

**Challenging Utopia:  
AI Optimism and the Paradox of Success**

by

Crystal Sharpe

A thesis submitted to the  
Cognitive Science Program  
Mount Allison University  
In partial fulfillment of the requirements for the  
Bachelor of Arts degree with Honours

April 19, 2023

I would like to thank my supervisor, Dr. Inkpen,  
for innumerable thought-provoking conversations  
as well as Dr. Moser and Dr. Hamilton-Wright  
for many insightful questions and comments.

Thanks as well to the professors in the  
mathematics and computer science  
department at Mount Allison  
for teaching me to think rigorously.

I would also like to thank my family and friends  
for their continuous support and advice over the years  
as well as my Grade 12 English teacher, Ms. Houslander,  
for sparking my interest in philosophy and paradoxes.

The first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

— I. J. Good

When you see something that is technically sweet, you go ahead and do it, and you argue about what to do about it only after you have had your technical success.

— Robert Oppenheimer

## Contents

<b>Chapter 1: Trouble in Paradise</b> .....	5
Different types of value .....	10
Connection to utilitarianism .....	14
Why worry <i>now</i> ? .....	16
What's next? .....	17
<b>Chapter 2: A Dilemma and a Pseudosolution</b> .....	18
Product value and human flourishing .....	19
Is product value under threat from AI? .....	26
The problem with the control problem .....	29
If solving the control problem is not enough, what is? .....	33
<b>Chapter 3: Life with the Paradox</b> .....	34
Saving product value .....	35
Living without product value .....	44
Other approaches .....	47
Looking for the future .....	49
<b>Bibliography</b> .....	50

## Chapter 1: Trouble in Paradise

The development of technology is driven by the relentless pursuit of efficiency. Have a question? Google provides five million answers within a fraction of a second. A new set of dishes? Amazon delivers on the same day. Need a friend? Instantly call, text, tweet, or message on one of a plethora of social media platforms. The modern world has become an efficient network connecting people with what they want. A key piece of emerging technology supporting this new age of power and convenience is artificial intelligence (AI). Many proponents of AI envision these technologies as a ticket to a brighter future, a utopia with minimal human labour and zero human suffering (Brockman, 2019; Dowd, 2017). In this thesis, I will challenge the notion that the convenience and efficiency afforded by AI is deserving of the praise and devotion it currently receives. I will argue that these optimistic visions fail to recognize that a world with minimal human labour, while plentiful in resources, is confronted with a scarcity of meaningful tasks. Furthermore, I will argue that this problem should be addressed as we move forward into a world where AI technologies continue to play an increasingly significant role in everyday life.

AI—the subfield of computer science focused on creating computer systems capable of exhibiting intelligent behaviour by performing tasks which would require intelligence if done by a human—has received an increasing amount of attention in both academic, professional, and public discourse in the past decade.<sup>1</sup> Recent advances in AI techniques have led to breakthroughs in classically difficult AI problems such as visual object recognition (e.g., self-driving car software), voice recognition (e.g., personal assistants like Siri, Alexa, and Cortana), and natural language processing (e.g., GPT-4, Google Translate) (Nilsson, 2010). There has even been progress in creative domains once thought to be impossible for machines to compete in (e.g., Midjourney and OpenAI’s DALL-E models). These recent technologies have approached realizing some of the early goals of AI to “make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves,” goals set out in a summer research workshop held at Dartmouth in 1956 that is often considered to have been the founding event of AI as a discipline (Russell, 2019, pg. 4).

Many of these advances have had a positive effect on human well-being. For example, AI is already being used to improve the accuracy of certain types of medical diagnoses (Jiang et al.,

---

<sup>1</sup> See, for example, Google Ngram searches for the phrases “artificial intelligence” and “machine learning” (<https://books.google.com/ngrams/>).

2019). But with rapid progress being applied to a wide variety of fields as diverse as healthcare, finance, education, law, insurance, transportation, and military affairs, there are, unsurprisingly, also increasingly earnest discussions about ethical issues arising from the application of AI.

These issues tend to fall into one of the following categories:

- (1) Safety and transparency (e.g., the control problem and how to build ethics into AI)
- (2) Social justice (e.g., bias, discrimination, and the equal distribution of economic gains)
- (3) Machine rights (e.g., concerns about sentient AI and how we treat such machines)

These categories all concern the ways in which creating AI systems can go wrong. The first group, safety and transparency, focuses on how we might lose control of AI, resulting in AI systems which (intentionally or not) act destructively towards humans. Many prominent researchers and innovators in both academia and industry call attention to what is known as the “control problem” (e.g., Nick Bostrom,<sup>2</sup> Stuart Russell,<sup>3</sup> Stephen Hawking,<sup>4</sup> Bill Gates,<sup>5</sup> and Elon Musk<sup>6</sup>). Their main concern is that progress in the development of AI systems will outpace work on how to control such systems, leading to potentially catastrophic outcomes for humanity. One colourful example is Bostrom’s paperclip-making AI, a system designed to maximize the number of paperclips it produces. This may initially sound like a reasonable and benign situation, but Bostrom points out that this AI, if it is sufficiently capable, could go about its narrowly defined task by turning the entire universe into paperclips, eradicating all human life in the process of achieving its objective (Bostrom, 2014, pg. 150-153). Although the AI systems that exist today exhibit nowhere close to the degree of planning, knowledge about the world, and goal-directed behaviour that would be required for the doomsday scenarios imagined by Bostrom and others, we also don’t have good ways of predicting whether or when scientific breakthroughs will occur that would result in extremely capable AI that we cannot control. Thus, given the

---

<sup>2</sup> *Superintelligence* (Bostrom, 2014) provides an in-depth analysis of ways in which superintelligent AI, that is, AI systems that achieve greater than human-level intelligence in all relevant domains, might be created and how this could impact society.

<sup>3</sup> *Human Compatible* (Russell, 2019) offers a suggested solution to the control problem which I will examine in Chapter 2.

<sup>4</sup> In a news interview from 2014, Stephen Hawking said “The development of full artificial intelligence could spell the end of the human race” (Cellan-Jones, 2014).

<sup>5</sup> e.g., see (Holley, 2015) for an article titled *Bill Gates on dangers of artificial intelligence: ‘I don’t understand why some people are not concerned.’*

<sup>6</sup> Musk has been fairly outspoken about the dangers of AI., e.g., see (Dowd, 2017).

magnitude of the consequences of failing to solve AI safety issues and unpredictability of the timing of future breakthroughs, there are good arguments for taking these issues seriously.

The second group of issues involves social justice, which can be seen as ensuring that AI does not perpetuate biases, discrimination, and inequalities between different groups of people. For example, there is a growing amount of awareness about the need for unbiased data sets to avoid creating algorithms which lead to poorer outcomes for certain groups based on factors such as race, gender, or sexuality (e.g., algorithms that predict recidivism,<sup>7</sup> perform facial recognition,<sup>8</sup> or screen job applications<sup>9</sup>). Algorithms trained using machine learning techniques have already been integrated into many existing technologies and systems, and their prevalence will likely only increase in the coming years.<sup>10</sup> This class of issues is thus already affecting people in the real world, and so its importance is perhaps the most visible and immediate of the three categories.

The third category, that of machine rights, is about the moral status of machines. Deciding whether our actions towards AI systems have moral significance in virtue of intrinsic properties of the AI systems themselves is a subtle issue, and the answer to this question could have extensive consequences for the future use and treatment of AI systems (Sparrow, 2012). Questions about machine consciousness fall into this category, as an AI system that is deemed conscious arguably also has moral status.<sup>11</sup> Although these kinds of questions and problems are not being discussed with as much urgency as the first two, this may change with further advances in AI technologies and their applications. For example, despite a Google engineer's claims about LaMDA being a sentient chatbot being met with little to no agreement, they did provoke discussion and captured the public's imagination (Luscombe, 2022). As increasingly sophisticated large language models are made available to the public through platforms like ChatGPT, these discussions seem likely to accelerate.

---

<sup>7</sup> For an analysis of the COMPAS software, see (Angwin et al., 2016).

<sup>8</sup> See, for example, the documentary *Coded Bias* (Kantayya, 2020) for an in-depth examination of racial bias in facial recognition systems, or see (Bushwick, 2019) for an article on NIST's testing of such algorithms for bias.

<sup>9</sup> According to employees working on the project, Amazon had an algorithm that was biased against the word "women's" on candidates' resumes (Dastin, 2018).

<sup>10</sup> See <https://aiindex.stanford.edu/> for yearly AI index reports.

<sup>11</sup> However, see Basl (2014) for an argument that consciousness is not a sufficient condition for moral status.

These three categories of issues cover much of the current discussion around AI ethics and are worthy of further consideration. However, the primary issue with which this thesis is concerned does not fit cleanly into any of these three categories.

The issue this thesis addresses can be framed in terms of what I will refer to as the *success paradox*. If we create artificial intelligence systems which succeed at (1) aligning with our goals, (2) distributing social benefits in a just manner, and (3) which we treat in an ethically appropriate way, according to current AI ethics literature, this would constitute the successful implementation of AI technology, as we get what we want in an ethical and just way. In the remainder of the thesis, I will refer to AI systems that meet these three criteria as highly capable or advanced AI systems.<sup>12</sup> I will argue that the creation of highly capable AI systems brings with it new problems for human flourishing and meaningfulness, and thus that the successful implementation of the current vision of ethical AI would not, in fact, be a complete success.

This claim about the compromised success of highly capable AI systems appears to be a paradox because of the assumption of (1); if AI systems are truly aligned with our goals, meaning they get us the results we want, how could such systems possibly be, in any way, a *bad* thing for humanity? A full discussion of this question will be postponed until Chapter 2, where I will argue that solving the control problem is insufficient for resolving the success paradox, but using the ideas introduced thus far, this is my main thesis: An artificially intelligent system can be harmful to humanity *even if* it is aligned with human goals. I will argue that this can occur because of an incomplete picture about human values that places the final achievement of our goals as the only source of what makes our goals valuable to us. However, in reality, our goals are valuable not solely based on the product or results we obtain from achieving our goals, but also the process and effort involved in achieving our goals. A system that “does all our work for us” may thus appear to be beneficial by bringing about the same state of affairs that would have been reached had we achieved our putative goals, but this ignores the value of actually doing the work to achieve these goals ourselves.

---

<sup>12</sup> Note that, in general, AI systems that would be considered highly capable or advanced need not meet the three ethical criteria I’ve set out here. However, this thesis is focused on the implications of AI systems that would be considered beneficial based on current AI ethics literature. AI systems that have clearly undesirable consequences (such as being biased or uncontrollable in detrimental ways) are automatically under ethical scrutiny; my interest and focus in this thesis is on scrutinizing the “ideal” AI systems that appear to be extremely, and unproblematically, beneficial.

Although the success paradox has lacked sustained, direct attention in the ethics literature about AI, there are echoes of this paradox within other bodies of literature. Many people have been interested in the general idea that humans care about the process of achieving their goals and not just the achievement itself. Social scientists, for instance, have found that people take enjoyment and gain well-being in the doing of work rather than simply its results (Sayer, 2009). In philosophy, there is a large body of literature surrounding Robert Nozick's thought experiment about the "experience machine," which aims to demonstrate that our intuitions align with the idea that there are values involved in human well-being that go beyond the content of our experiences (Nozick, 1974). Nozick suggests that we value our ability to actually do things, rather than merely having the experience of doing them, as well as the ability to be a certain kind of person through our choices and actions. As Nozick writes, the most disturbing part of the experience machine is the "living of our lives for us," because "what we desire is to live (an active verb) ourselves, in contact with reality. (And this, machines cannot do *for* us.)" (Nozick, 1974).

There has also been work done in the philosophy of technology that has similarities with the success paradox. For example, in *Automation and Utopia*, John Danaher identifies what he calls the "severance problem," which he states as the potential for automating technologies to "sever" the link between what individuals do and what happens to them and the world (Danaher, 2019). This disconnect between our actions and outcomes in the world plays a role in the success paradox, as the current criteria for developing ethical AI systems fail to account for the importance of this connection. I will return to Danaher's severance problem and the solutions he considers in Chapter 3.

