Human Achievement and Artificial Intelligence

Brett Karlan
Purdue University

Penultimate draft of a paper to appear in Ethics and Information Technology. *Please cite the published version if possible.*

Abstract

In domains as disparate as playing Go and predicting the structure of proteins, artificial intelligence (AI) technologies have begun to perform at levels beyond which any humans can achieve. Does this fact represent something lamentable? Does superhuman AI performance somehow undermine the value of human achievements in these areas? Go grandmaster Lee Sedol suggested as much when he announced his retirement from professional Go, blaming the advances of Go-playing programs like AlphaGo for sapping his will to play the game at a high level. In this paper, I attempt to make sense of Sedol's lament. I consider a number of ways that the existence of superhuman-performing AI technologies could undermine the value of human achievements. I argue there is very little in the nature of the technology itself that warrants such despair. (Compare: does the existence of a fighter jet undermine the value of being the fastest human sprinter?) But I also argue there are several more localized domains where these technologies threaten to displace human beings from being able to achieve valuable things at all. This is a particular worry for those in unequal societies, I argue, given the difficulty of many achievements and the corresponding amount of resources needed to achieve great things.

7561 words

1. Introduction: Sedol's Lament

In March 2016, the DeepMind algorithm AlphaGo defeated Go grandmaster Lee Sedol 4-1 in a best-of-five series.[1] The victory was considered a breakthrough for the development of artificial intelligence (AI) technologies in a domain that was previously the sole purview of human experts. Unlike the route tree searches and expert hand-coding used by Deep Blue to defeat chess grandmaster Gary Kasparov in 1996 (Hsu 2002), AlphaGo utilized state-of-the-art machine learning techniques (including more complicated Monte Carlo tree searches) to select its own strategies for play (Silver et al. 2016). These methods allowed the play of AlphaGo to at least appear more intelligent, and it more closely mirrored how human experts make similar

[1] See Lee et al. (2016) for a contemporaneous report.

decisions.[2] That Go is significantly more computationally complex than chess made AlphaGo's dominant victory all the more surprising.

The debate over whether AlphaGo and similar programs are truly rational or intelligent has dominated discussions since its victory. Halina (2021), for instance, presents a fascinating discussion of the idea that AlphaGo should be considered "creative" in its capacities. Halina argues that current machine learning algorithms represent a middle ground in our understanding of creativity. These models possess some (e.g. the ability to build and evaluate cognitive scenarios) but not all (e.g. domain generality) of the properties we normally associate with creative intelligence. Relatedly, Buckner (2019) presents a model of minimal rational inference and information-processing on which the play of AlphaGo is practically rational. It is not my intent to say much about the intelligence or rationality of machine learning algorithms here. I instead want to focus on a tangle of interesting but underexplored philosophical questions that arise from reflecting on Sedol's own reaction to his loss.

In 2019, Sedol retired from professional Go. He specifically tied his retirement to the growing dominance of AlphaGo:

> With the debut of AI in Go games, I've realized that I'm not at the top even if I become the number one through frantic efforts… Even if I become the number one, there is an entity that cannot be defeated.[3]

[2] More recent work has attempted to replicate superhuman performance at multiple games without hand-coding expertise of any kind (even the rules of the game) into the algorithm (Silver et al. 2018).
[3] Yonhap News Agency (2019). As the article mentions, there are also political reasons why Sedol might have announced his retirement from Korean Go. But whatever the true motivation for Sedol's retirement, the sentiments he expressed latch onto a real concern about the future of human achievement in an era of superhuman AI performance. It is this concern that motivates the remainder of this paper, not Sedol's actual motivations.

Sedol's pessimistic feelings about the future of human achievement in areas with current (or soon-to-be) superhuman performance by AI technologies seemed to resonate with many. A contemporaneous report of his retirement, for instance, made a similar point, using the language of accomplishment and achievement to frame the worry:

> Se-Dol's final bow in professional Go signals a more significant, existential concern. If a world champion, floating at the peak of personal *achievement*, starts to view *human accomplishment* and *machine accomplishment* as one and the same, it creates an environment for frustration, disappointment, and perceived loss of purpose. Se-Dol sits at the edge of this realization, but all of us are not far behind (Pranab 2019, emphasis added).

This concern deserves sustained normative reflection for several reasons. First, while the feeling seems plausible to many, there is little published philosophical reflection on the issue.[4] Many humans spend vast sums of time and resources to develop their skills in areas that now, or soon will be, the purview of superhuman AI performance. Is all of this investment a waste of time? Sedol's lament seems to suggest it is. If there are genuine reasons to be worried about superhuman AI performance undermining the value of human achievement, it would be useful to have an account of those reasons, and a framework to avoid negative outcomes wherever possible. If, however, the worries are not as pressing as they first seem, it would be good to know this as well.

Framing Sedol's lament in terms of the value of *achievement* also offers theorists the opportunity to connect work on value in technology ethics with the burgeoning literature on the

[4] The broader question of how to "align" the values of AI technologies with human interest is a rapidly-expanding field of research (Peterson 2019; Gabriel 2020), but there has been little published reflection connecting these concerns to the value of achievement in particular (though see Danaher & Nyholm 2020, discussed at length below).

nature and value of achievement in normative ethics.[5] This potential interaction between fields of study could be useful in both directions. For the theorist of achievement, AlphaGo represents an interesting test case. Sedol's lament, furthermore, may be expanded to many different kinds of human achievement, not merely competitive games like Go (including, as I argue below, scientific achievements). For the theorist in technology ethics, on the other hand, the language of achievement offers a way to precisify possible interpretations of Sedol's lament, and a general framework for dealing with cases similar to it. The AlphaGo case represents an opportunity for a fruitful dialogue between normative and applied ethics.

