought, and I share that perspective.

Goal Directed Thinking

The final difference I would like to mention between Dana and today's AI is that Dana is in the process of becoming more and more self-directed. I remember that when I was a PhD student this was a very big issue for me. My first advisor had strongly steered my first year research project toward a very specific issue that was of interest to him, and I found myself needing to set my own direction. When I work with my own PhD students, I always experience the same tension. How much should I steer them toward addressing my agenda? How much should I leave them alone to pursue their own direction? Since I always have things I am anxious to pursue, I certainly always share with them the things that excite me the most. However, I've also found that our mutual experience together is always better if I seek to work with my students to find a project that is of mutual interest. In Dana's case, as with many of my students, we settled on a primary research project, something that Dana expressed interest in the first time we met. Dana is seeking to make a contribution to knowledge by working to address the issue of how to learn new things quickly in a superpositional memory – a memory that does not stick each new item into a separate slot, but instead superimposes them, as in holographs or film exposed to several images. This is a largely unsolved problem in cognitive neuroscience, and Dana and I agree that interesting new steps are possible. As Dana's mentor I expect we will work together fairly closely initially, with Dana taking greater and greater ownership in the project as it progresses, though I hope to remain involved in finding the solution, rather than just helping to get the project going. In other words, the project is one that will, I hope, satisfy both of us as making progress toward an important goal, one that is good for science, for our reputations, and our careers.

In this regard, Dana is far different from current systems like *AlphaGo* and *GPT-3.* These systems have no independent agency whatsoever. Every computation they perform, every input they receive, and every output they generate, is entirely under the control of the scientists and engineers who design and run them. The designer creates what is called an objective function – a mathematical expression that characterizes the adequacy of the learner's performance, in terms the designer specifies. All of the learning in the system is directed toward maximizing performance as measured by this function (or minimizing the discrepancy from perfect performance, often called the *loss*).

It is true that there have been efforts underway for many years to create learning systems that explore their environments on their own and many thoughtful AI researchers are seeking to design systems with intrinsic goals that can lead to self-discovery. One such approach is to give a system the goal of producing novel experiences which then drive learning toward a deeper understanding than was possible based on the experiences the agent might have been exposed to passively. Progress is being made, and it will be interesting to see how far such research will go. I feel that an important place for the field to focus going forward will be on developing artificial systems that actually work toward

---

[9] Poincaré, Henri. *Science and Method*. (1908). (*Science et méthode* (1908), as translated by Francis Maitland (1914) and republished by Cosimo, Inc, New York) Part II. Ch. 2 : Mathematical Definitions and Education, p. 129.
[10] Samples, Bob. The Metaphoric Mind: A Celebration of Creative Consciousness. United Kingdom: Addison-Wesley Publishing Company, 1976. p 26.

particular goals, rather than simply focusing on improving performance by a global desire to experience novelty.

Being self-directed has, historically, been important for productive intelligence, where I define this as the ability to make a novel contribution. The history of science is the story of how independent thinkers revolutionized the way we understand the world around us. Galileo was found guilty of heresy for the new insights he contributed, and Newton and Einstein both revolutionized understanding of Physics. Likewise, Michelangelo and Beethoven are known as highly self-directed individuals who went beyond the achievements of their predecessors to achieve more than had ever been possible before.

Ex Machina raised this issue and a lot of other science fiction also touches on it. For an artificial being to be truly intelligent, must it also be completely self-directed? This is an important and interesting question. For us as humans, it often appears to be so, but I would offer two points that make me uncertain about whether this necessarily applies to all beings that can truly make innovative discoveries.

First, concerning humans, our goals are not, in my opinion, entirely our own. It is at least arguable that humans can have goals for others, or commitments to ideals, rather than just for themselves. Soldiers who are sent to war or health practitioners at the front lines of battling contagious diseases, as well as leaders of social justice movements may have goals that place the collective good ahead of their own personal ambition. Indeed, being able to pursue a goal that is greater than oneself is an important source of inspiration. Often throughout human history the truly innovative thinkers have appealed to someone or something greater than themselves, producing profoundly influential innovations.

Second, when it comes to artificial beings, the fact that a system like *AlphaGo* can come up with innovative moves leaves me wondering how much autonomy is strictly necessary. It seems arguable that deciding to seek an explanation for something that seems intuitively puzzling might require some degree of autonomy, but not necessarily the kind of autonomy that pits the artificial system against its creators. We should of course be wary of the possibility – one that *Ex Machina* and other science fiction has repeatedly raised – that we *might* be in danger of losing control. Speaking for myself, however, I am more worried about nefarious human uses of artificial intelligence than I am about losing control to autonomous artificial beings.

Final thoughts

From the thoughts I have expressed in this essay, it should be clear that, in my view at least, human intelligence still far exceeds artificial intelligence. However, I would like to note that artificial systems play an increasingly important role in *augmenting* human capabilities. Because these systems provide tools and resources that humans otherwise lack, they have enabled the development of systems that precisely target locations in the vast three-dimensional space of the outer reaches of our solar system or that allow us to predict how extremely complex chemical structures (usually proteins) will fold on themselves and interact with each other. More and more powerful extensions of human abilities will continue to be possible, thanks to the ever-increasing power of these systems. Of course, like other innovations, they can be used for good or ill and as citizens it is our crucial task to make sure there are governing bodies in place to oversee them as we oversee all other technologies. What remains to be seen is whether we come to see artificial systems as potential threats or competitors to ourselves. I am cautiously optimistic that we will be able to create systems that pursue prosocial goals, including the goals of encouraging our own sense of individual autonomy and agency.

Bibliography

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

Fodor, Jerry A., and Zenon W. Pylyshyn. "Connectionism and cognitive architecture: A critical analysis." *Cognition* 28, no. 1-2 (1988): 3-71.

Gordon, Peter. "Numerical cognition without words: Evidence from Amazonia." *Science* 306, no. 5695 (2004): 496-499.

Mao, Jiayuan, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision." *arXiv preprint arXiv:1904.12584* (2019).

McClelland, James L. Are humans still smarter than machines? Manuscript in preparation, Department of Psychology, Stanford University, February, 2021, based on the Graham Lecture at the University of Torono (same author and title) recorded October 20, 2020, on YouTube at https://www.youtube.com/watch?v=9ysH58hQ2n0&feature=youtu.be

Poincaré, Henri. *Science and Method*. (1908). (*Science et méthode* (1908), as translated by Francis Maitland (1914) and republished by Cosimo, Inc, New York)

Rumelhart, David E., McClelland, James L., and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MA: MIT Press (1986).

Samples, Bob. The Metaphoric Mind: A Celebration of Creative Consciousness. United Kingdom: Addison-Wesley Publishing Company, 1976.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser et al. "Mastering the game of Go with deep neural networks and tree search." *nature* 529, no. 7587 (2016): 484-489.

Simon, Herbert A. *Models of my life*. MIT press, 1996.