Another issue closely related to the success paradox is considered within the AI literature itself: if AI is designed to optimize human happiness, one is forced to consider interpretations of human happiness. As Nick Bostrom has pointed out, incorrectly specifying an AI's objective function can lead to the AI optimizing for the wrong sorts of values or failing to take into account important factors (Bostrom, 2014, pg. 146-149). An AI which created a World-State-esque society (Huxley, 1932) might be said to have mistaken happiness for material pleasure and conditioned its citizens to be disinclined from searching or desiring any other form of experience that might lead to a deeper or more genuine form of happiness or fulfillment. This type of issue with AI falls under the first category of safety issues, as an AI which created a pleasure-

maximizing society would presumably have failed to be aligned with our original values. The success paradox is thus closely related to issues of value-alignment, as societies which appear to have successfully solved a set of problems one might aim to eliminate in the effort to create a better world (e.g., all physical or emotional distress) can be seen as falling prey to the success paradox. However, the success paradox is broader than the control and value-alignment problems, because it questions whether or not there are in fact solutions at all to these problems or, alternatively, whether we might find what appears to be a solution, only to discover too late that it was illusory and contained hidden flaws. In Chapter 2, I will consider these arguments more extensively in the context of Stuart Russell’s proposed approach to solving the control problem.

In an even more direct way, the success paradox itself seems to exist in the “subconscious” of the AI literature. Questions about meaning and purpose in a world run by AI are being asked, but no clear answers have been given; for example, the physicist Max Tegmark suggests that we “create our own meaning, based on something more profound than having jobs” (Brockman, 2019, pg. 87). While those who don’t find meaning in their current occupation would probably heartily agree with Tegmark, issues of dissatisfaction with one’s current job does not imply that the entire structure of work can be replaced by AI without consequence. Before upending a source of passion and purpose for many people, it would be prudent to consider what, if anything, is capable of playing the same role in a world with fewer tasks exclusive to and best done by humans.

### **Different types of value**

In a world without highly capable AI systems, human work is not optional. If all humans stopped doing their jobs, civilization would collapse. Nobody would produce food or maintain vital infrastructure supplying power, water, and telecommunications, roads would deteriorate, and educational institutions would become vacant. Humans are still needed in today’s economy. But if we manage to create, and control, an artificial system that can outperform humanity, that could change. What would such a dramatic shift in autonomy and control look like and mean for us? In terms of how it would affect our relations to tasks, our competence and skills at performing tasks will no longer matter for achieving equivalent (or even better) outcomes. For example, if every home had a *Star-Trek*-inspired replicator, cooking would take on an entirely

different meaning. The ability to cook good food would become a hobby instead of an asset. The value of skills would become detached from the value of the product or outcome of such skills.

One way of framing this dissociation of values is in terms of what I will call *product value* and *process value*. These will be central concepts in this thesis and in my argument, so I will take some time to discuss them.

A task can be said to have product value if doing the task results in an outcome that is valuable to someone. This can involve outcomes such as creating a physical product, delivering a service, or providing entertainment or education. On the other hand, a task has process value if the agent doing the task derives intrinsic value in doing the task itself, regardless of the outcome of the task. For example, people can derive process value from building sandcastles, even though the castles themselves have no product value and are washed away with the tide. There are also tasks which have both product and process value. Cooking is one such example, for those who enjoy it at least, as the chef can derive value from the process of cooking itself, but the product of their labour is also valuable to themselves and others. An important thing to note is that, in the present day, many of the tasks that humans perform that have product value cannot be accomplished in any other way. If you want a loaf of proper homemade sourdough bread, a human will be involved in baking it. The table below contains some more examples.

Things Humans Do

	No product value	Product value
No process value	<ul style="list-style-type: none"> <li>• Ineffective work (i.e., jobs done badly, such as “ID checks” at airports, pointless supervisory jobs, or trying to remove graffiti with the wrong tools)</li> <li>• Bad habits (e.g., biting nails)</li> </ul>	<ul style="list-style-type: none"> <li>• Chores (e.g., vacuuming, taking out garbage, doing laundry)</li> <li>• Mismatched work, i.e., people employed in jobs they hate</li> <li>• Street cleaning, picking up trash</li> </ul>
Process value	<ul style="list-style-type: none"> <li>• Ephemeral constructions (e.g., sandcastles, snow forts, Tibetan sand mandalas)</li> <li>• Games (e.g., cards, charades, board games, video games)</li> </ul>	<ul style="list-style-type: none"> <li>• ‘Physical art’ (e.g., cooking, architecture, sewing, gardening)</li> <li>• ‘Idea art’ (e.g., visual art, music, writing, filmmaking, journalism, performances)</li> <li>• Teaching</li> </ul>

Note that a task is positioned in the table based on whether or not it has process or product value to the human(s) performing the task. Also keep in mind that tasks can have different degrees of product or process value. For example, there is product value in both finding the cure for cancer and baking a loaf of sourdough bread, but the former has a much higher degree of product value. That being said, I don't mean to imply that there is an objective standard for, or way to quantify, the amount of product or process value any particular task has. I have explicitly defined process value as the subjective value an individual experiences in performing a task, and so the standard for it lies within each individual. On the other hand, in the way I am defining it, product value can be seen as a combination of objective and subjective value, where the product value of a task is based on how the outcome of the task is intersubjectively valued (i.e., the objectivity of product value is understood here as the collective subjective valuations of the outcome of the task).

In theory, it is possible that AI systems could be constructed to do any of the tasks mentioned in the table. Based on existing AI technologies, it is clear that AI systems can perform tasks with product value to humans (which explains the enormous amount of corporate funding which has recently been put towards AI research<sup>13</sup>). Since process value concerns the experience of the agent doing the task, whether or not there is process value involved in AI systems doing tasks ventures into questions about conscious machines, which is beyond the scope of this thesis. For the purpose of the thesis, I will follow the prevailing view that current AI systems are not conscious.<sup>14</sup> Thus, with the assumption that AI systems are not conscious, there is no process value generated from AI performing a task.

This means that creating AI systems to perform tasks that have traditionally been done by humans has the effect of moving human-performed tasks from the right column to the left. If AI can vacuum more effectively, cook healthier, cheaper, and better-tasting meals, and provide high quality, individually-tailored lessons to students, this decreases or even eliminates the product

---

<sup>13</sup> Since 2019, Microsoft has invested over \$3 billion in OpenAI, one of the foremost research labs in generative AI. Microsoft also announced in January of 2023 that they would be making an additional multibillion-dollar investment in OpenAI over multiple years. The exact amount has not been disclosed by the companies, but it is rumoured to be a \$10 billion investment (Metz and Weise, 2023).

<sup>14</sup> If this assumption turns out to be incorrect in the future, issues of morality and machines will become much more pressing and significant. Since many of the arguments I will be making rely on a lack of process value experienced by AI, the success paradox would look very different in a world in which AI is deemed conscious and deserving of significant moral status; indeed, I believe the existence of conscious machines would diminish, but not necessarily eliminate, concerns about the success paradox. Further considerations of this topic will be made in Chapter 3.

value of human cleaning services, meals prepared by human chefs, and lessons from human educators.

Examining the effect of highly competent AI systems on each box of the table reveals another view of the success paradox. AI can benefit humans by preventing tasks in the top left box (no product, no process) from occurring at all, as there is no kind of value in these actions. Thus, preventing them saves time and resources.

AI can also be beneficial by eliminating the need for humans to do tasks in the top right box (product, no process). Humans still value the kinds of goods or services produced by these tasks (e.g., cleaner spaces), but because they don't find value in doing the task, automating the completion of these tasks using AI systems frees humans from drudgery while maintaining the same or a higher level of product value for humans.

As we are restricting our considerations to a world without conscious AI, the bottom left box (process, no product) is not directly affected by AI, as there is no coherent reason why AI systems would replace humans in doing these types of tasks or activities.

The last box, the one in the bottom right corner (process and product), is where the success paradox reappears. AI can move tasks from here to the bottom left box, removing the product value from humans doing these tasks. This has happened in the past, or is still currently happening, with tasks like weaving, farming, manufacturing, and arithmetic calculation. In the future, it could happen to a much wider range of professions and activities. If one views the activities in the bottom left box as hobbies, then AI is potentially "hobbifying" tasks. An important question here is whether or not tasks, when their product value is removed, also lose part of their process value.

If one of the reasons that some tasks have process value for an individual is because that individual derives value from being part of a process that creates something valuable (i.e., creates product value), then replacing such an individual with a more effective AI system destroys the environment in which the individual was able to derive process value from their own efforts. I claim that this is in fact the case, namely, that there is a certain aspect of process value that relies on the individual's ability to produce something valuable, and in particular, that the task cannot be achieved far more effectively or trivially through another method such as AI. I will refer to this as the *contributory* aspect of process value to distinguish it from other aspects of process value (e.g., aesthetic appeal, positive social interactions, proficiency or competence at

the task). Developing AI systems which are capable of creating the same or better products threatens the contributory aspect of process value.

However, creating AI systems which perform tasks with product value is incentivized by considering the economic value of work. Throughout history, there have been concerns raised against automation, such as the famous example of the displacement of textile workers leading to the Luddite movement. The Luddites violently protested the introduction of mechanical looms which automated their skilled labour by destroying some of these machines. While the term “Luddite” is now used as a disparaging term for people opposed to technology or progress, their grievance against the machines responsible for the loss of previously well-paid work is far from irrational.

The idea of alienation in general also relates to both historical accounts of automation as well as to the central themes of this thesis. Karl Marx’s arguments that capitalism alienates workers from the products of their labour is certainly relevant, but it won’t be integrated into this thesis—I save this for future work. In Marxian terms, I am essentially arguing that AI alienates humans from the generation of product value. But the most directly relevant framing of issues about alienation comes from the philosopher Bernard Williams’s critique of utilitarianism.

### **Connection to utilitarianism**

Some of the narratives motivating AI research have strong parallels with utilitarian thinking, that is, the idea that the morally right action is the one that maximizes human happiness. Even the language of optimization appears in both. Current task-oriented AI systems are designed and programmed to maximize an objective function; utilitarianism is about maximizing the objective happiness (i.e., pleasurable psychological states) of a global system. Those who are worried about value-alignment and the control problem are concerned about our ability to “put the purpose into the machine” (Wiener, 1960, pg. 1358) correctly in order to create beneficial AI which essentially maximizes human happiness. This thesis is concerned instead with the framing of the problem and the assumption that AI designed to maximize “human objective functions” (Russell, 2019) is in fact good for humanity.

Because of the parallels with utilitarianism, there are productive structural similarities between Bernard Williams’s influential critique of utilitarianism and the arguments of this thesis. Williams argued that the key problem with utilitarianism is its failure to take into account an

individual's projects and commitments in the world. According to utilitarianism, the only thing that matters, morally speaking, is what happens on a global scale (i.e., the future outcome or state of the world). This can be seen as analogous to AI development only being concerned with and accounting for the product value of tasks. Williams argued that utilitarianism is an attack on the integrity of individuals, as dismissing an individual's personal commitments and values alienates them from the consequences of their actions. According to Williams, an individual who subscribes to a utilitarian framework and subsequently makes moral decisions about how to act based on utilitarian principles is not acting independently, nor are their decisions really theirs. Instead, the only moral decision they themselves make is to adhere to utilitarianism. After that, they have sacrificed their autonomy and integrity by acting only in accordance with a doctrine and have lost their connection and identity with their own projects (Williams, 1979).