In this paper, I engage in this dialogue. I ask how we can best understand Sedol's lament, and what concerns like it can tell us about the threat AI technologies pose to practices we value. I present several different ways of developing the lament, and show that a clear understanding of the nature and value achievement tends to temper strong readings of it. I also argue, however, that the worry is not without merit in certain local cases. The rise of superhuman-performing AI technologies threatens to exacerbate existing inequalities in human achievement, making it easier for those already well-off to perform well in high-achieving areas, while making it even more difficult for those with less existing advantages to do so. I argue, in summary, that the future of human achievement in the age of AI is mixed. Though there is not much in the nature of algorithms like AlphaGo themselves that threatens achievements we value, there is serious harm that such advances make more likely (though not inevitable) when they are deployed against a background of inequality.

[5] Some examples of this recent work, many of which are discussed below, include Bradford (2015), Hirji (2019), von Kriegstein (2017), Hurka (2020), and Wang (2021).

2. Conceptual Preamble: the Nature and Value of Achievement

Before considering several ways of making Sedol's lament more precise, it will be helpful to have the basics of the theory of achievement on the table. What makes some activities, like scaling Mt. Everest, discovering a cure to a debilitating disease, or playing an excellent game of Go, achievements? And what makes some superficially similar activities, like taking a helicopter to the summit of Mt. Everest, not count as achievements? According to Bradford (2015)'s influential account,[6] achievements can be thought of as a process culminating in an outcome. In some cases, it is more natural for us to refer to the outcome itself as an achievement. The scientist who finds the cure for cancer has produced a great achievement, for instance. For other cases of achievement, however, a purely outcome-based approach will not suffice. The stock example of scaling a tall mountain (Bradford 2015, p. 12) does not seem to naturally be described as merely reaching a tall peak (otherwise, why not just take a helicopter?). Rather, it is the process of getting to the end state that represents an achievement. Achievements, then, involve both process and product without being reducible to either alone.[7]

Bradford adds two further commitments to a metaphysics of achievement, though they are controversial (see section 4 below). Achievements, for Bradford, must be *difficult*, and it must be *competently caused*. For our purposes, we will only need to focus on the nature and value of

[6] Though Bradford's account is controversial, the basic metaphysics and axiology of achievement (which she presents very clearly) are all we need to get on the table at this moment. If there are aspects of the view we need to modify in light of our reflection on AI performance, we can do so below.

[7] In focusing on Bradford's account, I am setting aside a dense assortment of theoretical questions concerning achievement. For one thing, Bradford (and authors who respond to her) take achievements to generate value irrespective of whether they contribute to the welfare of the achieving agent. One might, instead, understand an achievement as primarily being good for an agent's welfare (Scanlon 1998; Portmore 2008). There are complicated questions as to how welfare-based and intrinsic-value-based accounts of achievement might interact. While these debates are fascinating, the agent-neutral form of achievement seems to be what is most at stake with worries like Sedol's, so we shall focus on it here.

difficulty in achievement.[8] Bradford thinks that difficulty is part of what makes certain activities achievements. Proving some complicated mathematical result is an achievement, while calculating the sum of 3 and 5 is not, because of the difficulty of the former.[9] Bradford argues that achievements gain much of their value from their difficulty. Consider two novelists who each create a novel of similar aesthetic value, but where one novelist overcomes great personal and professional difficulty to do so, while the other produces theirs easily (Bradford 2015, p. 88). There is, according to Bradford, something more valuable about the novel that was produced after overcoming a great obstacle when compared to one (of similar quality) that was easily written. The only difference between the two cases is their difficulty. It is thus in virtue of overcoming more difficulty that the former achievement gains more value.

How can the theory of achievement help us illuminate the AlphaGo case? The playing of high-level Go as a significant human achievement. Developing the skills necessary to play Go at a 9-dan level is a process that results in particular outcomes (the playing of Go matches) where the process is both difficult and competently caused by the agent. What about the value of playing Go that Sedol thinks is lost in the era of superhuman AI performance? There are at least two possible sources. The first is the value of the product of Sedol and others' efforts, namely the games of Go themselves. Part of the value of playing Go at a high level is the production of records and recordings of games, which can be studied for their aesthetic and strategic value.

[8] For Bradford, for an achievement to be competently caused just is for the agent to have a significant number of justified true beliefs about that achievement (Bradford 2015, pp. 65-7). I ignore this condition for several reasons. First, it is only difficulty that ultimately contributes to the value of achievement for Bradford (see Hirji 2019, and ignoring complications about the value of organic unities (see also Hurka 2020) that would take us very far afield). The value of difficulty will be our exclusive focus in section 4. Additionally, I do not find the account of competent causation in terms of justified true belief compelling, preferring instead an account that centers the agent's capacities and dispositions (as in Sosa 2007).

[9] Ignoring that, for a creature with a different cognitive makeup, the latter might be quite difficult.

The other source is in the achievement itself. Those who theorize about achievement are often interested in accounting for its intrinsic value. They believe that achievements are valuable not only for the good outcomes they might produce, but in and of themselves. Often this is explained by a perfectionist account of value. Achievements are valuable because they represent an exercise of the rational will (Bradford 2015), creativity (Hirji 2019), and other kinds of distinctively human capacities that are intrinsically valuable to perfect (Hurka 2020). Playing Go at a high level is a valuable achievement because it requires the player to refine and perfect their rational capacities, and this is an intrinsically valuable thing for rational creatures to do.

