income) and encourages a reexamination of a theology of work. Finally, Peckham's last critique of AI centers on its implementation in video games and virtual reality. Peckham fears that these digital realities present a slippery slope for users who will be unable to differentiate between true reality and digital reality.

In the final two chapters (eleven and twelve), Peckham considers a Christian response to AI progress along with developing a Christian manifesto toward AI research and usage. Rather than utilizing AI technology mindlessly or carelessly, Peckham exhorts the reader to seriously consider the substantial influence AI has upon the individual and how AI development should be regulated moving forward. To properly consider and regulate AI, Peckham argues that a Christian worldview provides the best framework with which to understand humanity and our relationship with technological artifacts. Thus, his brief Christian manifesto serves to introduce how Christians can have a voice in the AI conversation.

Peckham's educational and vocational background in computer technology serves him well in writing this book. He has worked on computer and AI technology in both the government and commercial sectors. With his background in various AI technologies, Peckham understands how AI technology is built, how it functions, and the intentions behind the design. This is a strength of the book since many Christians who discuss AI often lack the requisite training and expertise.

Although Peckham does understand AI technology well, he does not examine the ontological considerations of AI. Peckham looks mostly at the effects of AI technology and then tries to develop a critique of that technology rather than relying on more philosophical arguments. Peckham's critique throughout the chapters would be stronger if he considered an ontology of AI or provided a more detailed explanation of what AI is before presenting his critique. At several points throughout the book, Peckham implores the reader to consider the harmful consequences of AI technology, but he does not look into the deeper fundamental philosophical presuppositions.

In addition, chapter ten, addressing video game AI and virtual reality technology, comes across as outdated, restating many of the traditional Christian arguments against video games. While Peckham does helpfully highlight the new AI technologies used in video games (such as augmented and virtual reality), his criticisms of video games ignore the numerous variations of games as well as the communities built around video games. By presenting a familiar Christian critique, Peckham risks dismissing some of the more-recent developments in the video game industry as well as alienating readers who are active within that community.

Overall, *Masters or Slaves?* is a welcome addition to the growing Christian literature on AI. In comparison to other recent Christian publications on AI, such as Jason Thacker's *The Age of AI* or John Lennox's *2084*, Peckham's contribution has a stronger technical foundation due to his extensive background in the technology. Peckham expresses moral concerns similar to those of other authors about the development of AI, while covering a large number of areas that AI currently, or will inevitably, affect. Although Peckham could certainly provide even more background on specific AI technologies, his book serves as an excellent introduction to a Christian response to AI.

Reviewed by Eddy Wu, IT Operations Manager and PhD student at Southeastern Baptist Theological Seminary, Wake Forest, NC 27587.

THE ALIGNMENT PROBLEM: Machine Learning and Human Values by Brian Christian. New York: W.W. Norton, 2020. 344 pages. Hardcover; $28.95. ISBN: 9780393635829.

The global conversation about artificial intelligence (AI) is increasingly polemic—"AI will change the world!" "AI will ruin the world!" Amidst the strife, Brian Christian's work stands out. It is thoughtful, nuanced, and, at times, even poetic. Coming on the heels of his two other bestsellers, *The Most Human Human* and *Algorithms to Live By*, this meticulously researched recounting of the last decade of research into AI safety provides a broad perspective of the field and its future.

The "alignment problem" in the title refers to the disconnect between what AI does and what we want it to do. In Christian's words, it is the disconnect between "machine learning and human values." This disconnect has been the subject of intense research in recent years, as both companies and academics

continually discover that AIs inherit the mistakes and biases of their creators.

For example, we train AIs that predict recidivism rates of convicted criminals in hopes of crafting more accurate sentences. However, the AIs produce racially biased outcomes. Or, we train AIs which map words into mathematical spaces. These AIs can perform mathematical "computations" on words, such as "king – man + woman = queen" and "Paris – France + Italy = Rome." But they also say that "doctor – man + woman = nurse" and "computer programmer – man + woman = homemaker." These examples of racial and gender bias are some of the numerous ways that human bias appears inside the supposedly impartial tools we have created.

As Norbert Wiener, a famous mathematician in the mid-twentieth century, put it, "We had better be sure the purpose put into the machine is the purpose which we really desire" (p. 312). The discoveries of the last ten years have shocked researchers into realizing that our machines have purposes we never intended. Christian's message is clear: these mistakes must be fixed before those machines become a fixed part of our everyday lives.

The book is divided into three main sections. The first, Prophecy, provides a historical overview of how researchers uncovered the AI biases that are now well known. It traces the origins of how AI models ended up in the public sphere and the history of how people have tried to solve the problems AI creates. Perhaps one of the most interesting anecdotes in this section is about how researchers try to create explainable models to comply with GDPR requirements.

The second section, Agency, explores the alignment problem in the context of reinforcement learning. Reinforcement learning involves teaching computer "agents" (aka AIs) to perform certain tasks using complex reward systems. Time and time again, the reward systems that researchers create have unintended side effects, and Christian recounts numerous humorous examples of this. He explains in simple terms why it is so difficult to correctly motivate the behaviors we wish to see in others (both humans and machines), and what it might take to create machines which are truly curious. This section feels a bit long. Christian dives deeply into the research of a few specific labs and appears to lose his logical thread in the weeds of research. Eventually, he emerges.