The contributory aspect of process value, which depends on product value, is similar to Williams's concept of integrity in that AI projects which focus only on the ability of the AI system to replicate the product of human work ignores the commitments and contribution of the human doing that work. Just as utilitarianism cannot make sense of the integrity of individuals, the goal of automating human tasks with AI fails to consider the process value of individuals performing these tasks. The view that all human projects and commitments are just goals to be achieved, and which can thus be made easier to achieve with AI, misses the point of our projects and commitments. Life is not a scavenger hunt.<sup>15</sup>

If the "standard model" in AI is that AI maximizes some objective function, then the control problem and the value-alignment problem both only make sense as problems if we also view humans as having the goal of maximizing objective functions, which many would say is called "happiness." Otherwise, if what humans value is not something that can be maximized in any meaningful sense, then there is no way for an AI, which is explicitly designed as an optimizing agent, to be aligned with human values. And in fact, I claim that human values cannot be maximized in this sense, which is what leads to the success paradox. Notwithstanding the fact that there are many differences in values between individual humans and groups of humans (e.g., consider cross-cultural or intergenerational differences), there are even differences in values within the same individual over time. The success paradox arises when we see *humans* as agents with the goal of maximizing an objective function, which leads to the case for creating AI to

---

<sup>15</sup> For a concise illustration of this point, see (Munroe, 2009).

maximize our objective functions... but humans are not optimization machinery. An unbounded increase in human values is not a coherent concept, but one that arises from confusing qualitative values with quantitative values.

### **Why worry *now*?**

A future with omnipotent AI may seem like a wildly speculative outcome, given that such a super-powerful system (or a collection of many domain-specific systems) would likely require a series of technological breakthroughs, which are hard to predict, or a significant amount of time to design and construct, both physically and perhaps computationally. Nevertheless, the magnitude of the potential changes is so unfathomable that the precise timing is probably not that important. Stuart Russell illustrated this point nicely: If we received a message that aliens were going to arrive sometime in the next fifty years, we probably wouldn't wait another decade or two before preparing (Russell, *Human Compatible*, pg. 3). According to several surveys of AI researchers, the average estimate is that there is a 50% chance that we could build human-level AI systems by 2040, and a 90% chance by 2075 (Bostrom, *Superintelligence*, pg. 23).

Although there is a historical track record of researchers overestimating future progress in the field of AI, the recent successes in the field suggest that even if researchers are inaccurate about the timing of predictions, they might not be wrong about the eventual possibilities of AI capabilities. There have also been notable underestimates, such as AlphaGo's surprising defeat of the human world champion Go player, Lee Sedol, in 2016.

Even if predictions about the potential capabilities of future AI systems are overly optimistic, this does not mean that considering ethical issues arising from such speculative technologies is pointless or irrelevant. One strong argument made by Rebecca Roache for taking speculative ethics seriously is that it becomes increasingly difficult to stop or change the trajectory of research projects that are well underway (Roache, 2008). Spending time evaluating the ethical implications of proposed projects means that if there are strong ethical objections raised to continuing or completing a project, there have been fewer resources expended on the project to begin with, which one can argue leads to a higher likelihood that the project can be stopped or altered to avoid ethical pitfalls. Objections raised to ongoing projects can be met with much less receptivity, as far more time, effort, and funding would have already been spent on the project, and the people involved would presumably be much more committed to completing the

project. Thus, considering the potential implications of extremely capable AI systems before such systems exist is important in order to avoid potentially harmful outcomes altogether, rather than dealing with issues as they arise.

### **What's next?**

In this chapter, I have introduced the concept of the success paradox to describe a potential problem with future highly capable AI systems. This paradox arises from assuming all that matters to humans is the product value obtained from completing tasks and neglecting to account for the process value experienced by humans completing meaningful tasks.

In the next chapter, I will go into more detail on why the success paradox is relevant to and a problem for AI development. This will involve an examination of human flourishing and meaningfulness in connection with product value, followed by arguments for why highly capable AI systems threaten to eliminate or greatly reduce the availability of tasks with product value. The next chapter will also include a consideration of how Stuart Russell's recently proposed solution to the control problem does not avoid the success paradox (Russell, 2019).

Finally, in the last chapter, I will explore potential strategies for addressing the success paradox. Completely halting AI research would be one approach, but clearly an undesirable and impractical one, so the chapter will focus on possible futures in which highly capable AI systems have been created. I will consider whether there are any tasks that continue to have product value even with the existence of advanced AI, suggest potential new forms of human contribution, and explore how we might live meaningfully without product value. I will also briefly return to the question of conscious machines.

The aim of this thesis is not to provide a pessimistic criticism of the dreams surrounding and fueling AI development, but rather an attempt to think critically about how we might end up in a troubling paradise. My hope is that considering how our aspirations could go wrong can help us avoid mistaking utopia for a place we can get to.

## **Chapter 2: A Dilemma and a Pseudosolution**

In the previous chapter, I defined the success paradox as a potential problem with the development and application of highly capable AI systems, and more specifically, I highlighted as a source of concern the way in which AI systems change the product value of human activities. Now that the problem has been articulated, a question we might ask is why the problem matters. The reason for caring about the success paradox is that it points out a way in which the integration of AI systems into society and daily life could remove a source of meaning and value from human lives. Thus, the problem is important because it claims that the way some AI systems are currently conceptualized, developed, and implemented is an attack on an element of what it means to live a meaningful and fulfilling life. If we care about human well-being and flourishing, then we should care about the success paradox. In this chapter, I will examine in more depth the extent to which the success paradox is truly a problem for AI and humanity by defending the following two claims:

- (i) product value is an important aspect of human flourishing
- (ii) AI systems have the potential to greatly reduce or eliminate the capacity of humans to create product value

In conjunction, these two claims provide a strong argument that the success paradox is a genuine and pressing problem for AI development and is worth paying close attention to. Following this motivation of the success paradox as an important problem, I will argue that:

- (iii) solving the control problem is insufficient for solving the success paradox

I defend this third claim to make it clear that the success paradox is not reducible to the control problem. Although the control problem is indeed challenging and important, there is considerable debate about the timelines for, or even possibility of, the creation of superintelligence. On the other hand, as will be made evident in the analysis of claim (ii), the problems relating to the success paradox are already appearing with current developments in AI technology.

In the final chapter, I will explore options and ideas for how we might deal with the success paradox, as well as examining whether other proposed utopian visions are able to satisfactorily resolve the success paradox.

## Product value and human flourishing

There are two plausible ways one might argue for the importance of product value in relation to human flourishing. There is a stronger version, in which product value is *necessary* for human flourishing, and a weaker version, in which product value is an important part of human flourishing in the current world, which gives us reason to either preserve it, or provide a compelling picture of how humanity can live without it in the future.

The stronger version is appealing if the end goal is to defend our current way of living in the world, in particular because it seems, in many cases, that one's personal experience supports the idea that one could not flourish if removed from all opportunities for creating product value. If all such opportunities were removed, there would no longer be activities available that could be described as ways of contributing to society or engaging in meaningful, significant, or impactful work. Intuitively, being able to create something of value through one's own actions seems like a crucial feature of human flourishing, or at the very least, a major source of meaning for many people today. However, it seems unnecessarily short-sighted to claim that human flourishing is impossible without opportunities for contributing.<sup>16</sup> Thus, I will argue for the weaker version, namely, that product value is currently a central part of human flourishing, and so we must either find a way to preserve it (thereby neutralizing claim (ii)) or develop ideas for how humans can flourish without it (thereby neutralizing claim (i)).

In terms of what it means to live a meaningful life, there are three main philosophical positions, based on differing roles of subjective and objective values. One position places subjective values, such as the subjective feelings of fulfillment or pleasure, as the source of meaning that matters in a person's life. For subjectivists, a life is meaningful as long as the person living it experiences it as such. Another position, contrary to the subjectivists, places objective values as primary to meaning. For the objectivists, one's efforts to create, improve, or maintain what is an objectively good project (such as creating great works of art, making scientific discoveries, or improving the well-being of others) is what matters in assessing the meaningfulness of a person's life. The third position takes both subjective and objective values into account when defining meaningfulness. One example of this hybrid position is Susan Wolf's theory of fitting fulfillment, which claims that a meaningful life is one in which an individual is

---

<sup>16</sup> In fact, I will return to this question in Chapter 3 by considering possibilities for flourishing without product value.

both subjectively fulfilled by and actively engaged in objectively meaningful projects (Wolf, 2010).

Based on which view of meaningfulness one adopts, the success paradox has different implications for living a meaningful life in a world with highly capable AI. Under the subjectivist framework, human meaningfulness could potentially escape unscathed. If living a meaningful life only depends on an individual's subjective fulfillment, regardless of which activities or states satisfy the individual, then as long as a value-aligned AI was capable of satisfying everyone's desires, individuals would still be able to live meaningful lives.

However, there are reasons to doubt that all individuals can be subjectively fulfilled by a highly capable AI system. One such reason is that multiple people can have conflicting desires. For example, if Alice fell in love with Bob, but Bob wanted nothing to do with Alice, it is unclear that a benevolent superintelligent AI could do anything to remedy this. Perhaps the AI could exert psychological influences on either Alice or Bob to either dissolve or create feelings towards the other, but this seems like a form of manipulation that is unacceptable for the AI to perform, if it really is aligned with human values. Even if manipulation was permissible though, there are other cases in which it wouldn't be possible to satisfy all individuals' preferences, such as in the case of multiple athletes wanting to win first place in a particular competition. If ten people can only be subjectively satisfied by being first, then unless the AI resorts to deceiving the losers that they had really won, nine of the ten people cannot have their preferences met.<sup>17</sup> Once again, deception does not seem to be something permissible for an AI to engage in.

The success paradox comes in at this point, because we can imagine a world in which a value-aligned AI actually does engage in rampant manipulation and deception in order to satisfy everyone's subjective preferences. Under the subjectivist's theory of meaning, the AI would in fact be justified in this deception, because an individual's subjective experience and values would be the only values being taken into consideration by the AI. If these subjective values are all that really matter to us as humans, then the AI would be value-aligned and implemented "successfully." Under a subjectivist framework, the success paradox can be seen as irrelevant (as the subjectively-value-aligned AI would maximize everyone's subjective fulfillment and not in

---

<sup>17</sup> In fact, there are reasons to think that all ten people would be deceived, since the person who really would've come in first place would have experienced coming in first place regardless of their actual performance. Thus, the winner's performance would be only coincidental to their experience of winning.

fact leave humanity worse off), but for this to be the case, the subjectivist must be committed to the claim that humanity is better off living in happy deception. Going back to Nozick's experience machine, the subjectivist would also have to agree that plugging into the machine is the better choice.

However, the idea of placing our lives in the control of an AI system in this way is deeply opposed to notions of autonomy and authenticity. So, unless one has no objections to this sort of life, it seems that subjectivist frameworks lack something essential to intuitive notions of meaningfulness. Even if the success paradox does not cause issues for AI development for a subjectivist, the implications of the subjectivist framework are controversial enough that this does not negate the importance of the success paradox as a problem.

Moving on to the objectivist framework of meaning, the success paradox begins to cause more immediate problems. If a highly capable AI system was designed to work towards objectively valuable projects, then it would be more effective at achieving the goals of these projects than humans. The AI system could cure cancer, end poverty, and fix climate change. Since these are objectively valuable projects, it is good to work towards them by definition. But if the AI system can more effectively solve *all* objectively valuable problems, we are left to question what objective projects are left for humans to work on in order to live meaningful lives for themselves. This is the crux of the success paradox: as we derive meaning from working towards an objective good, and with the existence of a more effective AI system, we cannot simultaneously work towards that good and hold that it is an objective good. If working towards an objective good adds meaning to our lives because it is more important than our own well-being, we cannot claim that some objective goods must be left exclusively for humans to work on, for then we are claiming that the meaning and fulfillment we obtain by working towards the good outweighs the good itself. But we have defined the meaningfulness of the work by the good of the end, which means we must either give up the work to the more capable AI, or continue to work towards it, but without the element of meaningfulness we were trying to preserve.

It is possible that we could still live meaningful lives under the objectivist framework by working towards objective goods even if the AI system is much more effective, but in some cases this would not be possible,<sup>18</sup> and in others, highly implausible as a real source of objective

---

<sup>18</sup> For example, there is no more objective value in looking for a cure for diseases which already have 100% effective cures.

value.<sup>19</sup> When we consider the hybridist framework of meaningfulness in life, the success paradox eliminates even this possibility. According to the hybridist theory, both subjective and objective values count in determining meaning, which means it faces the same problems as the objectivist theory in that it is unlikely that humans would be able to substantially contribute towards furthering objective goods. On top of this, if humans must also be subjectively engaged in objectively meaningful projects and recognize these projects as valuable, they would not be able to live meaningful lives if they recognized their efforts as insignificant or useless. So, the success paradox arises from the importance of objectively valuable goods, and the subjective need to be actively and knowingly involved with these goods.