I will assume a perfectionist account of the intrinsic value of achievement, though I do not think most of the arguments given here rely too heavily on it.[10] By doing this, I will not be considering a certain class of objections to the creation and use of AI technologies. If one were to object to the prevalence of AlphaGo because it meant no human being could win the top prize money at a tournament which AlphaGo entered, this loss of practical value would not be what I mean to isolate in Sedol's lament. Additionally, though Sedol's lament is couched in the language of competitive games, focusing on the value of achievements allows us to expand our focus to many different valuable human achievements that might be threatened by superhuman AI performance (e.g. the scientific achievements discussed in section 5). We will avoid the worry, found in Hurka (2006), that difficult games represent a valuable diversion that is best pursued once more pressing practical issues have been solved, perhaps in some future Suitsian

[10] One reason I think this: the underlying axiology of perfectionist value is flexible enough that many antecedent commitments can fit within it. For example, consequentialist leanings are compatible with versions of perfectionism (Hurka 1993). One can also imagine how to adjust Hurka's consequentialist theory to take into account the agent-centered prerogatives of nonconsequentialist theories. The important point for technology ethics is one that Rawls (1999, p. 325) makes: perfectionist goods are a kind of good that should be built into *any* moral theory (to be weighed against other goods).

work-free utopia (Suits 1978). In short, formulating Sedol's lament in this way allows us to ask whether, and in what way, the superhuman performance of AI technologies threatens to undermine the value of human achievement. This will be our concern going forward.

3. The Value of Being the GOAT

I want to start with what seems to me the most natural interpretation of Sedol's lament. Sedol is focused on the value of being "at the top," and he worries that the existence of superhuman Go-playing algorithms undermines the achievement of being "the number one," since no matter which human being is the best (human) Go player, there is "an entity that cannot be defeated" (Yonhap News Agency 2019). Sedol thinks that being the *best* at what one does is a particular kind of valuable achievement. AlphaGo removes from the realm of human possibility this kind of achievement. With this, Sedol thinks, comes a significant loss of value, significant enough that one might conclude playing competitive Go is not worth the effort.

Being better at playing Go than anyone else seems like a distinctive kind of achievement. Call this achievement being the "greatest of all time," or the GOAT, for short. It seems plausible that being the greatest Go player of all time is a more significant achievement than being a very good Go player. It is more difficult to achieve GOATness, and it involves more competent causation on the part of the agent as well. The existence of AlphaGo threatens to remove the possibility of being the GOAT (without qualification) from the realm of human attainability. Even if a human were to become a better Go player than any other human who currently exists (or ever has existed), they would still not be the GOAT. We can thus formulate our first reading of Sedol's complaint:

The GOAT reading. There is a distinctive achievement value in being the greatest of all time in some domain. The invention of AlphaGo means that a human being will never again be the GOAT at Go, and this is a significant loss of value.

What should we make of the GOAT reading? An initial worry: it is not obvious that the *mere fact* that someone happens to be the greatest in some area (relative to some comparison class) carries much value, compared with the value of the underlying achievements themselves. The underlying facts about what a person was able to achieve are what give the GOAT's achievements their value, not where it lands them on a list of the greatest achievers of all time. This intuition seems to operate at both extremes. First, suppose there were only five untalented Go players in the whole universe. One of those five might be the GOAT, but it is not plausible that much (if any) value accompanies this fact. At the other extreme, suppose two very talented Go players played a grueling series of games, with one player ultimately coming out on top by the slimmest margins. Further suppose the victorious competitor is able to repeatedly do this, besting their opponent just slightly after playing at the highest standard. The victorious player would be the GOAT, and their achievement might be slightly more valuable as a result. But surely the vast majority of the value in this case is to be found in the great and competitive games the two players have played, not the fact that one is always able to slightly best the other. This suggests that, if there is any value to be had in being the GOAT, it is a relatively small amount. The majority of the value is to be found in the performances that make one the GOAT, not the mere fact that one is the GOAT.

It is also unclear whether, if we adopt a perfectionist account of the value of achievement, the performance of AlphaGo (and other contemporary deep neural networks) should be considered achievements *at all*, and thus whether it makes sense to include them in a

ranking of the greatest achievers in Go. It was not *difficult* in any obvious sense for AlphaGo to learn Go, nor is producing high-level play particularly difficult for it.[11] It's also not obvious that AlphaGo competently caused the moves that constituted its winning games, since competent causation is something agents do when they gain justified true beliefs or otherwise respond to their reasons. It is implausible, on many accounts, that AlphaGo has the right kinds of beliefs and desires to act on reasons in the standard way.[12] It is unclear that AlphaGo is achieving anything, let alone anything of value, when it plays Go.

The mere fact that some technology can outperform a human agent at something (which would be an achievement *for humans* but is not for the technology itself) does not represent much of a threat to the value of that achievement. There are many vehicles that can move faster than the fastest human sprinter. Does the fact that a fighter jet could outpace Usain Bolt in the 400-meter dash undermine the achievement of being the best sprinter in the world? It is hard to see why it should. The comparison class is of the wrong kind. It is a significant achievement to be the fastest human sprinter, and this is not undermined by the fact that some machine could perform at a higher level. The mere superhuman performance of non-agential technologies has no bearing on the value of human achievement in the same area.

The defender of the GOAT reading might respond to this line of reflection in several ways. First, they might point out that *some* value, however small, is attached to being the GOAT

[11] The program is able to play millions of games in the time it would take a human being to play tens or hundreds (Silver et al. 2018). If anything, playing Go is the easiest thing in the world for AlphaGo.

[12] This is not to take a stand on the thorny question of whether a sufficiently complicated AI technology *could* have cognition or agency in the right way. Contrary to classic arguments from Searle (1984), I do not see any in-principle reasons why this could not be a possibility, and there are some interesting extant accounts for how this might happen (e.g. List 2021). Nonetheless, almost everyone agrees that machine learning algorithms as they currently exist lack most of the capacities necessary for agency, and thus for competent causation (though see Danaher 2020).

in all of these cases, and the loss of that value still seems something that could be rationally lamented with reference to the class of "anything that can play Go" (not merely those agents who can play Go).[13] Being the GOAT in any category is an achievement of some kind, and it is an achievement increasingly becoming unavailable to humans due to AI technologies. This seems true, to an extent. There is a particular value of being the best at something, and there is definitely *some* value lost when some other kind of being takes over the top spot. But it is unclear that this limited value loss should bother us, let alone be enough to generate widespread despair, especially when we are being so liberal about the categories we use to rank the best performers in some field. The vast majority of value to be found in an achievement is to be found in the process and outcome itself, not how it compares to others within a given category, let alone when that category is massively permissive. Such refinement is freely available in the age of superhuman AI technologies.

