Ryan Filgas
AI Ethics Response

The book “Artificial Intelligence: A Guide for Thinking Humans” goes through 4 different subject areas relating to the challenge of creating a general artificial intelligence. Looking and seeing, Learning to Play, Natural Language Processing, and the Barrier of meaning.

Part 1: Like other overviews on AI and machine learning the author prefaces by describing some of the history of ML and AI. Particularly, much focus was given for Turing tests and the running bet between Kurzweil and Kapor on whether or not their upgraded turing test could be satisfied by 2029 using 2 hour sessions rather than 5 minutes. This benchmark was moved because the 5 minute mark was too easy to solve (p 49). Following this is the debate on when “the singularity” will happen – an AI that can create smarter AI. The argument against this is that hardware may be growing at high rates, but software has not. For this reason, some argue it's unlikely we will reach this achievement by the 2040s.

Part 2: The events that kicked off the current AI spring revolve around computer vision, specifically CNNs. Based off research by Hubel and Weisel on how the brains visual system works Yann LeCun proposed CNNs in the 1980s (p73). Following this the next explosion in growth was triggered by the advent of ImageNet created using mechanical turk. Alex Krizhvsky supervised by Geoffrey Hinton and working with Ilya Sutskever achieved 85% accuracy on ImageNet, an 11% jump from the previous year. Since then, the last imageNet competition achieved 98% and convNets have become commercially popular.

Due to the data quantity needed to train convNets the incongruencies between human and machine learning indicated a general intelligence requires unsupervised learning. Models trained on image sets may be overfitted to the data making them less practical. Other problems presented are that neural networks exhibit systemic bias, adversarial attacks are difficult to mitigate, regulation hasn't been implemented for error prone facial recognition (p 117), and that it's difficult to agree on morals to program in AI systems (p128).

Part 3: Part three focused on reinforcement learning and overviewed Atari games and games like chess and go. Specific game focused algorithms covered included minimax and monte carlo tree search. One notable point in this section was that random search via initialized neural networks and genetic algorithms were able to outperform Deep Q Networks in less time than it took to train them. This indicated that perhaps these problems aren't difficult benchmarks.

Part 4: The author discusses some of the current problems in NLP. Notably NLP algorithms pick up on systemic bias and they have difficulty interpreting contextual clues and sarcasm. In addition, problems arise with language translation as cultural expressions or ambiguous words provide stumbling blocks to good translations. This is being partially addressed using recurrent neural networks, but there remains a lot to be done.

Part 5: Part five centered around meaning. To the extent that algorithms can interpret input and match that up with answers in a database they don't understand meaning and rely primarily on immediate context and frequency. Computational creativity is an area where the author argues machines may exhibit characteristics of creativity, but the crucial point is that they don't understand what is good or meaningful and therefore lack the capacity of creativity, at least right now. The author ends by going over some burning and unanswered questions about the trajectory essentially saying that the field is less worried about a mythical evil AI and more worried about a "dumb AI" that is prevalent everywhere, essentially referencing that they are error prone and easy to attack.

Ryan Filgas
AI Ethics Response

Response: I found the main challenges presented to match up well with what I've understood in prior classes. Part 1 discussed Turing tests which is an interesting topic because it's a measure of how human-like an AI is. Without delving into speculation too much the obstacles to achieving this benchmark are high. The author presented an opinion held by many that to achieve a certain level of humanity an AI would have to go through many of the same cultural experiences as a human and make observations not only about cultural subjects, but also the innate experiences we take for granted: traditional cause and effect scenarios, an intuitive understanding of physics and how that relates to our environment – that is to say a database of knowledge may not be enough. I would agree that there is a lot of truth to this, and while it will be possible to satisfy a variety of needs with AI, acting perfectly human may not be possible with clever programming alone.

The main ideas in part 2 surrounded the challenges in creating robust vision systems as and regulating them. The example that stood out was when members of congress were run through a facial recognition system, and some were falsely identified in a criminal database with a disproportionate amount being black. Given inaccuracies in these systems the technology of facial recognition should be regulated and taken with a grain of salt as the consequences of misidentification are serious. If the problems concerning adversarial attacks can't be mitigated, many applications of computer vision models will be hampered until this is addressed.

In part 3 it was interesting that random weight assignments were faster to come up with solutions to classic games as well as genetic algorithms. It suggests that trivial enough problems can be solved using relatively aimless random methods over intelligent methods that are guaranteed to eventually converge to a workable solution. It also demonstrates that simple solutions can exist for seemingly complex problems and begs the question “is there an easier way”. I believe this is important in the context of ML models in general, as it may not make sense in some cases to apply a complex model to a problem when a simple one will do.

Part 4 exposed a lot of the current problems in NLP. The systemic bias in these algorithms was nothing new to discover, but the deeper dive into why models make mistakes was insightful. The discussion focused on IBM Watson for example was a lesson in how answering questions doesn’t necessarily demonstrate understanding. The discussion about translation software demonstrated the same thing. True understanding for AI models is still a long way off, however it has been making measurable progress.

In terms of the computational creativity discussion in part 5 this topic hits home as my first degree is in Studio Arts. Understanding on the surface how these models work the creative process is essentially the same for a human in many ways who has a specific goal. The differences as I understand them come in two parts. The first the author presented is that computers can’t yet interpret whether what they’ve created is good or has meaning. The models also lack the expression of human life. One thing made important in the art world is that beauty isn’t the goal of expression. There should be a goal, a purpose, an expression of a thought or feeling. These measures have not all been achieved and might not be for some time. The technology does threaten traditional creation processes as well as art whose only purpose is beauty and this has caused some concern in the art community. This set of art also includes small and independent makers and risks job displacement essentially. Overall, the book was a great overview on the benefits and risks of AI as well as the history of advancement.

Ryan Filgas
AI Ethics Response

Sources Cited

Mitchell, M. (2019). *Artificial intelligence: a guide for thinking humans.* First edition. New York, Farrar, Straus and Giroux.