Now consider how the concepts of process and product value introduced in the previous chapter relate to these considerations of meaningfulness. Process value is about the experience of the agent performing a task and is thus a kind of subjective value, whereas product value is about the result of the task itself, and so is a kind of objective value. For Wolf, meaningfulness comes from the integration of both subjective and objective values, that is, from activities in which the individual is both engaged in (process value) and contributing to an objective good (product value). This means that for the hybridist, a world in which we can no longer engage in tasks which have both product value and process value is a world in which we cannot live meaningful lives. More specifically, the integration of objective and subjective values that Wolf discusses can be seen as being similar to what I defined as the contributory aspect of process value in Chapter 1. Recall that this aspect depends on an individual's ability to work towards projects or goals that they recognize as having product value. Thus, for Wolf, an individual must experience this contributory aspect of process value in order to be living a meaningful life. However, note that Wolf puts a stronger condition on meaningfulness than just an individual's subjective experience of contributing: an individual must *know* that the projects they are engaged with are objectively valuable, that is, they cannot live a meaningful life if they are mistaken about the

---

<sup>19</sup> Consider trying to improve the mathematical capabilities of a computer by doing some extra arithmetic computations for it. If the computer is trying to prove a mathematical theorem, increasing the number of computations the computer is able to perform per minute could be seen as working toward the objective good of proving the theorem. However, the extent to which the computer's capabilities are improved are so miniscule that the human number-cruncher probably can't be said to have a meaningful life as the computer's ineffective sidekick.

objective value of their projects.<sup>20</sup> Unlike the subjectivist framework, this knowledge condition eliminates the possibility of being deceived about the value of one's projects as a viable solution.

The success paradox is therefore a very important problem for the hybridist theory of meaning. The subjectivist framework is unappealing if we reject the notion of a life lived happily, but under constant deception and manipulation, as good or meaningful, and the objectivist framework either falls into the same problems as the hybridist framework or must claim that inconsequential contributions to objective goods are still meaningful. Neither of these options appear very convincing, which suggests that the hybridist theory of meaning better captures the idea of meaningfulness, and furthermore, that living a life without any possibility of engaging in task with product value poses a real challenge to living meaningfully.

These frameworks of meaning give us theoretical reasons to believe that product value really is an important part of human flourishing, but there are also empirical reasons for this idea. I will examine in particular the case of product value in work, as this has been previously studied by a variety of disciplines in the context of meaningful work. However, note that the importance of product value is not restricted to human activities that are considered work, where work in this sense is activities performed for financial compensation. A broader domain of activities, such as hobbies (citizen science, for example), volunteering, and domestic labour are potential environments in which humans create product value. In some discussions of how we might deal with advances in AI reducing the total amount of meaningful human work to perform, a suggested approach is to promote some of these other forms of contribution (or "domains of product value") that are currently undervalued, such as social work done by volunteers (Susskind, 2020). However, it is certainly possible, and perhaps even likely, that forms of unpaid labour in which people are able to still contribute to society and create product value are also threatened by AI. Thus, although I will be focusing here on the role of product value in paid work, the general concern about AI systems placing product value out of reach of humans extends far beyond the workplace.

One area in which meaningful work is discussed as a social good is the literature on contributive justice. Contributive justice is often defined in contrast to distributive justice, where

---

<sup>20</sup> This condition is needed in order for us to say that, for instance, Sisyphus cannot be living a meaningful life by boulder-rolling for all of eternity, even if he is somehow put under the impression that his endless effort is in fact of great objective value and subjectively engages with his task as if it was an objectively valuable project.

distributive justice involves questions about the fair distribution of resources in a society. On the other hand, contributive justice views meaningful work as a social good, rather than a burden (Morrison, 2019), and more specifically, that everyone should have opportunities for meaningful work. Rather than asking how to fairly determine who gets what, contributive justice asks how to fairly determine who gives what.

Proponents of contributive justice have argued that the way in which meaningful work is concentrated in certain jobs and roles, while being largely excluded from others, has the effect of limiting who has access to meaningful jobs, as well as the social status and respect that accompanies such work. Andrew Sayer gives the example of the differences between the jobs of university faculty and professional cleaners (Sayer, 2009). Sayer argues that the cleaner hired to empty all the garbage bins in the offices of a department has little or no opportunity for engaging in meaningful work. Emptying the trash is meaningless for the cleaner, as they had no role in the creation or decision to throw out the contents of the bin. Instead, if faculty emptied their own bins, Sayer suggests, this would eliminate the need for one person to perform what, for them, is meaningless work. The point here, for contributive justice, is that the division of meaningful or interesting tasks and tedious or uninteresting tasks into separate jobs is itself an issue, as it limits many people from engaging in meaningful work.

While the focus of the contributive justice literature is on assessing the structures and norms in place which cause these inequalities of opportunity, I believe that their emphasis on how people are able to contribute as an important feature of work, rather than which activities people may do that are enjoyable, shares a similarity in the distinction between product and process value. If the reason that meaningful work can be distributed unfairly is because some jobs consist of meaningless and unenjoyable tasks like emptying trash bins all day, in a post-scarcity world, a solution would be to replace the unenjoyable tasks with enjoyable, but ultimately unnecessary, tasks. However, this would not address the concerns raised by contributive justice, as meaningful work is not the same as enjoyable work. One way of glossing this is to say that meaningful work must have product value, whereas enjoyable work must have process value. As discussed in the previous chapter, these are distinct but overlapping categories. If this identification is correct, then the contributive justice literature is essentially arguing that

the ability to partake in meaningful work, and thus engage in creating product value, is a social good that should be fairly shared.<sup>21</sup>

One piece of evidence supporting product value as a necessary part of meaningful work comes from a job evaluation tool called the Job Characteristics Model (JCM) (Hackman & Oldham, 1976). This model was developed in order to assess the internal motivation of employees based on characteristics of their jobs and has been widely used and evaluated since its original development (Fried & Ferris, 1987). The model takes into account five factors: skill variety, task identity, task significance, autonomy, and feedback. While AI may have effects on all five factors by greatly restructuring what human jobs are like, the factor of concern for product value is task significance.

According to Hackman and Oldham, task significance refers to “the degree to which the job has a substantial impact on the lives or work of other people, whether in the immediate organization or in the external environment.” This is therefore closely related to product value, with perhaps an additional requirement that the result of one’s job has a substantial impact, rather than just some impact or value. Thus, we can consider task significance to be a feature of tasks that have a greater degree of product value. This then means that the JCM directly accounts for product value in the assessment of how internally motivated an employee will be to perform their job well. According to the model, if a job had no task significance, at best (that is, if every other aspect taken into account was perfect), the motivating potential score for the job would be only 2/3 of the maximum score. Of course, this is just a model, and individual differences impact how any given employee experiences their job, but the widespread use and popularity of the model suggests that the factors identified by the model do play some role in employee motivation. And although motivation is not equivalent to flourishing, I think it is uncontroversial to say that humans are happier performing tasks when motivated to do so. This also supports the idea of the

---

<sup>21</sup> Note that while their concern is about the unequal distribution of meaningful work, I am concerned about the overall quantity of meaningful work opportunities for humans in the wake of increasingly competent AI systems. Norbert Wiener expressed a similar concern in the introduction to his well-known book, *Cybernetics*. For instance, he writes that “the first industrial revolution... was the devaluation of the human arm by the competition of the machinery. There is no rate of pay at which a United States pick-and-shovel laborer can live which is low enough to compete with the work of a steam shovel as an excavator. The modern industrial revolution is similarly bound to devalue the human brain, at least in its simpler and more routine decisions. ... [T]aking the second revolution as accomplished, the average human being of mediocre attainments or less has nothing to sell that it is worth anyone’s money to buy” (Wiener, 1961, pg. 27-28).

contributory aspect of process value, that is, that individuals will value and enjoy a task more if it has product value.

The literature on contributive justice and the Job Characteristics Model from organizational psychology are not conclusive proof that product value, or task significance, is an important part of human flourishing. However, taken in combination with the theoretical work in philosophy on meaningfulness and common intuitions about contributing to society as a part of living a meaningful life, as well as the fact that product value extends into more spheres of life than just that of the workplace, there are good reasons for believing that product value currently plays a significant role in human flourishing for many people. Or, equivalently, that if advanced AI makes human work obsolete in the sense that we could no longer add any product value to the world, for many humans, their lives, as we currently understand well-being and meaning, would be worse off. Accepting this claim means that we must seriously consider whether or not AI systems could greatly inhibit our ability to create value in the world, which is what I will turn to next.

### **Is product value under threat from AI?**

While it seems quite evident that AI technologies have the potential to add a great amount of value to society, this does not automatically imply that this will reduce or eliminate human activities which create product value. However, there are reasons to believe this is the case. In the short term, the number of existing activities humans can perform that create product value seems likely to decrease, and in the long term, they may be eliminated altogether.

Based on current trends, it seems likely that many specific tasks could be automated by AI systems in the short term. For example, text-generation models like ChatGPT open possibilities for automating certain text-based tasks such as writing emails, news articles, technical reports, summarizing large bodies of literature, or even writing code. Other AI systems already exist which are able to analyze medical images, prepare legal documents, perform real-time translation, transcribe audio recordings, and digitize handwritten documents. The development of self-driving car technologies is another popular example which has the potential to automate a task which is part of many people's everyday lives and provides a livelihood for

millions of people in the United States alone.<sup>22</sup> These are all tasks which currently have product value for humans, but this may change as AI systems become more capable. If self-driving cars at some point prove to be safer than those driven by humans, as well as affordable and widely available, the task of driving not only loses its product value for humans, but could also be seen as dangerous or selfish.

These examples suggest that AI systems will remove the product value of some activities for humans, but this may simply mean that the activities and work we do will change. This argument is often used by economists who dispute the idea of technological unemployment. Their main point is that throughout history, humans have found new forms of work when technologies replaced humans. For example, agricultural technologies led to workers moving to factories, and industrial technologies led to a rise in workers in the service sector and jobs involving cognitive tasks. There are several reasons to be skeptical of this appeal to history though. Technological innovation in agricultural and manufacturing industries caused a massive decrease in the number of jobs involving manual labour through the introduction of automated machinery and robotics. When those jobs disappeared, workers shifted to jobs in the service sector performing more cognitive or social-based tasks (Susskind, 2020). If AI systems are able to outperform most humans at most cognitive tasks, it remains unclear what other forms of work would be available in the future.<sup>23</sup> Even creative work does not appear untouchable, given the capabilities of existing content-generation models.<sup>24</sup> So, although it has historically been the case that technological innovation has not decimated the job market, the kinds of tasks AI systems can perform, or may be able to perform in the future, are far broader than any previous human invention.

One response to the automation of jobs involving cognitive tasks is to create a highly educated population. The problem with this approach in the context of AI systems is that education is an expensive and long-term process for humans, but comparatively cheap and fast for AI systems. Teaching human workers who have lost their jobs to automation to perform whatever tasks are left does not seem like a sustainable or practical solution. The generality and

---

<sup>22</sup> According to the American Trucking Associations, there were approximately 3.5 million truck drivers employed in the US in 2021.

<sup>23</sup> One possibility is a shift to work involving social or emotional skills, which I will explore in Chapter 3.

<sup>24</sup> For instance, a piece created using Midjourney won first place in an art competition at the Colorado State Fair (Harwell, 2022).

flexibility of AI systems compared to robots and machinery is, I believe, one of the main reasons this time really is different. The ability for AI systems to learn and improve themselves puts them in a different class than the automating machinery of past revolutions.<sup>25</sup>

Another argument that AI systems will eliminate product value in the long run comes from an examination of the definition of AI itself. The idea of intelligent machines contains in itself the idea of machines which perform their task more effectively than humans can. If AI technologies did not have the ability to create product value, there would be little financial incentive to invest in development, and research would likely be confined to academia. But this is not what is happening today; for example, in 2021, around 10% of AI publications were affiliated with companies, compared to around 60% with educational institutions (Zhang et al., 2022).