Another line of response might point to the fundamental importance of the cognitive in human life.[14] Unlike physical skill, where we have known for a long time that we could not stand up to other animals and basic advances in automotive technology, the rational and cognitive capacities necessary to play Go are ones that we (up until now) thought were solely the purview of human beings. Losing this kind of cognitive supremacy is itself a loss of a kind of value. This response amounts to a rather implausible rendering of the fundamental insight of the perfectionist tradition, however. Perfectionism tells us to develop the capacities that are *essentially* or *distinctively* human, not those that are *uniquely* human. If we discovered tomorrow a species of aliens that could exercise their rational will in the same way human beings do,

[13] My thanks to Josh Shepherd for pushing me on this point.
[14] My thanks to Jake Quilty-Dunn for discussions of this line of reflection.

would this undermine the value of perfecting the will for human beings? It is hard to see why it should. *Mutatis mutandis* for the case of designing a superhuman-performing AI.

I ultimately do not find the GOAT reading very convincing. It does pick out a particular kind of value (the value of being the greatest of all time) and notes that the performance of AlphaGo undermines that value. While this might be true in the abstract, two objections seem to diminish the value of being the GOAT: the relatively small amount of additional value the GOAT property adds to an achievement (over and above the value of the underlying achievement itself), and the fact that AlphaGo is not the kind of agent that could compete with, and achieve success over, human beings. The GOAT formulation doesn't do enough to justify the anxiety Sedol and others have expressed when considering superhuman AI performance. Another formulation is needed.

4. The Value of Difficulty

I said above that the GOAT formulation is the most natural reading of Sedol's lament. It is also one where the dread felt by many who reflect on superhuman AI performance is not borne out by our theory of the value of achievement. I next want to step away from the language of Sedol's comments themselves to examine another possible threat that has recently occupied some authors in the technology ethics literature.

A prominent line of reflection in AI ethics has focused on the idea that an achievement must be, and gains value to the extent that it is, difficult. Bradford, recall, has an "egalitarian" (Hirji 2019, p. 8) account of the value of achievement, on which its value increases as the difficulty the agent overcomes increases. Anyone can reasonably expect to be able to achieve

something, provided they are willing and able to overcome some significant obstacles. For Bradford, this is for perfectionist reasons. Achievements are valuable because they are expressions of the rational will, and the more difficult an achievement is to obtain, the more the agent must express her rational capacities in order to achieve (Bradford 2015, pp. 122-3). There are several lines of reflection that suggest the coming dominance of AI technologies entails our achievements will be, in some important sense, *easier*, and therefore less valuable. This is far away from the reasons given by Sedol for worrying about superhuman AI performance. Given the deflationary account of those worries we arrived at in the last section, however, perhaps the realm of difficulty is where worries about superhuman AI performance can be substantiated.

Danaher & Nyholm (2020) offer a version of this in what they call the argument from the "achievement gap." They are primarily interested in the achievements that workers often accomplish in the workplace. They worry that the rapid automatization of work will lead to a subsequent loss of opportunities for achieving things of value in the workplace. In typical non-automated work, many workers have opportunities to participate in complex, difficult activities that culminate in significant achievements (for instance, participating in an assembly line that produces a car). With the advance of automation in the workplace, however, a shift in the nature of work has begun to occur. Danaher and Nyholm note that collaborations between humans and machines in the workplace often involve maintenance and order-following on the part of human workers (Danaher & Nyholm 2020, pp. 6-8). Processes that, in a non-automated workplace, would be quite difficult and involved for human beings, become quite easy (and

unengaging) when most of the work is automated. They argue these are not achievements.[15] While the particularities of the workplace setting are not immediately relevant to the AlphaGo case, the general trend of AI technologies making things easier might make sense of the general worries that gave rise to specific complaints like Sedol's.

A more specific version of the argument from the achievement gap can be found in Wang (2021).[16] Wang is particularly interested in the ways that reflections on the nature of achievement interact with the broader debate surrounding cognitive enhancement.[17] But the idea is strikingly similar to the worries Danaher and Nyholm raise. With the advent of AI technologies, either in the form of advanced tutoring from AI, or from AI chip implants meant to augment our existing cognitive capacities, technological advances will make doing many things we now consider difficult much easier in the future. This might seem like something to be welcomed. But if we think that valuable achievements require difficult effort on the part of human beings, this might not be so. Wang concludes that analyses of cognitive enhancement should take into account the lost value of achievement that might result as things become easier. Wang explicitly includes advanced modes of instruction with technologies as a kind of enhancement (Wang 2021, p. 121), so the broader point applies to the AI case as well, even if we think the notion of brain implants is a remote possibility.[18]

---

[15] There are some reasons to push back here, since keeping a "well-oiled machine" running might itself be a genuine achievement.The empirical facts concerning the spread and ubiquity of "bullshit jobs" (Graeber 2013), however, make this a rather theoretical response.

[16] Similar arguments have also been given outside of the perfectionist account of achievement, most obviously in Experience Machine arguments (Nozick 1974).

[17] Some standard citations include Persson & Savulescu (2008) and Levy (2007, ch. 2 & 3).

[18] I raise some particular issues for these ideas below, but they are rather applied in scope. A more systematic critique of the supposed undermining of achievement by enhancement can be found in Forsberg & Skelton (2020).

The main worry is this: in virtue of being aided by AI technologies, many things we now consider achievements will be too easy in the future. They will fail to count as genuine achievements at all, or failing that, AI aid will at least significantly decrease the value of said achievements. Call this the *easy* reading of Sedol's lament (though the content of the worry is not particularly Sedolian):

> **The easy reading**. The advent of superhuman-performing AI technologies will cause, through whatever causal mechanism, human actions (which in the absence of such technologies would be considered valuable achievements) to become too easy to count as genuine achievements in a given domain, or will otherwise undermine their value.