The final section, Normativity, provides perspective on current efforts to understand and fix the alignment problem. Its subchapters, "Imitation," "Inference," and "Uncertainty," reference different qualities that human researchers struggle to instill in machines. Imitating correct behaviors while ignoring bad ones is hard, as is getting a machine to perform correctly on data it hasn't seen before. Finally, teaching a model (and humans reading its results) to correctly interpret uncertainty is an active area of research with no concrete solutions.

After spending over three hundred pages recounting the pitfalls of AI and the difficulties of realigning models with human values, Christian ends on a hopeful note. He postulates that the issues discovered in machine-learning models illuminate societal issues that might otherwise be ignored.

> Unfair pretrial detection models, for one thing, shine a spotlight on upstream inequities. Biased language models give us, among other things, a way to measure the state of our discourse and offer us a benchmark against which to try to improve and better ourselves … In seeing a kind of mind at work as it digests and reacts to the world, we will learn something both about the world and also, perhaps, about minds. (p. 328)

As a Christ-follower, I believe the biases found in AI are both terrible and unsurprising. Humans are imperfect creators. While researchers' efforts to fix biases and shortcomings in AI systems are important and worthwhile, they can never exorcise fallen human nature from AI. Christian's conclusions about AI pointing to biases in humans comes close to this idea but avoids taking an overtly theological stance.

This book is well worth reading for those who wish to better understand the limitations of AI and current efforts to fix them. It weaves together history, mathematics, ethics, and philosophy, while remaining accessible to a broad audience through smooth explanations of detailed concepts. You don't need to be an AI expert (or even familiar with AI at all) to appreciate this book's insights.

After you're done reading it, recommend this book to the next person who tells you, with absolute certainty, that AI will either save or ruin the world.

Christian's book provides a much-needed dose of sanity and perspective amidst the hype.

Reviewed by Emily Wenger, graduate student in the Department of Computer Science, University of Chicago, Chicago, IL 60637.

THE MYTH OF ARTIFICIAL INTELLIGENCE: Why Computers Can't Think the Way We Do by Erik J. Larson. Cambridge, MA: Belknap Press, 2021. 312 pages. Hardcover; $29.95. ISBN: 9780674983519.

The Myth of Artificial Intelligence (AI) offers a technical and philosophical introduction to AI with an emphasis on AI's limitations. Larson, a computer scientist and tech entrepreneur, keeps his central claim modest: true general AI is neither inevitable nor imminent, and if it is possible, it will require fundamentally new approaches. It is an easy read, combining references to fiction, history, and science. It lays out a bird's eye view of the origins and ideas behind current AI methods, focusing on general AI, a category of AI that would need to learn and engage with a wide variety of problems.

Separated into three parts, *The Myth of AI* begins with the history and algorithmic logic of AI, largely through the lens of the Turing test. Larson argues that we are not near the singularity (superintelligent computers able to create ever more intelligent machines) and that, in fact, the basic premise of the singularity is flawed.

The second part discusses inference. AI falls short of human intelligence because it can work with hard rules, but cannot make the guesses necessary to formulate new ones or handle uncertain rules. In attempts at the Turing test, AI can throw data at the problem but will always lack understanding. Achieving the understanding necessary for true intelligence will require an approach fundamentally different from recent advances made in AI, which are only effective for narrow AI (a category of AI for solving specialized problems) and not general AI.

The final, and relatively brief, part examines AI in science. According to Larson's assessment, new scientific research relies heavily on newly available computation power and big data in order to use narrow AI to its full extent. Larson claims that this approach will hinder development of new theories. He also claims that this leads to treating scientists as if they were computers as well, which causes overvaluing the system of science above people. He criticizes "swarm science," which he describes as a large group of scientists approaching one problem with a variety of projects, emphasizing this collaboration over the individuals. Instead, he claims, we need our culture to continue to emphasize individual discovery and intelligence, as it is the key to innovation.

Through the discussions of the history, philosophy, and logic of AI in the first two parts of the book, Larson disentangles the hype of AI from what is actually possible with current technology. Even as he sheds light on the gap between the singularity prediction and what machine learning is truly capable of, he emphasizes the significance of the myth. "The myth is an emotional lighthouse by which we navigate the AI topic" (p. 76). The stories we tell through predictions and science fiction define AI in the public eye and set the goals for AI research.

Our underlying philosophy matters as much as the current state of AI research, when we consider the social role of AI and what we predict for our future. In the development of AI, we must define intelligence and explore what it means to be human. While this is not a book with overtly religious claims, it does acknowledge the spiritual claims inherent in discussions of personhood. It also frames technoscience as replacing philosophy and religion and as the oversimplified understanding of humanity and the precursor to expectations of the singularity.

Beyond the stated goal of disenchanting the reader of the inevitability of AI, the book highlights the significance of stories to both society and science and emphasizes the importance of understanding for both humans and AI. We need to understand not only the technical aspects of the technology we build but also the philosophy that defines our goals.

While I found the first two sections of the book to be an engaging and accurate discussion of the tension between the science and hopes of AI, I had concerns about the warnings of "swarm science" in the third. Larson is placing a strong emphasis on individual genius in science; however, science has never been a truly independent endeavor. Many times in history, from evolution to DNA, multiple teams of scientists independently made the same discoveries at nearly the same time, based on previously published work. Though these discoveries were not inevitable, they