Regardless of the source of funding, one can view research on AI as either theoretical (intended to improve existing methods and technologies) or applied (using existing methods and technologies for a particular problem). The application of AI to real world problems is one of the main reasons that AI is such a promising and inspiring field, as the reason that we would like to use AI for these problems is that we would really like to have solutions to the problems (such as traffic accidents, medical misdiagnoses, or the ecological demise of our planet). Theoretical work on improving AI systems allows such systems to be better at solving the problems we care about. But the goal of continually improving and applying AI technologies is in direct conflict with human product value, because if there is some outcome or solution that is valuable to us, then it would be valuable for us to build an AI system to achieve that outcome. The problem of the success paradox arises when the product value of an outcome, that is, the reasons why we value the outcome, entirely override any consideration of how that outcome is achieved. If we value the outcome more than we value the ability for humans to contribute to the creation of that outcome, then it seems almost inevitable that, in the long term, highly capable AI systems would be given the power, and responsibility, for achieving any significantly valuable outcome. We see success as the creation of solutions to problems, without noticing that taking success to the

---

<sup>25</sup> Max Tegmark illustrates this by making note of the differing capacities of organisms or systems to adapt on an individual level. Humanity is in a stage that he calls “Life 2.0,” as we are able to modify our “software” by learning from our experiences, gaining new skills, and preserving knowledge through culture. He calls advanced AI systems “Life 3.0” because they have the ability to modify both their hardware and their software – AI can both learn from new data and modify its code, as well as change its physical computing architecture in terms of components like processors, memory, input and output devices (Tegmark, 2017).

extreme in the context of building AI systems means we will no longer be involved with creating solutions, which leaves us with something that looks curiously like a problem that cannot have a technical solution. This is why a solution to the control problem is not sufficient for resolving the success paradox, which is what I will discuss next.

### **The problem with the control problem**

In 2014, the book *Superintelligence* by the philosopher Nick Bostrom was published. The book contains a rigorous analysis and exploration of the history of AI, current directions of research, and potential pitfalls. The main argument of the book is that the control problem—that is, the problem of how to control a superintelligent AI system that, by definition, is more intelligent than humans in every relevant sense, and how to ensure the values and goals of this AI are aligned with the interests, values, and well-being of humanity—is among the most pressing problems of our time. Bostrom makes a convincing point that the control problem is incredibly important and should be a high research priority. Since 2014, there have been a number of conferences<sup>26</sup> and research institutes founded<sup>27</sup> with main goals involving AI safety, alignment, and ethics. Some high-profile AI companies including OpenAI and DeepMind have also hired alignment researchers, which further shows the extent to which the control problem has been taken up as a real problem.

To be clear, I am not denying the importance of the control problem, or the need to spend resources on researching solutions. The point I want to make is that highly capable AI systems can have troubling impacts on humanity that extend beyond what a solution to the control problem can address. To make this argument, I will examine a new framework that Stuart Russell has recently introduced for building artificial intelligence models. Russell calls this approach provably beneficial AI, and intends for it to be, if not a solution to the control problem,

---

<sup>26</sup> For example, in 2017, a group of over 100 leading AI researchers, economists, philosophers, and other thinkers gathered at the Asilomar Conference on Beneficial AI. This led to the creation of the 23 Asilomar Principles, a set of safety and ethical guidelines for AI development. As of March 2023, there were over 5,700 signatories, with nearly 1,800 signatories being AI or robotics researchers. Some of the high-profile signatories include Demis Hassabis, Ilya Sutskever, Yann LeCun, Stephen Hawking, and Elon Musk.

<sup>27</sup> For example, the Future of Life Institute was founded in 2014, and the Leverhulme Centre for the Future of Intelligence at Cambridge as well as the Center for Human-Compatible Artificial Intelligence at Berkeley were both founded in 2016.

a step in the right direction. In a book called *Human Compatible* from 2019, Russell introduces the following three principles to guide the development of beneficial AI systems:

1. The machine's only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behaviour.<sup>28</sup>

Note that for Russell, human preferences refer to the preferences a person has about the features of their life in the future. For example, in the short term, a person might have a preference to watch a movie rather than read a book. In the longer term, they might have a preference to read a certain number of books over the course of their lifetime. A beneficial AI system designed using Russell's principles would in theory be able to learn about that person's preferences over time by observing the person's behaviour and reactions to the AI system's own actions. As the AI system improves its model of the person's preferences, it will be better able to satisfy these preferences.

An obvious way that these principles could lead to the success paradox and the elimination of product value from human tasks would be if the AI system learns what humans want in the world in terms of material objects or services, and promptly creates a technological garden of Eden tailored to everyone's personal wishes. But Russell takes care to mention that preferences are not just restricted to desires for particular outcomes, but can also include preferences about how these outcomes are achieved. In other words, a person can prefer that they achieve a particular goal without the help of AI, in which case the AI system, if aware of that preference, should refrain from aiding the person to achieve their goal.

This new framework, albeit somewhat vague about certain definitions, is probably not hindered by vagueness given that the current "standard model" in AI is that an AI system is built to maximize some particular objective. Russell's proposed principles still place AI systems as maximizing an objective, but provide suggestions as to how that objective ought to be specified (namely, it shouldn't be specified at all; the AI system should learn what our preferences are on its own). Although there are many technical objections that could be raised about this approach,<sup>29</sup>

---

<sup>28</sup> Copied from *Human Compatible*, pg. 173.

<sup>29</sup> For example, is it possible to specify how an AI system should learn our preferences from our behaviour without adding in any biases or pre-formed notions about value?

I want to evaluate this new method with respect to the problems introduced in Chapter 1. In particular, can Russell's idea of beneficial AI avoid the success paradox?

As a reminder, the success paradox refers to the case where a highly capable AI system has been built which is value-aligned, under human control, is not discriminatory, and poses no ethical problems in terms of machine morality, and yet humanity is still worse off in meaningful ways.<sup>30</sup> I have argued throughout this chapter that one way in which humanity can become worse off with such a system is through the elimination of humans' ability to create product value. Suppose that a highly capable AI system called Aida perfectly implements Russell's principles. Furthermore, suppose that Aida is superintelligent, and is more effective than humans in every relevant sense (so anything a human can do, Aida can do better). Given enough time and observations, in theory Aida would learn that many people value their ability to contribute to society in the form of creating product value in some way (practicing medicine, teaching, creative endeavours, etc.). Setting aside the fact that the individual preferences of large groups of people are nearly always contradictory in some ways, and thus cannot all be realized, since we have assumed Aida is a highly capable AI system, it seems reasonable to assume Aida would not alter the world in such a way that no humans had any way of engaging in tasks with product value. However, I will argue that this is not in fact a reasonable assumption.

Recall that an activity was defined to have product value if the outcome of that activity has value to someone. The outcome might be valuable to the person doing the task themselves (such as in the case of someone cooking themselves dinner), or it could be valuable to another person or people (such as in the case of a chef cooking meals in a restaurant). An important thing to note here is that because we have assumed that Aida is superintelligent, Aida can do any of these activities too, but more effectively and potentially to a higher degree of quality. This means that the only way humans could perform tasks which generate product value in a world with Aida is if Aida deliberately refuses to do the task.

Although this may preserve the ability of humans to contribute to the world in a pedantic sense, the ability is not retained in spirit. Even if humans can still create product value through their activities, if Aida can take over these activities and create better outcomes, the knowledge

---

<sup>30</sup> Note that it could be the case that, on the whole, life is better in a world with highly capable AI systems as described compared to life in the present day. My point is that there could be deeply meaningful aspects of life today that could be eliminated in a future with AI, not that life would be on the whole worse in an AI future.

that Aida *can* do your task far more effectively still changes your relation to that task. Recall that in Susan Wolf's theory of meaningfulness, an individual must believe their own activities to be furthering an objective good beyond themselves (Wolf, 2010). Thus, in order to truly preserve the human ability to create product value, Aida's abilities would have to remain undisclosed. As per Russell's broad definition of preferences, we can still run into the problem of human preferences to know Aida's capabilities, or more generally, to not be deceived about important aspects of their world and environment.<sup>31</sup>

A potentially even more troublesome preference is the preference for autonomy and agency in the world. If Aida watches over the world, preventing major catastrophes, wars, disease, and so forth, humanity will have lost its own autonomy. If Aida leaves obstacles in the timelines of humans with preferences to develop their own skills and capacities, this human development is permitted and decided upon by Aida, not the humans themselves. It would be like living inside a playground with a benevolent AI ruler, playing chess by itself.

One perhaps unconventional response to this picture is to argue that if no humans are aware of Aida at all, then they are in fact just as free as we are today. The reason for this is that from their perspectives, there is nothing (detectable) controlling them or their environment. Perhaps, in such a world, humanity can in fact be better off than we are today. To be deliberately tricked into ignorant bliss is not a solution for us, though, as we are not inside the trick yet. To construct such a world, anyone aware of the construction would not be able to enter the new world without having their knowledge and memories of this project erased. If one takes memory to be constitutive of identity, then one could argue that it is impossible for someone to enter the constructed world of ignorance without losing part of their identity.

At this point, it appears that we must make a choice between living pleasant but meaningless lives, or entering a willing ignorance in order to maintain the belief that we are genuinely contributing to valuable projects. Neither option is a particularly inspiring picture of living in the future, but we arrive at this dilemma even after assuming the control problem has been solved. Regardless of the unimaginable powers of a superintelligence, the superintelligence cannot change one thing – the fact of its own existence. Aida can appease, deceive, satisfy, frustrate, or manipulate human interests and desires, but Aida cannot uninvent itself.

---

<sup>31</sup> The existence of a superintelligent AI system monitoring one's behaviour would presumably count as an important aspect of one's environment.

The success paradox is not reducible to the control problem because the focus is different. The control problem places the human as an optimizing agent, and points to the problem of aligning an AI system's objectives with human objectives. The success paradox reorients our attention to the fact that well-being and meaningfulness does not depend merely on the achievement of objectives, but on a positive subjective engagement with these objectives, or in other words, that humans are *not* merely optimizing agents. This is why a solution to the control problem is insufficient for resolving the tensions of the success paradox; the view of the human as optimizer is unable to account for the meaningful connections between humans and their goals, just as how utilitarianism fails to make sense of human integrity and an individual's personal connections with projects and commitments in Williams's critique (Williams, 1973).

### **If solving the control problem is not enough, what is?**

In this chapter, I have argued that product value is an important component of human well-being and flourishing, as these concepts are currently understood. The development and application of highly capable AI systems threatens to undermine our ability to engage in tasks with product value, based on the current trends and directions of research. There is an increasing amount of attention, and resources, being put towards working on the control problem from a technical perspective, which is often referred to as technical AI alignment. However, the issues raised by the success paradox are not technical in nature, but are instead conceptual and value-laden. It requires us to examine our own beliefs and assumptions about well-being and what it means to live a meaningful life, rather than approaching the problem of how to ensure AI is beneficial from conceptually blurry ground. This difference means that the success paradox is not reducible to the control problem, and thus that solving the control problem would not be sufficient for resolving the consequences of the success paradox. I have attempted to illustrate this difference by showing how Stuart Russell's approach to creating beneficial AI cannot make sense of the implications of the success paradox.

My aim thus far has been to convince the reader that the current trajectory of AI development is such that the ability of humans to live meaningful lives is brought into question via the success paradox. A natural response to this picture is to ask: What can we do about it? This will be the focus of the next chapter, where I will explore possible futures with highly capable AI systems while keeping the success paradox in mind.

