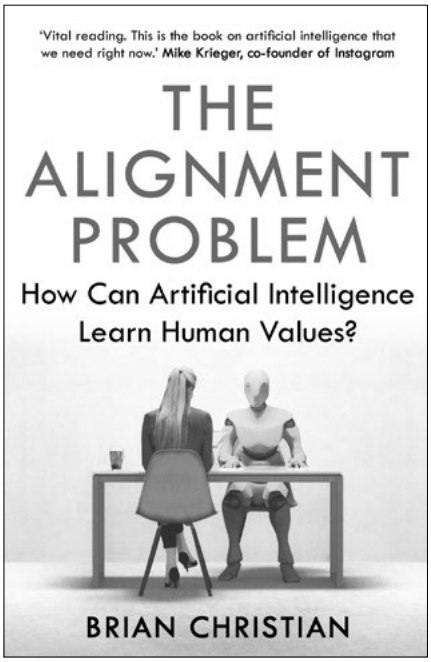
Brian Christian: The Alignment Problem: How Can Artificial Intelligence Learn Human Values?

Reviewed by Michael Haiden



In *The Alignment Problem,* author Brian Christian talks about humans, animals, and – centrally – artificial intelligence. As book's title suggests, his focus is the 'alignment problem', more precisely the task of ensuring that artificial agents behave the way we want them to. The author at one point quotes MIT researcher Norbert Wiener, who in 1960 stated the problem as follows: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it […], then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it." (295). As Christian makes clear, this is easier said than done.

The first part of the book (*Prophecy*) fulfils two purposes: it shows us the problems we might encounter when we deploy artificial intelligence, and it explains why these problems matter. The author confronts us with the amazing abilities of artificial intelligence, learning faster than any biological agent, recognising patterns better than the most intelligent human – sometimes even seeing things we miss altogether.

One is easily convinced that we are standing in front of a powerful tool. However, the positive outlook is dimmed by the problems Christian outlines. The crucial challenge in this part is the possibility of algorithms making 'wrong' decisions – meaning that the machine acts not as we want it to. There are various reasons for this. For example, a lack of training data leads a Google AI to classify images of black people as Gorillas, because it does not have enough pictures of black people in its database (25–26). Another instance is that when AI is used to decide whether criminals deserve parole, it treats black people much harsher than white people (60). This is not because black people actually are more likely to offend again – rather, the algorithm makes a decision based on the data we provide, in which black people are more likely to be caught offending, due to the over-policing of black neighbourhoods (76).

It is not that the algorithm is knowingly biased, as Christian stresses. It simply makes decisions based on our (biased) data. Thus, the complexity of these technical issues mirrors the complexity of the societal problems that underlie them. And for such complex problems, there are no easy solutions. For instance, if one wants to avoid biased outcomes based on race, it is not enough to remove race as an attribute from the data, because the impressive pattern-recognition of AI allows it to still see relationships between race and the attributes that correlate with it – something called 'redundant encodings'. In a society where black people are arrested more often than white people, the number of arrests can be tied to race. Put simply, those who are arrested more often will be judged more negatively by the algorithm – and those people will happen to be black. Removing race can make matters even worse, since it makes us blind to the racial bias behind the number of arrests (64).

The beginning of the book thus sets the stage, outlining how our social problems could be perpetuated and even worsened by AI. By using it, we are not only modelling the world, but changing it – potentially leading to dangerous feedback loops. An algorithm to rank job applicants that is biased in favour of men – because its data was collected in a professional world that is biased – will prefer men in the hiring process, which then further enlarges the gender gap, as the algorithm influences the training data for future iterations (49). As Christian puts it: "Our human, social, and civic dilemmas are becoming technical. And our technical dilemmas are becoming human, social, and civic. Our successes and failures alike in getting these systems to do 'what we want,' it turns out, offer us an unflinching, revelatory mirror." (13).

It is only natural that the second part of the book (*Agency*) tries to understand how agents – biological and artificial – actually learn. While the first part outlines the problems, the second provides the necessary knowledge to understand how we may solve them. The chapters in this part are populated by algorithms trying to drive cars or play complicated video games, which serve as examples to discuss different ways to teach them the behaviour we want them to exhibit.

Christian offers a comprehensible guide in these chapters, which discusses different strategies to teach agents. Can we give them rewards for acting in the desired way, chapter four asks? This question seems straightforward, but it faces problems. More precisely, how do you keep agents motivated through long, complicated tasks, where the reward only waits at the end? Very often, agents give up before reaching their goal – it comes as no surprise, for instance, that PhD students suffer from depression and procrastination, since they have little intermediate rewards but only the promise of their doctorate at the end (179).

To solve this, chapter 5 suggests "shaping", or structuring the environment in a way that encourages the desired behaviour (151). Instead of rewarding a job well done, we reward limited actions that approximate the desired behaviour (154-155). Simply put, if you want to teach a pigeon how to bowl, do not reward it only for moving the ball. A good start may instead be rewarding it

for looking at the bowling ball, at which point you can gradually work your way forward (153).