It is easy to see how the easy reading would apply to the Go case. Playing Go with the *aid* of superhuman AI technologies is the main area of concern. What once took many years of training and play for a human to master would now be an activity that anyone with access to the best machine learning technologies could simply memorize and apply. Much like comparing the best gymnastic vault performers in the 1940s and today, what we now consider an achievement might look childish once human performers are consistently trained with the aid of AI technologies. Playing a high-level game of Go would become something more like playing pushpin than anything that counts as an achievement.

While the easy reading has found several proponents in the literature, worries about its soundness can be raised. For one thing, arguments for the easy reading seem to ignore the importance of *ceiling effects* (or, more specifically, a *lack* of nearby ceiling effects) in the development of human skills and talents.[19] The easy reading rightly points out that, for many things we *currently* think of as difficult, the ability to tap into the resources of superhuman AI

[19] There is plenty of philosophical work on the nature and function of human skill (e.g. Stichter 2007; Shepherd 2019), but comparatively less on the notion of talent, though they are intimately connected. I am here relying on the excellent and novel account of talent in Robb (2020).

technologies might make *those particular activities* easy enough that they do not count as achievements on a difficulty-based account. But why think that human beings will stop developing their talents at the levels we are currently at? Provided that a human skill is not at or near ceiling (that is, not near the limits of what is physically possible for a human to achieve), agents will, with the help of AI technologies, be able to play *better* games of Go, discover even more complicated scientific facts, and perhaps even climb taller mountains. If using AI tutors makes getting to some level of achievement easier, the best human agents at that skill will simply leave previous levels of achievement behind. It is true that what we *now* understand as an achievement may come to be viewed as relatively easy by those with access to better training, but this happens all the time as human skills are perfected. There is no threat from AI technologies to human achievement on this reading. In fact, the easy reading might function as an argument *for* developing AI technologies, since better training will allow human beings to achieve more impressive results without other significant downsides.

The world of achievement in chess is an interesting comparison class for Go. Chess has been dominated by AI technologies for more than twenty years. Has this led to a slow and depressing abandonment of chess as an arena of human achievement? Quite the opposite. As measured by online registrations for the world's largest chess website, chess has never been more popular (Brookwell 2020). Human grandmasters continue to compete against other humans in prestigious tournaments for large cash prizes, even though chess engines could consistently defeat any human competitor. The ubiquity of superhuman AI chess engines has also changed the way chess is played. New openings and other moves have been adopted from chess engines, opening up even more complicated styles of play (Levene & Bar-Ilan 2007). The

popularity of speed, bullet, and other faster kinds of chess has also increased. This is another interesting aspect of human talent development that helps us avoid ceiling effects. If a particular area of achievement is becoming stale or easy (e.g. too many draws in classical chess), agents will find a heretofore unexplored nearby area of possible achievement and work to hone their skills in that area instead.[20] What has not happened in chess is a dissolution of valuable human achievement under the relentless assault of AI champions. The picture of achievement in chess offers a much rosier assessment of the future of Go than the easy reading might suggest.

Another reason to be skeptical of the easy reading is that achievement and difficulty are not as intimately related as the difficulty-based view assumes. Hirji (2019) presents a compelling counterexample, focusing on brother and sister poets in Elizabethan England (p. 6). Both are naturally adept poets. The brother, in virtue of living in a society that presents men with opportunities to develop their skills, becomes a world-renowned poet who produces works of great value. The sister, in virtue of that same fact, struggles to produce poetry in addition to the domestic responsibilities heaped on her. She does not produce poems nearly as aesthetically valuable as her brother's. The difficulty-based account claims there is more intrinsic value in the sister's struggles, or at least that the difficulty of the poem might offset its aesthetic flaws. But this is an odd thing for a perfectionist to say, given that the sister's perfection of her capacities has been radically impeded by a sexist society. The verdict gets things the wrong way around. It is precisely in the deprivation of her ability to achieve valuable things that the sexist society wrongs her. Added difficulty does not always add value to an achievement, or if it does, it

[20] Machine learning can in turn be used to evaluate different variants of the rules of chess, creating a feedback loop that pushes players towards new variants that will keep and attract interest within the broader space of "chess-like games" (Tomasev et al. 2020).

might be swamped and counteracted when the difficulty is sufficiently high that the agent cannot develop her capacities enough to produce a valuable product.[21]

It is also not obvious that difficulty is *necessary* for all cases of achievement. The existence of savants and virtuosos, who achieve great things with little effort, bolsters this idea. Consider SwamPerlman, a version of the familiar Swampman who comes into existence in the swamp with the abilities of a virtuoso violinist. Suppose SwamPerlman comes into being three seconds before going on stage to perform the Paganini Concerto, which he does flawlessly. SwamPerlman's flawless performance of a devilishly hard violin concerto seems to be a real achievement, even though he has not struggled at all in the three seconds of his life before the performance. Critics like Hirji have pointed out that the connection between achievement and difficulty, while certainly present, is not as simple as Bradford and others have assumed. The existence of savants and virtuosos who seem to have struggled little but still achieve great things also suggests that difficulty is not necessary to generate a valuable achievement.

For these reasons, I do not think the easy reading is the best way to understand the threat AI technologies pose to human achievement. Most human skills are not at ceiling, so the most likely outcome of advanced AI training will be the expansion of what we consider an achievement, not the disappearance of it. It is, furthermore, plausible that easy achievements can be valuable achievements anyway. This does not mean that all is well for human achievement in the age of AI, however. I catalog a version of Sedol's lament in the next sections that I believe, in certain localized forms, represents a genuine threat to valuable human achievement in superhuman-AI-dominated fields.

[21] My thanks to an anonymous reviewer for pushing me to make this formulation more precise.