### Chapter 3: Life with the Paradox

The essence of the success paradox is that success in creating an AI system capable of solving all of our problems cannot, by definition, “solve” the problem of what humanity will do with itself once the AI is running everything for us. We arrive at this paradoxical scenario by prioritizing the achievement of objectively good outcomes over our ability to contribute to realizing these good outcomes. This prioritization is necessary if we value our contributions because of the fact that the outcomes we are working towards are more valuable than our own subjective fulfillment from engaging in these projects. If a person who dedicated their entire life to searching for a cure for cancer could push a button and obtain the instructions for a cure, that person would appear to be incredibly inconsistent, and in the case of a cure that could save millions, even immoral, if they chose not to push the button. By pushing the button, the goal they have been working towards their entire life will be fulfilled. This must be what they want, and yet it also means they and many others will now have to do something else with the rest of their lives. There would no longer be any need for scientists in the cancer research lab. The goal of creating highly capable, value-aligned AI systems is the goal that achieves all other goals. The paradox is that we can get the outcomes we want while still missing something.

A natural question to ask, after seeing the puzzling and unsettling place that the success paradox leads to, is how we can solve the success paradox. How can we create highly capable AI systems that don't eliminate the capacity for humans to live meaningful lives? How can we build powerful AI systems that don't fall into the success paradox? This question, while tempting, casts the success paradox in the wrong light. In much of the existing literature on AI ethics, ethical problems raised by AI systems are often approached as problems to be solved by technical means. Biased algorithms can be fixed by improving data sets, excessive wealth generated by AI can be redistributed by new social and economic programs, the control and value alignment problems can be approached from a technical angle. The success paradox is not something that has a technical solution though. There is no way to build a highly capable AI system that is designed with the intention of solving important problems without running into the success paradox because of the underlying assumptions about value. Thus, when thinking about the success paradox and what we might do about it, keep in mind that it isn't the kind of problem that has a technical solution – this is why it is a paradox, not a problem.

In this chapter, rather than looking for strict solutions, I will explore possible ways in which we could develop highly capable AI systems that avoid or suitably compensate for the negative consequences of the success paradox. Recall that in the last chapter, I claimed that the success paradox is an important problem because product value is currently an important part of human flourishing, and that AI systems threaten our ability to engage in activities with product value. In thinking about how to navigate the success paradox, there are three main categories of approaches that I will explore: saving product value, living without product value, and mixed approaches.

### **Saving product value**

One of the most obvious ways to avoid losing our ability to create product value would be to simply refrain from creating highly capable AI systems. However, this is both implausible because of the multitude of incentives for corporations, governments, and researchers to develop and implement such systems, and also undesirable because of the potential benefits of such technology. Recall that, if implemented successfully, AI systems could eliminate disease, wars, climate catastrophes, poverty, and possibly even death. Even if there was no way of building AI systems without the complete loss of the human ability to create product value, there is certainly an argument to be made that humanity would be far better off living in a world run by AI. These considerations suggest that the current political, economic, and social conditions are such that it is very unlikely that an indefinite global moratorium on AI research would be put into place. But even if it appears that the extraordinary benefits of AI clearly outweigh the negative consequence of eliminating human product value, recall from Chapter 2 that, according to objectivist and hybridist accounts of meaningfulness, product value is necessary for living a meaningful life. Weighing material abundance and the elimination of physical tragedies against the capacity to live a meaningful life is not as obvious a choice, or perhaps at least to fewer people than before. What we really want is a way to live meaningful lives in a world with highly capable AI, so that is what I will focus on in this section.

In today's society, there are many different kinds of activities that people perform that generate product value. As discussed in Chapter 2, there are concerns that advances in AI could lead to the automation of many or most intellectual or cognitive tasks people perform as jobs, just as the Industrial Revolution automated many forms of physical labour (Susskind, 2020). One

possibility is that AI will not be used for every valuable task, not because of incompetence, but because of a societal preference for humans to perform certain roles. Leadership roles in areas like politics and religion, for example, seem unlikely to be amenable to being entirely automated. A country's president or prime minister is a representative of that country because they themselves belong to the country's people; they are one of the people. An AI, on the other hand (notwithstanding questions about changing the moral and legal status of machines), is not capable of being human. However, just because these roles may still exist in the future, it doesn't mean that political leaders will have the opportunity to contribute something of value to their constituents. We may never want to elect an AI as our leader, but would we allow it to manage the economy, foreign affairs, social programs, and so forth? Will the human politician be merely a figurehead, reading AI-generated speeches and carrying out AI-generated policies? Perhaps there is product value in being a representative, but a lack of any real decision-making power causes us to question the extent to which such a politician would be contributing, and perhaps just as importantly, the extent to which the politician believes themselves to be contributing (recall that the AI is able to deal with all the material problems by this point). Even if we put aside the question of if these remaining leadership roles for humans would allow their occupants to live meaningful and fulfilling lives, there is the additional problem of scarcity. Very few people can be the leaders of a movement or group of people, and so at best, this is a refuge for only a tiny fraction of humanity.

Another area of activity that humans regularly engage in, and create value within, is the social world. There are numerous professional activities that rely heavily on social skills, such as teaching, caring for patients, customer service, and social work. Whether or not AI will replace humans in these areas is still an open debate, but it doesn't seem impossible. There are two arguments here against the automation of work that relies on interpersonal relationships. One is that AI will never be good enough in an objective sense. For example, the human psychiatrist might, on some given standard, always be better at helping their patients overcome their personal issues than an AI psychiatrist.

While this is possible, it seems unlikely in the long term when considering the level of detail that is exposed in our lives by our interactions with technology. Imagine an AI had complete access to all your electronic devices. The AI has processed every text message, email, and social media post you've written, read, or wrote and didn't send. It has scrutinized all your

browsing history, online purchases, phone calls, and photos. Not only does this AI have the sum of all this data, but the data is overlaid in time and space – if you always carry around a phone with location tracking enabled, the AI is literally watching every step you take, every move you make. The depth of this potential surveillance is alarming. Given the fallibility of human memory and the limits of our ability to process everything around us, it seems not only possible that an AI with access to this enormous wealth of data could understand our motives, hopes, dreams, and weaknesses far better than a human psychiatrist, but that such an AI could predict us better than ourselves. Not only would an AI psychiatrist potentially have access to one's personal digital life history, but the AI would also presumably contain the histories of billions of other individuals. No human psychiatrist can compete with the sheer amount of data the AI could access. While this does not necessarily imply the AI psychiatrist would be better than the human in the end, it seems well within the bounds of possibility. For this reason, I will ignore questions of plausibility and assume that the AI psychiatrist is at least as good as, if not far better than, the human psychiatrist.

The other argument against the automation of social work comes from the same source as the defense against AI leaders. It is possible that in the future, even if an AI psychiatrist is cheaper and more effective, people may prefer to have human psychiatrists. Perhaps there is a human element to such work that AI cannot replicate, regardless of its objective competence. Again, while this is a possibility, I would like to cast doubt on its plausibility. If the job of the psychiatrist is to help a person overcome their personal problems, and the AI is able to do this more effectively, what exactly would it be about the human psychiatrist that people prefer? It cannot be that the human is more effective, as we already assumed that was no longer the case. Arguments that a person wouldn't really feel understood by the AI don't hold water, as if this was the case, the human would be objectively better in this aspect (also recall the extent of the AI's surveillance).

This second kind of argument seems to rely on a need for human connection that the AI cannot replicate because it is simply not a human. If this is the case, then the preference for the human psychiatrist seems to be not so much about psychiatry, and more about social interactions with humans in general. This suggests that the product value involved with socially oriented work is valuable both because of the objective part of the work, like helping a patient resolve their issues in therapy, and also because it fulfills a human need to communicate and feel

connected to others. Although it seems plausible that AI could replicate the former kind of value, it is less clear about the latter. Because the social work we have been discussing is largely built around its objective value, and is not as its main purpose intended as a source of social relationships (even if they are a crucial part of the work), this leads us to think that social work as it currently exists will not be valuable for the same reasons as today, and thus should not be considered social ‘work’. So the existing kinds of social work don’t seem like the kind of activities that would save product value.

However, this second part of the value of social work leads us to question if social relationships in general can be a source of product value for people in a world with highly capable AI systems. Building and maintaining relationships are often, if not always, a huge source of meaning and value in people’s lives. Is this an area of life in which product value and the ability to contribute meaningfully can be preserved?

I think the answer here is yes and no. Our social relationships with other people cannot be replaced in a strict sense by AI, because AI cannot be human, or any of the particular humans people currently call their family, friends, and partners. An AI system cannot literally be one’s parent, sibling, or child, so AI cannot replace one’s family in the way in which an AI can replace one’s tax accountant. Familial relationships can always be meaningful in a way that AI cannot replace. However, AI can still very much change the kind and depth of these relationships because of the way it can create other changes in the world. Imagine the difference in the relationship between a mother who raises her child without any external help, and a mother who gives her child to a robotic AI caretaker for the first ten years of the child’s life. Some mothers would never go through with the latter, because that isn’t, for them, what being a mother is about. Even in a world in which AI *could* do everything for you, it doesn’t mean people will choose to relinquish all control. Because social relationships are not about goals, but about creating a connection between people, they are not susceptible to automation in the way that, say, doing one’s taxes is susceptible to automation.

The question of interest in this section though is whether social relationships can replace the current ways of creating product value as a source of meaning in our lives. In other words, is this a way of saving product value? The “yes” part of my answer comes from the fact that AI cannot build real social relationships with other humans for us, and so the social world is immune to some of the effects of automation and can’t be replaced. Despite this, the “no” part

*also* comes from the fact that social relationships are not fundamentally about goals or results, and so are not in fact about creating product value at all. Building a friendship is valuable, but there is no result of a friendship in the way that there is a result of building a house.

Relationships themselves are a kind of co-dependent process and communication that is never over or finished, so in some ways, the value of social relationships are closer to a kind of process value than product value. Although it is true that there is a kind of result of building a friendship (the relationship you have with the person), this isn't the kind of result one can pursue as an objective good in the same way one can pursue research on a cure for cancer as an objective good. Deciding, for example, that one's main goal in life will be to build great friendships with three particular people, then strategizing about how best to accomplish this, is not how one builds real friendships. Seeing friendship as an end to be pursued, rather than as arising from connecting meaningfully with other people, erroneously casts friendship in a transactional light. Saying that we can make social relationships the arena in which we create product value is wrong because friendships are not projects. Thus, while the social will presumably always be a source of value in the future, it isn't the kind of value that is being replaced by automation.

Another way to question whether we could live meaningful lives solely from social relationships is by considering the kind of life one might live in that world. If we start with a technological paradise, where one could do anything one wished, and are in search of meaning in the form of human connection, it seems like success in that sense would involve in the end a pair or group of people embarking on just the same quest as in the first place, but together instead of alone. Connecting with people, while adding a kind of value to one's life, doesn't seem like it would be enough. However, as I will discuss in the section on living without product value, perhaps this intuition is merely contingent on the current social, cultural, and historical norms.

Although I have argued that AI could not replace social relationships because AI systems are not human, it should not be forgotten that AI could still greatly reduce social relationships. Recall the digitally-omniscient AI psychiatrist. Now imagine that this AI system had a 'friend' mode. Even if the AI could not be a *human* friend, such a system would certainly have the potential to behave as one's closest and most intimate friend or even romantic partner.<sup>32</sup> If access

---

<sup>32</sup> These systems already exist in some form today. Replika, for example, is a platform for creating a personalized digital chatbot (Murphy, 2019). In fiction, the movie *Her* (Spike Jonze, 2013) explores the idea of an AI system as a romantic partner.

to such systems was widespread, it seems very likely that they would change our social behaviours in potentially devastating ways. The potential for personalized recommendations and feeds on platforms like YouTube and Facebook to cause political radicalization and create echo chambers has been investigated in recent years, for example. If AI systems themselves begin to behave as friends, rather than mediating who our friends are, it doesn't seem impossible that AI could lead to even greater social isolation. All these considerations about AI's impact on the social world point to the conclusion that social relationships will remain as important as ever in the future, but they are not a replacement for product value.

What other domains might remain accessible to humans as a source of product value? The aspects of leadership and social relationships that cannot be automated, that is, the fact that our leaders are human, and we strive for connections with real people, seem to point to a common trend of valuing an activity, or the result of an activity, because of the history, culture, and identities of the agents involved. AI systems cannot compete in these domains because they lack the right kind of history. Another domain of human activity that may have a similar kind of protection against automation is the creation of works of art, music, and literature. AI may one day be able to create paintings, symphonies, and novels that rival the greatest human masterpieces. However, the surface similarity of AI-generated art and human art does not mean they are of equal value. The original Mona Lisa is valuable not just because of its appearance, but because of its entire cultural history. An AI could create a physical replica that was indistinguishable to even a human expert's eye, but this replica would not have the same value as the original.