This avoids depression and procrastination, but is hardly safe from complications. For instance, Christian notes, if we reward the approximation of a desired action, we may encounter 'reward hacking', where agents repeat the rewarded act over and over. The cognitive scientist Tom Higgins recounts in the book how he would praise his daughter for cleaning the floor, until the child emptied the collected dirt on the floor, only to clean it up again (165–166). Thus, we should instead reward a state of affairs – the fact that the floor is clean, rather than the act of cleaning. We reward progress towards the goal and subtract rewards for moving away from it – in this case, dirtying the floor again (169–170).

Chapter 6 tackles another issue: How do we make agents explore things on their own? How do we make our agent interested in cleaning the floor in the first place? Especially with rather complicated tasks, this becomes a key issue.

These excursions into the world of learning form the backdrop for the book's main focus: How can we teach AI the values we want it to have? Part three (*Normativity*) ties the insights of the previous part into the wider theme of the book. It begins with another chapter (chap. 7) on learning, this time by imitation. It quickly becomes apparent why this chapter is located in part three: through imitation, we are now asking the machine to draw its own inferences.

Learning by imitation means that humans tell the machine to "watch me and do as I do." This avoids many of the problems above, such as reward-hacking, but carries its own issues, such as how many data points a machine needs to imitate us in all potential circumstances. For instance, an algorithm that learns how to drive by imitating a human driving in the middle of their lane may make terrible errors once it is not in the middle of the lane (229).

Chapter 8 delves deeper into algorithms drawing their own conclusions. We learn of inverse reinforcement learning, where algorithms observe our actions and infer our goals from that (255). A promising way is to let humans and machines work together, towards a reward that only the human knows in the beginning. Dubbed 'cooperative inverse reinforcement learning', this offers an engaging way to address the alignment problem – not guiding the machine to the right behaviour, but letting it infer it for itself (268-272). A good side effect: humans tend to trust the machine-colleagues more when they work together first (272).

It is in chapter 9 that Christian opens up more frightening issues, starting by recounting the story of the Soviet soldier Stanislav Petrov. In 1983, serving in a Moscow bunker, Petrov received a warning of five incoming American nuclear missiles. The system instructed him to launch a counter-strike. But instead of reporting to his superiors, Petrov started thinking: would the United States not send more than five missiles if they attacked? Luckily, his doubts were legitimate. The warning system had erred and no strike was happening. Thanks to his doubts, humanity potentially avoided nuclear war.

The element which had no doubt in this entire scenario was the system – reporting that the reliability of its assessment was "highest" (277–278). The issue of *Uncertainty* (the title of the chapter) pervades through the alignment problem. Since algorithms do not express epistemic humility, how far should we actually trust them if they are sure about their own assessments? This affects many issues, albeit usually in less dramatic ways than with Petrov. For instance, Christian recounts how an image classifier will tag every image you give it, even if it is random static. Instead of opening up about its inability, or saying that it is unsure, the algorithm will give you a classification with more than 99% confidence (279).

In a different sense, uncertainty affects human agents as well. Specifically, we know that we have no perfect knowledge of the values we want to teach AI – and this is a problem. As the philosopher William MacAskill noted in his famous book *What We Owe the Future (2022)*, it is dangerous to think that we already know the correct moral values. What we see as normal may be completely abhorrent in the future. This entails, of course, a certain danger. Not only do we have to ensure that AI follows our values, but we must first define what these values should even be (306–307).

MacAskill identifies with the long-termist movement, a collective of people who think about ensuring a decent life for humans living in the very far future. MacAskill argues that one of the biggest existential risks to these people – and maybe the most likely one – is being under the spell of the wrong moral values (309). And it is possible that AI could solidify the wrong values, making it more difficult to improve on them. As Christian himself puts it: "We must take caution that we do not find ourselves in a world where our systems do not allow what they cannot imagine – where they, in effect, *enforce* the limits of their own understanding" (327, emphasis in original).

The scale of the alignment problem grows as one reads this book, as its implications and the obstacles to solving it become clearer with every page. Christian tries and mostly succeeds in giving an overview of the problem, while giving the reader enough knowledge on the underlying issues to understand it. There is little for which one could fault the book, except that an additional chapter on the human obstacles to the alignment problem might have been worthwhile. It would have been interesting to explore the human side of the alignment problem more deeply. We read much about incentives for humans and machines to learn, but little about *incentives for humans to teach.* Will autocratic states have a different view on alignment than democracies? Do all firms understand it the same way? Are there incentives for researchers to neglect alignment for the sake of quick deployment?

The book could have used a discussion on how to make humans follow the optimal course for AI alignment. Without this, the book seems to be missing an essential part of AI alignment – which is clearly noticeable in a work that gives such a comprehensive overview otherwise. This is a regrettable state, since one cannot help but ask these questions after having finished reading the book. The first step of the alignment problem is aligning our own ideas about alignment. If the reader is interested in exploring this, they will sadly have to reach for another book after having finished *The Alignment Problem*.

Brian Christian (2020): The Alignment Problem: How Can Artificial Intelligence Learn Human Values. London: Atlantic Books: 476 pages. ISBN: 9781786494313. Price 9,99 € (paperback).