5. Displacing Achievement

In this section, I want to motivate a version of Sedol's lament that can survive objections the GOAT and easy readings could not. This reading starts with a series of considerations that Danaher & Nyholm (2020) also mentioned in their paper. Instead of making activities in the workplace too easy to count as achievements, automation also *displaces* workers. This is most obvious in the case of automation-driven job losses, but it can also be true for workers who keep their jobs in an automated workforce, since they will often be shifted to maintenance and upkeep work on machines and away from the complex procedures that previously occupied them. On this version of the worry, automation does not make previously difficult activities too easy to count as achievements. Rather, AI technologies completely replace human actors in domains where previously they had been achievers.

Taking this suggestion at face value, we are presented with the *displacement* reading of Sedol's lament:

> **The displacement reading.** The advent of superhuman-performing AI technologies will cause, through whatever causal mechanism, human actors to fail to engage in achievement-worthy activities at all in a given domain.

Applying the displacement reading to the Go-playing case is a bit tricky, since games are often thought to be one of the activities we would continue to engage in in a technological utopia where work has been eliminated.[22] But the effect might instead be psychological. If the advance of superhuman Go-playing algorithms affects many people like it did Sedol, they may give up playing Go, no matter what philosophers tell them about the axiological basis for their concerns.

[22] This is the classic argument of Suits (1978), though how to precisify the idea is not always clear (Yorke 2018; Wildman & Archer 2019).

There are also many valuable achievements outside of competitive gaming that seem to be vulnerable to AI displacement. Consider the achievement of discovering some significant scientific breakthrough. If a neural network will soon be able to make all of the significant discoveries in a field, it is not hard to imagine human beings primarily being involved in the field by maintaining the networks, not making any discoveries themselves. While building and maintaining complex computational models is almost certainly an achievement in itself, the actual achievements of science would no longer be something humans could participate in, perhaps because the science would grow too complex for human cognizers to track. While this is still a science fiction scenario, such a future might not be far away. Ilyas et al. (2019) argue that adversarial examples (baffling results where deep neural networks will, for instance, classify images that look like a panda as a "stop sign" with high confidence after minor perturbations to the pixels of an image (Szegedy et al. 2013)) might represent neural networks accessing high-dimensional patterns in data that human beings are unable to see. They imagine a future science where machine learning networks track connections in hundreds of dimensions and make precise predictions without the aid of human theorists (who lack the relevant perceptual abilities to make discriminations).[23]

It is also possible that some valuable achievements will be displaced because they are primarily supported for their practical value, which will be more easily met by AI technologies. The AlphaFold algorithm, capable of making accurate predictions about the structure of folded proteins given their amino acid sequences (Jumper et al. 2021), presents a test case for this idea. Before AlphaFold, predicting the structure of folded proteins from their sequences was a

[23] For more on this possibility, and its impact on science, see Buckner (2020).

complex human achievement that, due to the small and unpredictable energetic effects of local folding environments, was impossible to axiomatize and required skill and ingenuity to figure out. In a future where machine learning algorithms can predict the structure of folded proteins as well as, or better than, human molecular biologists, learning those skills might still be a valuable achievement for humans (as I argued above, it is unclear why the existence of a superhuman AI technology for some human skill should say much of anything about the value of learning that skill for humans). But molecular biology departments support the salaries of these experts because of the practical value they provide to the practice of biological science. If algorithms like AlphaFold become widespread in their use, learning the complex, intuitive rules of protein folding will become a game that could be played for its own sake, but will likely not be supported by the scientific establishment. It is not hard to see how such achievements would be displaced in this future.

6. The Inequality of Achievement

Once we shift our focus to the displacement reading of Sedol's lament, a host of more contingent, but also more worrying, possibilities come to light. In an antecedently resource-unequal society like the contemporary United States, for instance, an uncritical focus on achievement as a good can come to have the stench of elitism.[24] Achievements, after all, are often (though not necessarily) difficult. Becoming the type of person who can compete at the highest levels of an activity, or who can achieve some great scientific or literary feat, takes a significant amount of resource investment. For an egalitarian who recognizes the value of

[24] This is the standard objection to political forms of perfectionism; see Nagel (1995), Brink (2007), and Wall (2009) for book-length treatments of these topics.

achievement, talent development should be available to all that want it. In practice, however, many high-achieving people in unequal societies tend to come from socioeconomic groups that have enough access to resources to provide necessary training. This, by itself, is lamentable, as it represents massive amounts of human talent squandered.

The use of AI technology has tended to only exacerbate these underlying inequalities (O'Neil 2016; Noble 2018), often through two separate (though not exclusive) causal pathways.[25] First, technologies can be used by human decision-makers in ways that exacerbate underlying inequalities between groups. This happens, for instance, when AI technology companies outsource a significant amount of the work that goes into making "autonomous" technologies to vulnerable human workers (e.g. Gray & Suri 2019), or when AI technologies are used to make the lives of already well-off humans better at the expense of those worse-off (e.g. Mohamed, Png, & Isaac 2020). Second, AI technologies might themselves encode and exacerbate biases and inequalities. The literature on algorithmic bias is rife with examples. Algorithms disproportionately recommend Black individuals receive harsher parole decisions (e.g. the COMPAS recidivism algorithm; Brennan, Dieterich, & Ehret 2009), are slower to recognize faces of minority group members (Buolamwini & Gebru 2018), and construct proxies for a subject's race and gender to operate even when using such variables is disallowed by programmers (Adler et al. 2016). Though a theoretical debate exists as to whether these biases exist in the

[25] As an anonymous reviewer points out, though the specific empirical facts cited here are widely discussed and (mostly) accepted, it is possible to contest them. Even so, I think the project sketched in this section is interesting regardless, if for no other reason than as a conditional claim. If the social and political facts are as this section claims, then a version of displacement represents a real threat to the value of widespread human achievement in the era of superhuman AI. How various institutional and social realities intersect with the normative theory of achievement in the era of AI is a broad research project on which I have much more to say, but can only gesture at here due to space limitations.

algorithms themselves or are just reflections of the data that algorithms work on (Johnson 2020), the effect of algorithmic bias in exacerbating inequality is well-established.