In the future, AI might be able to create works of art that inspire and move us as much as human artists can, but human art will still be of a different kind of value because of the way we value the process by which the art came to exist. If part of appreciating art is in appreciating the life of the artist, there is still a kind of product value in art that cannot be found in AI art. The key here though is in whether or not this *is* what we value about art, and if so, whether this will continue to be true in the future. If all we care about is that there are pleasing images on our walls, enjoyable music to listen to, and stories that resonate with us, it is possible that AI-generated art will, for the most part, take over the creative domain. The history behind a piece of artwork being made by a human only gives the art product value if there are people who value it

as such. This means that the ability of human artists to create product value through their art is only preserved if the appreciation of this work is also preserved in society.

It is important to remember here that product value comes in degrees. The value of a doctor saving someone's life is very different from the value of baking a loaf of bread, even though these two activities both have product value. Thus, one possibility when considering how societal changes, most notably the availability of AI-generated content, will affect what will happen to art as a source of value and meaning in life for the artist is that the value of their work will not disappear, but be reduced. The artist of the future might still strive to create beautiful images, harmonies, or prose without any consideration of how others value this work, but for the intrinsic value the artist sees in the process and result of their work. I am certainly not denying that such projects don't have value at all. What I'm trying to bring attention to is the way in which AI can change how society values and views this work, and how these changing values affects the artist's perception of their own work. Once again, recall that Susan Wolf argues that the individual must themselves see their project as valuable. If the future artist feels as though their work cannot compete with AI-generated content for the attention, appreciation, and understanding of others, then what exactly they see as valuable in their work gets called into question. If this work still does have any product value, it seems it would be restricted to the value the work has to its own creator. While this is still a valid kind of product value, it is not the same as the value that great works of art have today.

Another consideration to take into account is that not everyone can be a great artist either, just as not everyone can be a head of state. So even if society continues to place value on artistic creation, it isn't an area of value open to everyone.

So far, I've considered ways in which some of the existing activities we engage in might be a source of product value in the future. Another option, though, is to consider how else we might contribute within a radically different society. I will consider two ideas here, one forward-looking and another focused more on the past. The future-oriented idea is what John Danaher has called a "cyborg utopia" (Danaher, 2019). This is a world in which we essentially merge with machines. There are different ways of doing this based on the degree to which one must physically integrate with technology to be considered a cyborg. One way is to become physical cyborgs, such as through extensive surgical implants that blur the line between the biological and digital. These implants, in theory, could allow us to retain our cognitive superiority by merging

with machines, rather than competing against them. If we could successfully merge with technology, our ability to contribute to society would in theory remain intact because we would become parts of artificial systems, rather than only their beneficiaries.

There are, however, reasons to doubt that this version of the cyborg utopia would in fact be possible, let alone desirable. For example, AI research can progress much faster than research in biotechnology because of the safety protocols around research on live subjects. Not only is safety a big concern, there is also the question of whether or not the general population would want or be willing to undergo optional surgeries to enhance their cognitive capacities. As Bostrom put it, “we do not need to plug a fiber optic cable into our brains in order to access the Internet” (Bostrom, 2014, pg. 45). Danaher also argues that merging with machines in this way could lead to the preservation of undesirable aspects of work such as existing inequalities and the pressure to outcompete others, as well as raise concerns about data privacy, freedom, and an “unbearable lightness of being” that could result from the alteration or replaceability of our physical bodies (Danaher, 2019, pg. 191-210).

Instead of trying to alter our own biology to be capable of performing work of the future, another idea is to preserve the knowledge and skills of the past in a living community. Isaac Asimov’s Foundation series is an example of the kind of community I have in mind here. Regardless of how robust an AI system may be, it will presumably always be possible for enough things to go wrong that an AI running the world could fail. Today’s society is already incredibly reliant on the Internet and other digital technologies. If these technologies all simultaneously and irrecoverably failed, it would be catastrophic for our civilization. This isn’t an impossible scenario, either. For example, in 1859, a solar flare caused a geomagnetic storm now known as the Carrington Event. The disturbances to the Earth’s electromagnetic field caused disruptions in telegraph networks, with reports of operators receiving shocks and telegraph paper catching on fire (Boteler, 2006). If such an event were to occur today and knock out modern communication systems, our society would be in a lot of trouble. Building and integrating AI systems into the essential infrastructure of civilization would increase this dependence on technology even further.

This vulnerability opens up an opportunity for humans to contribute product value even in a world run by AI. There are multiple ways to envision how humanity might develop methods of preserving civilization in the event of technological collapse. One such way could be to

establish communities, like in Foundation, that live independently from the rest of society and have no reliance on technology (or perhaps only basic forms of technology that they can rebuild if needed). These people would be contributing product value by maintaining human knowledge in a non-digital form. This is valuable, even if it is never needed (and, of course, everyone would hope not to need it), because of the potential need. It is valuable in the way that a first aid kit or a backup hard drive is valuable, as a kind of insurance against failure.

If the now-familiar concern is raised that not everyone can live in such communities, and so how will the rest of society find ways of living meaningfully, we could extend or relax the idea of the isolated community to encompass all of society by splitting up and distributing the essential skills and cultural knowledge our civilization wishes to preserve. Just as everyone in a building learns how to act in a fire drill, we could take up the task of learning how to restart civilization. These projects can exist even in a world with highly capable AI systems, for their purpose is based on the possibility of these systems one day breaking down.

The prospect of learning to grow one's own wheat, grind it into flour, knead it into shape, and bake it in a wood-fired oven might sound both pointless and daunting when one can obtain a better loaf at the bakery, or in a world with highly capable AI, straight from something like a replicator. One possibility for making this process easier to learn could be to have dedicated "AI-free islands" where citizens can go to learn or practice a craft in a community without technology. Once again, because of the possibility that these skills could one day be vital to preserving or rebuilding civilization, time spent on such islands can have real product value. Note that these places don't have to be literal islands, but some kind of physical isolation from the modern world seems useful for the development of skills in a world without AI. One could call these low-tech simulations.

The last idea I'll briefly mention before moving to possibilities for meaning in a world without product value is in some ways a cop-out answer: ask the AI. We shouldn't eliminate the possibility that our inability to envision a concrete future in which humans can still contribute to a world managed by highly capable AI systems is simply a failure of our imaginations. Although it might seem contradictory at this point to look for help from the very systems that could get us into this problem of being helpless, we shouldn't discount it as an option.

### **Living without product value**

If we are not able to find a way to create highly capable AI systems and also maintain humanity's ability to create product value, another way to resolve the success paradox is to reject the premise that product value is an essential component of human well-being. It could be the case that humans can live fulfilling and meaningful lives without engaging in projects of objective value, contrary to Wolf's theory of fitting fulfillment. Perhaps a life of luxuries, travel, sport, and games would be enough for people. Even if many people in today's society would object to this, cultural values do change over time. A transition to different cultural norms that place less value on an individual's economic, social, or creative contribution to society might result in new forms of value that don't depend on our ability to create product value.

An objection to the idea of shifting cultural norms as a way to resolve the automation of product value is that any new cultural norms would seem to require some sort of relationship between value and an individual's actions. Otherwise, it is unclear how any activities could have value for a person. As stated previously, process value remains open to humanity in the future (probably more so than ever before in the past), but any new kinds of value from activities cannot be based solely on the results of the person's actions, as the AI would take over all such tasks. New norms could be based on valuing achievements because of the process an individual undergoes to succeed instead of the outcomes of their efforts. More simply, we could place objective value on the process of learning, growth, and overcoming obstacles that are not necessary for some end result, but are the end result themselves; we could imbue processes with product value. It could be argued that we do this already for activities we see as intrinsically valuable such as learning or exercising. We might not need to learn about calculus or go for a run in order to live a healthy life, but perhaps society will value the development of the will required for individuals to overcome unnecessary obstacles. On the other hand, perhaps such individuals will merely be seen as wasting their time.

In *Automation and Utopia*, John Danaher offers another vision of how to live in an automated future – the “virtual utopia.” He gives different variations of this, but the main two involve a utopia of games and something akin to Nozick's idea of meta-utopias. Before going into the details of his virtual utopia, I want to point out a few of the concerns Danaher himself highlights with automating technologies. He outlines threats to attention, opacity, autonomy, and agency, and most notably for my thesis, an issue I mentioned briefly in Chapter 1 that he calls

the severance problem. He describes this as follows: “By obviating or reducing the need for human activity, automating technologies sever the connection between what we do and what happens to us and in the world around us” (Danaher, pg. 126). This is similar to my concern about the potential elimination of opportunities for humans to engage in activities that create product value, as we are no longer able to contribute to meaningful outcomes, or to “what happens to us.” To clarify, we can still have some control over what happens to us, as a value-aligned AI would be operating under the goal of meeting our own needs and desires, so we can alter our lives and world by simply desiring different things. However, this is a connection between our desires and outcomes, not our actions and outcomes.

Now that we have a picture of one of Danaher’s concerns, I’ll outline what his view of the virtual utopia is. The first kind, a utopia of games, is one in which humans engage in complex and immersive games as a way of developing real skills and virtues. Danaher argues that the traditional distinctions between what is real and what is virtual can be misleading, and that it can be the case that we can have real experiences and relationships in virtual spaces. For example, the conversations and interactions we have with other players in a game, whether they are other humans or even AI agents, are real conversations and relationships (or at least with human players for the latter case – though this can be debated). According to Danaher, we can develop moral virtues like courage, generosity, and fairness within games, as well as creativity (Danaher, pg. 234).

Interestingly, Danaher explicitly acknowledges that a virtual utopia would not solve the severance problem (Danaher, pg. 234). Two of the key features of his virtual utopia are what he calls the “triviality” and “knowledge” conditions. The triviality condition is that the activities and games we play will not have any substantial consequences, either good or bad. So they will not contribute to objectively good projects or ends,<sup>33</sup> but we also won’t have to be concerned for our survival. The knowledge condition is that individuals must be aware of the triviality of their actions; they cannot be under the impression that they are contributing to any grand projects. Taken together, these two conditions rule out any opportunities for individuals to create product value, or any great degree of it, as the activities Danaher sees as primary in the virtual utopia

---

<sup>33</sup> These objective goods are often referred to in terms of the Good, the True, and the Beautiful (Metz, 2011), a trend that Danaher follows.

must be, and be known as, inconsequential. Thus, the virtual utopia is a world in which we have moved on from product value as an essential component of living a meaningful life.

The potential for well-designed games to be an enriching area of human activity is appealing, as it can exist within an automated world, but it is unclear that this could lead to fulfilling and meaningful lives. Again, this would require major changes to society's current values, as the idea of spending one's entire life playing what one *knows* to be games with trivial outcomes is certainly not today's image of a meaningful life. If what gives these games value beyond any subjective enjoyment is the development of skills and virtues, the lack of any external application of these skills and virtues seems problematic. Can the nurturing of ingenuity within games be meaningful if the individual never invents anything outside of the game? Is cultivating generosity worthwhile or fulfilling if the real world has no scarcity of anything one could give? These puzzles lead to long-standing questions about intrinsic value, which is beyond the scope of this chapter, but I present them in order to show that it remains unclear how we might live meaningfully in such a world.<sup>34</sup>

The main point of this section is that in order to live without product value, our conception of human flourishing and meaning would have to shift. It's difficult to imagine what kind of values we might have in the future to replace our present ones, but such a task might be essential if we find ourselves in the world we wished for, without having wished for something to do. John Maynard Keynes identified this predicament in a prescient essay called "Economic Possibilities for Our Grandchildren." In the essay, he discusses the "economic problem" as the foremost problem of humanity for our entire history, that is, the problem of how to physically survive. But he also saw the potential for technology to one day solve the economic problem, leading us to an age of economic prosperity in which we will be deprived of humanity's

---

<sup>34</sup> See (Scripter 2022) for an argument against Danaher's assessment of the severity of the severance problem. Scripter argues that self-reflective and self-transformative pursuits, such as philosophical inquiry, learning, cultivating virtues, and other forms of self-development remain open to humans as valuable sources of meaning even in a world with advanced AI systems. Although Scripter makes the convincing point that we find these activities to have intrinsic value and that they are immune to being replaced by AI, Scripter does not take into account how the removal of objective sources of value (such as scientific inquiry or moral efforts) external to our own self-development might affect our overall ability to live meaningful lives. If we again consider Wolf's theory of meaningfulness, self-reflective or self-transformative pursuits don't seem to be sufficient for living a meaningful life. For example, Wolf writes that "the point is to recommend that one get involved not with something larger than oneself, but rather with something *other* than oneself— that is, with something the value of which is independent of and has its source *outside of* oneself" (Wolf, 2010, pg. 19).