My argument, given this empirical preamble, is fairly straightforward. As long as AI technologies are more likely to be accessible by those who have the resources to pursue high-achieving activities, an *inequality of achievement* is likely to emerge. Those who have resources will tend to be favored by the inequalities created by AI technologies, and they will tend to have more access to resources that will allow them to take full advantage of AI technologies in furthering their advantage. Though two agents might both try just as hard to access some scientific or sporting achievement, the one whose effort is aided by increasingly powerful AI technologies will be the one much more likely to come out like the brother poet in Hirji's critique of Bradford. Increased use of AI technologies will make more achievements like achievements in motorsport, where a massive amount of resource investment in young drivers is needed to produce high-achieving individuals. Without the right background, achievements with the aid of AI will just not be possible for most people.

This inequality will not impact all areas of achievement equally. It will be damaging for those areas of achievement where AI aid will be particularly helpful, and where the value of the achievement is *comparative* (or relative to the performance of other human beings). Many sporting achievements are comparative in this sense. What matters is not how good a football team is *ex nihilo*, but how well they are able to outperform their opponents. Go has this comparative structure, and it is an area where training with AI technologies could significantly improve our performance (since, as I argued above, it is implausible that we are near the human ceiling for games like Go). The future this threatens to set up is one where those with the

resources to access AI-assisted training will be able to develop their skills to the point where those without such access are not able to compete. The highest rungs of competitive Go will be filled only with those from the right kind of background, and only those with such a background will be encouraged to spend the time to develop their skills in the first place. In many ways, competitive sports are already like this. AI technologies threaten to accelerate and exacerbate these inequalities in other areas as well.

There are areas where such an effect might not be as drastic. This will be true if (a) the AI technologies necessary to train oneself are not difficult for the majority of potential achievers to obtain, and (b) the technology itself has a kind of egalitarian function, bringing those from disparate groups up to a level playing field rather than exacerbating their different starting points. Some might argue the function of AI technologies in chess represents a useful example of both (a) and (b). While prospective chess players still have to have the ability and knowledge to access websites like chess.com, chess engines and their recommendations are embedded in the website's architecture. They allow for talent development with the aid of AI technologies that might provide training for someone with no other resources available. Why not think this will also occur for other AI technologies in areas like Go, scientific research, or violin playing? There is no reason why, in principle, they could not. (This objection, remember, is not one concerning the nature of the technology itself, but how it is currently used.) But the amount of computational power needed to make and use AlphaGo is much greater than these simple chess engines. AI technology companies also have a distinct incentive to monetize and keep control over their products in a way that the creators of simpler chess engines do not. Thus, while not impossible, I would be very surprised to see freely-available AlphaGo clones with anything like

the computing power of AlphaGo in the near future, let alone for the networks that aid scientific research.

Another reaction might be: so what? While achievements like playing high-level Go are certainly things that people care about and derive significant value from, these are not the most important achievements we as human beings could aim for (recall Hurka 2006; though see Nguyen 2019). Why should we care about rich people being able to play Go at a level that less well-off individuals will not be able to? Two responses suggest themselves. First, one might admit that playing games at a high level isn't the only possible valuable thing for a human to achieve, while nevertheless lamenting it becoming less available to a broader class of people. Why should *any* valuable activity be the sole purview of a special monied class? The fruits of some valuable activity being only available to the rich is something to be lamented, even if the activity is not the most valuable one we can imagine. Second, it is not clear that this will only be true for "trivial" achievements like games. High-level training in the sciences and humanities, for instance, is also a competitive good,[26] where only a certain number of students are admitted to grad schools, offered postdocs and jobs, and published in journals. If being able to access AI training in these fields will make it more likely that some will pass these hurdles while others will not, who has access to something as meaningful as great scientific achievement will become even more of a rarified achievement than it is already (Teachman 1987). This, too, is an outcome to be avoided wherever possible.

This situation, which we already find ourselves to some degree, is what I suggest should generate anxiety at the advance of superhuman AI technologies. It is not, to be clear, an

[26] As theorists of the "leaky pipeline" in academia have long noted; see Cheryan et al. (2017) for a recent example in STEM fields in particular.

existential concern. There is nothing internal to the nature and value of human achievement that suggests the march of algorithms like AlphaGo should do much to dampen our fundamental drive for valuable achievements. The threat is instead one related to justice and equity. Absent a change in our current trajectory, only those with significant resources will be able in the future to achieve and perfect their capacities. For anyone who cares about the value of human achievement in a just society, this result should be unacceptable.[27]

7. Conclusion

I have not considered here every possible argument for the internal threat of superhuman AI performance to the value of human achievement, so I cannot claim to have shown that such a threat does not exist.[28] I have, however, argued that two natural readings of Sedol's lament do not offer a bleak picture of our future with superhuman AI technologies, provided we keep a plausible theory of human achievement in focus. Instead, I have argued the main threat is more particular and local. In an already unjust society, the adoption of AI technologies makes the ability to perfect one's abilities the domain of only a select few with antecedent wealth. This should be particularly distasteful for a perfectionist who thinks every

[27] Attempts to mitigate the results of algorithmic bias in particular have been attempted, especially in "algorithmic auditing" (see the framework in Raji et al. 2020). But it is unclear how much these internal fixes, originating at and being implemented in companies whose interests are clearly aligned with inequality-driving forces, will be sufficient to alleviate the problem.

[28] For instance, if one thinks the value of achievement is partially or wholly grounded in the enjoyment that we get out of the process of achieving, then the fact that we despair when contemplating the rise of superhuman AI might *itself* be enough to undermine the value of our achievements. This could be true even if all the arguments presented in this paper are on the right track. While I think this view represents an implausibly subjective view of the value of achievement, more work is needed to tease out the threads of these kinds of downstream issues. I am thankful to an anonymous reviewer for suggesting this line of future work.

human being has the ability and duty to perfect their talents, since such inequalities make it impossible for most of us to achieve the highest goods we could as human beings.[29]

References

Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2016). Auditing black-box models for indirect influence. In *2016 IEEE 16th international conference on data mining* (ICDM) (pp. 1–10). IEEE.