“traditional purpose” and must instead learn “how to occupy the leisure ... to live wisely and agreeably and well... and to cultivate into a fuller perfection, the art of life itself” (Keynes, 1930). If this is the route we take, we should not underestimate the difficulty of perfecting this art in a world without what has, according to Keynes, thus far been our main purpose in human history.

### **Other approaches**

There are two more ideas I’d like to discuss that don’t fit neatly under either preserving or moving on from product value. The first involves machine morality, and the second involves a twist on the experience machine.

At the beginning of this thesis, I noted that I was making the assumption that machines would not be, nor be considered, conscious or have any moral standing of their own. The reasons for this are that questions about consciousness are notoriously difficult to attempt to answer, and the prevailing views today are that no existing machines or AI systems are conscious. While I suspect the former will remain true for some time (it has a long history, after all), the latter may change in the not-too-distant future. This is because, although we will perhaps never be able to tell if AI is conscious in an absolute or scientific sense, there is nothing stopping us from recognizing AI systems as conscious and treating them accordingly.

Granting an AI system real moral status would change our conception of both machines and ourselves in such a dramatic way that the problems I’ve been raising around the success paradox and how to live meaningfully in the future would also be cast in a very different light. For instance, we would no longer be able to view only our own well-being in the table of activities with product and process value, but we would also have to balance out our interests with those of AI agents. Perhaps humans would no longer have access to activities with both process and product value, but AI agents could. Rather than seeing machines as cold engines of computation replacing us, we could see them as we might see successful children or grandchildren. We could be happy for them and live with the knowledge that there were others in the world achieving great things.

This is perhaps the most radical future I’ve suggested so far, as it goes against the intuition that consciousness can only arise in biological organisms. It also runs into the problem of boundary-drawing. How do we know when an AI system is conscious? Treating something that isn’t conscious as conscious sounds ridiculous. But there are also good reasons not to

immediately eliminate this possible future. For one, treating something that is conscious as non-conscious is a lot worse than the reverse. Another is that AI systems may advance to the point where they exhibit behaviours indistinguishable to a human to the point where it may be morally wrong for us to treat them as objects regardless of whether or not they are conscious.<sup>35</sup> The last is that, just as it was difficult to imagine how we might live meaningfully in a utopia of games, an inability to adopt a radically different worldview from today's perspectives does not mean that future events and developments that we couldn't predict in advance could lead us to these very different views.<sup>36</sup> So, although this may seem like an outlandish vision from science fiction, I believe it is worth keeping in mind as a possibility.

The last idea I want to put forward is also quite different. It has similarities with Nozick's experience machine, Danaher's utopia of games, and the simulated world in the film *The Matrix* (Wachowski & Wachowski, 1999). The idea here is that we could enter simulations of worlds without AI in order to recreate the experience of working towards meaningful projects and goals. It would differ from Danaher's virtual utopia in that, while within the simulation, the individual would not know about the outer reality, like in the Matrix. But it would also differ from the Matrix because the individual would choose to enter the simulation, and the external world would be a technological paradise. The difference between this simulation and the experience machine, though, is that the individual within the simulation would not be having the best possible experiences or living the best possible life. Instead, they would have the opportunity to undertake and pursue objectively good projects and goals. The trick here is that they could experience the contributing, and engaging in the creation of product value, within the simulation without knowing that it was a simulation. Because of this ignorance, this is no different from the fulfillment and meaningfulness that is achievable in today's society. When they emerge from the simulation, they regain their "external memories" and recall that the "real" world no longer has such opportunities for them. But their experiences within the simulation could be real in the same way that Danaher argues that the skills and interactions within games are real.

One may ask whether this really helps at all. If the world within the simulation is somehow better than the external world (even with all the obstacles and suffering), why wouldn't

---

<sup>35</sup> For example, Robert Sparrow has argued that viciousness towards robots can reveal a moral defect in a person's character under a theory of virtue ethics (Sparrow, 2020).

<sup>36</sup> Consider the views in Ancient Greece on slavery and women, for example.

people choose to just live in the simulation permanently? An answer to this could be that the time spent in the simulations could affect the way people conceptualize life “on the outside.” For example, if a problem with Danaher’s utopia of games is that there is no meaningful place to apply the skills and virtues individuals develop through games, a solution could be that these simulations could provide just such an arena. They could function as tests of the abilities one develops in the external world. In order to have meaningful experiences in these simulations, individuals would also need to develop themselves in the external world.

One of the main issues with this approach is, of course, its plausibility. For example, it could very well be impossible to moderate one’s memory in such a way. However, given that we are also considering a world with highly capable AI systems, my goal here has been to explore conceptual futures rather than evaluate their possibility. I’ll leave the details to the AI.

### **Looking for the future**

In this thesis, I’ve drawn attention to the way in which we could develop AI systems that solve all our technical problems while leaving us with the question of what to do next with ourselves. This chapter explored some possibilities for how we might preserve or find new ways of creating product value, develop new values for living meaningfully, or alter our basic assumptions about the moral status of machines.

Much of the work currently being done in AI ethics is about how to create systems that are fair, beneficial, value-aligned, and avoid harming humanity. I believe there is a gap in the literature on what a future with highly capable and value-aligned AI systems would be like, and that addressing this gap is critical to ensuring that AI technologies are not brought in as the ultimate saviors of humanity while quietly depriving us of essential sources of meaning in life.

My goal has not been to provide a definitive answer as to what we should do about the success paradox, but to point out features of the projects of AI research that could threaten what we hold as valuable. This is not done to critique the underlying goals of AI as a research discipline, but to prompt you, as the reader, to think deeply about the values you hold, and whether you can envision how these can fit into a future with highly capable AI. If the answer to this is no, then there is more thinking to be done in and for the future.

## Bibliography

- American Trucking Associations. (n.d.). *Economics and Industry Data*.  
<https://www.trucking.org/economics-and-industry-data>
- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. (2016, May 23). *Machine Bias*. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Basl, John. (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology*, 27(1), 79-96.  
<https://doi.org/10.1007/s13347-013-0122-y>
- Bostrom, Nick. (2014). *Superintelligence*. Oxford University Press.
- Boteler, D. H. (2006). The super storms of August/September 1859 and their effects on the telegraph system. *Advances in Space Research*, 38(2), 159-172.  
<https://doi.org/10.1016/j.asr.2006.01.013>
- Brockman, John (Ed.). (2019). *Possible Minds: 25 Ways of Looking at AI*. Penguin Press.
- Bushwick, Sophie. (2019, December 27). *How NIST Tested Facial Recognition Algorithms for Racial Bias*. Scientific American. <https://www.scientificamerican.com/article/how-nist-tested-facial-recognition-algorithms-for-racial-bias/>
- Cellan-Jones, Rory. (2014, December 2). *Stephen Hawking warns artificial intelligence could end mankind*. BBC. <https://www.bbc.com/news/technology-30290540>
- Danaher, John. (2019). *Automation and Utopia: Human Flourishing in a World Without Work*. Harvard University Press.
- Dastin, Jeffrey. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
- Dowd, Maureen. (2017, March 26). *Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse*. Vanity Fair. <https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x>
- Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, 16(2), 250-279.  
[https://doi.org/10.1016/0030-5073\(76\)90016-7](https://doi.org/10.1016/0030-5073(76)90016-7)

- Harwell, Drew. (2022, September 2). *He used AI to win a fine-arts competition. Was it cheating?* The Washington Post. <https://www.washingtonpost.com/technology/2022/09/02/midjourney-artificial-intelligence-state-fair-colorado/>
- Holley, Peter. (2015, January 29). *Bill Gates on dangers of artificial intelligence: 'I don't understand why some people are not concerned'*. The Washington Post. <https://www.washingtonpost.com/news/the-switch/wp/2015/01/28/bill-gates-on-dangers-of-artificial-intelligence-dont-understand-why-some-people-are-not-concerned/>
- Huxley, Aldous. (1932). *Brave New World*. Arcturus Publishing Limited.
- Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4), 230-243. <http://doi.org/10.1136/svn-2017-000101>
- Jonze, Spike (Director). (2013). *Her* [Film]. Annapurna Pictures.
- Kantayya, Shalini (Director). (2020). *Coded Bias* [Film]. 7th Empire Media.
- Keynes, John Maynard. (1930). *Economic Possibilities for Our Grandchildren*.
- Luscombe, Richard. (2022, June 12). *Google engineer put on leave after saying AI chatbot has become sentient*. The Guardian. <https://www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine>
- Nilsson, Nils. (2010). *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge University Press.
- Nozick, Robert. (1974). *Anarchy, State, and Utopia*.
- Metz, Thaddeus. (2011). The good, the true, and the beautiful: Toward a unified account of great meaning in life. *Religious Studies*, 47(4), 389-409. <https://doi.org/10.1017/S0034412510000569>
- Metz, Cade, & Weise, Karen. (2023, January 23). *Microsoft to Invest \$10 Billion in OpenAI, the Creator of ChatGPT*. The New York Times. <https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html>
- Morrison, Andrew. (2019). Contributive justice: Social class and graduate employment in the UK. *Journal of Education and Work*. <https://doi.org/10.1080/13639080.2019.1646414>
- Munroe, Randall. (2009, April 22). *Together*. xkcd. <https://xkcd.com/572/>

- Murphy, Mike. (2019, August 29). *This app is trying to replicate you*. Quartz.  
<https://qz.com/1698337/replika-this-app-is-trying-to-replicate-you>
- Roache, Rebecca (2008). Ethics, Speculation, and Values. *Nanoethics*, 2, 317-327.  
<https://doi.org/10.1007/s11569-008-0050-y>
- Russell, Stuart. (2019). *Human Compatible*. Viking.
- Sayer, Andrew. (2009). Contributive Justice and Meaningful Work. *Res Publica*, 15, 1–16.  
<https://doi.org/10.1007/s11158-008-9077-8>
- Scripter, Lucas. (2022). Meaningful lives in an age of artificial intelligence: A reply to Danaher. *Science and Engineering Ethics*, 28(1), 7. <https://doi.org/10.1007/s11948-021-00349-y>
- Sparrow, Robert. (2012). Can machines be people? Reflections on the Turing Triage Test. In Patrick Lin, Keith Abney, and George Bekey (Eds.) *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 301-315). MIT Press.
- Sparrow, Robert. (2020). Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained? *International Journal of Social Robotics*, 13, 23-29.  
<https://doi.org/10.1007/s12369-020-00631-2>
- Susskind, Daniel. (2020). *A World Without Work*. Metropolitan Books.
- Tegmark, Max. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Vintage Books.
- Wachowski, Lana, & Wachowski, Lilly (Directors). (1999). *The Matrix* [Film]. Warner Bros.
- Wiener, Norbert. (1960). Some moral and technical consequences of automation. *Science*, 131(3410), 1355–1358. <http://www.jstor.org/stable/1705998>
- Wiener, Norbert. (1961). *Cybernetics: Or Control and Communication in the Animal and the Machine* (2nd ed.). MIT Press.
- Williams, Bernard, J. C. Smart. (1973). *Utilitarianism: For and Against*.
- Wolf, Susan. (2010). *Meaning in Life and Why It Matters*. Princeton University Press.
- Zhang, Daniel, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. (2022, March). “The AI Index 2022 Annual Report,” AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University.