Bradford, G. (2015). *Achievement.* Oxford University Press.

Brink, D. O. (2007). *Perfectionism and the Common Good: Themes in the Philosophy of TH Green.* Clarendon Press.

Brookwell, I. (2020). The hottest new video game is… chess? *Fast Company.*

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36(1), 21-40.

Buckner, C. (2019). Rational inference: The lowest bounds. *Philosophy and Phenomenological Research*, 98(3), 697-724.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence*, 2(12), 731-736.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others?. *Psychological bulletin*, 143(1), 1.

Danaher, J. (2020). Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and Engineering Ethics*, 26(4), 2023-2049.

Danaher, J., & Nyholm, S. (2020). Automation, work and the achievement gap. *AI and Ethics*, 1-11.

Forsberg, L., & Skelton, A. (2020). Achievement and Enhancement. *Canadian Journal of Philosophy*, 50(3), 322-338.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, 30(3), 411-437.

Graeber, D. (2013). On the phenomenon of bullshit jobs: A work rant. *Strike Magazine*, 3, 1-5.

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[29] My deepest gratitude to Joe Moore, Thomas Lambert, Jake Quilty-Dunn, Josh Shepard, Anncy Thresher, Jon Vandenburgh, Michael Ball-Blakely, Ting-An Lin, Diana Acosta-Navas, Henrik Kugelberg, Valerie Soon, Anne Newman, Rob Reich, Tom Kelly, Colin Allen, Tony Chemero, and Zvi Biener for comments and conversation that improved this paper immensely. Thanks are also due to audiences at Stanford University, the University of Cincinnati, the University of Pittsburgh, Florida Atlantic University, and the 2021 iteration of the Society for Philosophy and Psychology annual meeting.

Halina, M. (2021). Insightful artificial intelligence. *Mind and Language*.
Hirji, S. (2019). Not always worth the effort: Difficulty and the value of achievement. *Pacific Philosophical Quarterly*, 100(2), 525-548.
Hsu, F. H. (2002). *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton University Press.
Hurka, T. (1993). *Perfectionism*. Oxford University Press.
Hurka, T. (2006). Games and the good. *Proceedings of the Aristotelian Society* 106 (1):217-235.
Hurka, T. (2020). The Parallel Goods of Knowledge and Achievement. *Erkenntnis*, 85(3), 589-608.
Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv preprint arXiv*:1905.02175.
Johnson, G. M. (2020). Algorithmic bias: on the implicit biases of social technology. *Synthese*, 1-21.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., ... & Hassabis, D. (2020). High accuracy protein structure prediction using deep learning. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction*, 22-24.
von Kriegstein, H. (2017). Effort and Achievement. *Utilitas* 29(1), 27-51.
Lee, C. S., Wang, M. H., Yen, S. J., Wei, T. H., Wu, I. C., Chou, P. C., ... & Yan, T. H. (2016). Human vs. computer go: Review and prospect. *IEEE Computational intelligence magazine*, 11(3), 67-72.
Levene, M., & Bar-Ilan, J. (2007). Comparing typical opening move choices made by humans and chess engines. *The Computer Journal*, 50(5), 567-573.
Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge University Press.
List, C. (2021). Group Agency and Artificial Intelligence. *Philosophy and Technology*, 1-30.
Nagel, T. (1995). *Equality and partiality*. Oxford University Press.
Nguyen, C. T. (2019). Games and the art of agency. *Philosophical Review*, 128(4), 423-462.
Noble, S. U. (2018). *Algorithms of oppression*. New York University Press.
Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.
O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
Pranam, A. (2019) Why The Retirement Of Lee Se-Dol, Former 'Go' Champion, Is A Sign Of Things To Come. *Forbes.*
Persson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of applied philosophy*, 25(3), 162-177.
Peterson, M. (2019). The value alignment problem: a geometric approach. *Ethics and Information Technology*, 21(1), 19-28.
Portmore, D. W. (2008). Welfare, Achievement, and Self-Sacrifice. *Journal of Ethics and Social Philosophy*, 2(2).
Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).
Rawls, J. (1999). *A theory of justice: Revised edition*. Harvard university press.

Robb, C. M. (2020). Talent dispositionalism. *Synthese*, 1-18.
Scanlon, T. (1998). *What we owe to each other*. Belknap Press.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3), 417-424.
Shepherd, J. (2019). Skilled action and the double life of intention. *Philosophy and phenomenological research*, 98(2), 286-305.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2016). Mastering the game of go without human knowledge. *Nature*, 550(7676), 354-359.
Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140-1144.
Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (Vol. 1). Oxford University Press.
Stichter, M. (2007). Ethical expertise: The skill model of virtue. *Ethical Theory and Moral Practice*, 10(2), 183-194.
Suits, B. (1978). *The grasshopper*. University of Toronto Press.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint* arXiv:1312.6199.
Teachman, J. D. (1987). Family background, educational resources, and educational attainment. *American sociological review*, 548-557.
Tomašev, N., Paquet, U., Hassabis, D., & Kramnik, V. (2020). Assessing game balance with AlphaZero: Exploring alternative rule sets in chess. *arXiv preprint arXiv*:2009.04374.
Wang, J. (2021). Cognitive Enhancement and the Value of Cognitive Achievement. *Journal of Applied Philosophy* 38(1), 121-135.
Wildman, N., & Archer, A. (2019). Playing with Art in Suits' Utopia. *Sport, Ethics and Philosophy*, 13(3-4), 456-470.
Yonhap News Agency (2019). Go Master Lee Says He Quits, Unable to Win over AI Go Players. *Yonhap News.*
Yorke, C. C. (2018). Bernard Suits on capacities: games, perfectionism, and Utopia. J*ournal of the Philosophy of Sport*, 45(2), 177-188.