Chapter 7

Goals

The mystery of human existence lies not in just staying alive, but in finding something to live for.

Fyodor Dostoyevsky, *The Brothers Karamazov*

Life is a journey, not a destination.

Ralph Waldo Emerson

If I had to summarize in a single word what the thorniest AI controversies are about, it would be “goals”: Should we give AI goals, and if so, whose goals? How can we give AI goals? Can we ensure that these goals are retained even if the AI gets smarter? Can we change the goals of an AI that’s smarter than us? What are our ultimate goals? These questions are not only difficult, but also crucial for the future of life: if we don’t know what we want, we’re less likely to get it, and if we cede control to machines that don’t share our goals, then we’re likely to get what we don’t want.

Physics: The Origin of Goals

To shed light on these questions, let’s first explore the ultimate origin of goals. When we look around us in the world, some processes strike us as *goal-oriented* while others don’t. Consider, for example, the process of a soccer ball being kicked for the game-winning shot. The behavior of the ball itself does not appear goal-oriented, and is most economically explained in terms of Newton’s laws of motion, as a reaction to the kick. The behavior of the player, on the other hand, is most economically explained not mechanistically in terms of atoms pushing each other around, but in terms of her having the *goal* of maximizing her team’s score. How did such goal-oriented behavior emerge from the physics of our early Universe, which consisted merely of a bunch of particles bouncing around seemingly without goals?

Intriguingly, the ultimate roots of goal-oriented behavior can be found in the laws of physics themselves, and manifest themselves even in simple processes that don’t involve life. If a lifeguard rescues a swimmer, as in figure 7.1, we expect her not to go in a straight line, but to run a bit further along the beach where she can go faster than in the water, thereby turning slightly when she enters the water. We naturally interpret her choice of trajectory as goal-oriented, since out of all possible trajectories, she’s deliberately choosing the optimal one that gets her to the swimmer as fast as possible. Yet a simple light ray similarly bends when it enters water (see figure 7.1), also minimizing the travel time to its destination! How can this be?

This is known in physics as *Fermat’s principle,* articulated in 1662, and it provides an alternative way of predicting the behavior of light rays. Remarkably, physicists have since discovered that *all* laws of classical physics can be mathematically reformulated in an analogous way: out of all ways that nature could choose to do something, it prefers the optimal way, which typically boils down to minimizing or maximizing some quantity. There are two mathematically equivalent ways of describing each physical law: either as the past causing the future, or as nature optimizing something. Although the second way usually isn’t taught in introductory physics courses because the math is tougher, I feel that it’s more elegant and profound. If a person is trying to optimize something (for example, their score, their wealth or their happiness) we’ll naturally describe their pursuit of it as goal-oriented. So if nature itself is trying to optimize something, then no wonder that goal-oriented behavior can emerge: it was hardwired in from the start, in the very laws of physics.

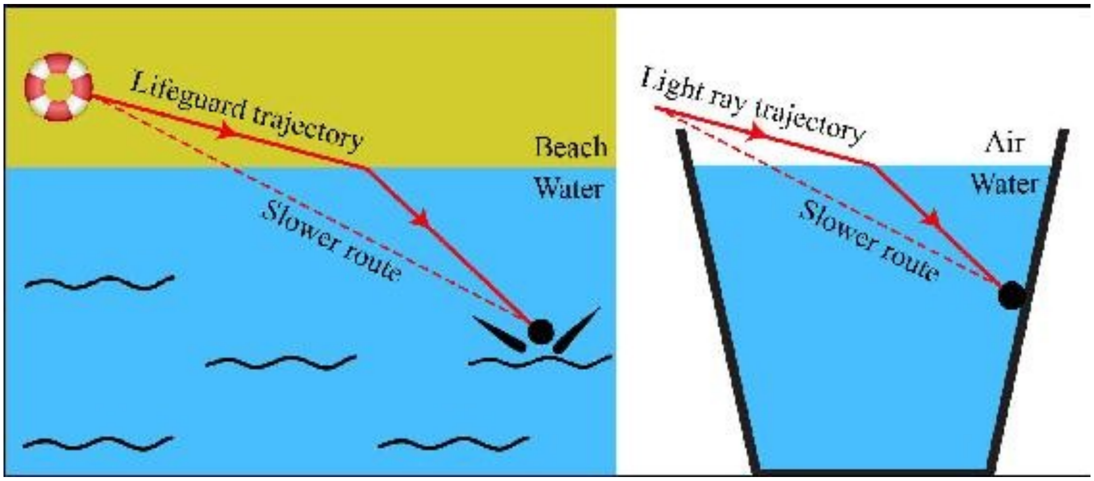

Figure 7.1: To rescue a swimmer as fast as possible, a lifeguard won’t go in a straight line (dashed), but a bit further along the beach where she can go faster than in the water. A light ray similarly bends when entering the water to reach its destination as fast as possible.

One famous quantity that nature strives to maximize is *entropy,* which loosely speaking measures how messy things are. The second law of thermodynamics states that entropy tends to increase until it reaches its maximum possible value. Ignoring the effects of gravity for now, this maximally messy end state is called *heat death,* and corresponds to everything being spread out in boring perfect uniformity, with no complexity, no life and no change. When you pour cold milk into hot coffee, for example, your beverage appears to march irreversibly toward its own personal heat death goal, and before long, it’s all just a uniform lukewarm mixture. If a living organism dies, its entropy also starts to rise, and before long, the arrangement of its particles tends to get much less organized.

Nature’s apparent goal to increase entropy helps explain why time seems to have a preferred direction, making movies look unrealistic if played backward: if you drop a glass of wine, you expect it to shatter against the floor and increase global messiness (entropy). If you then saw it *unshatter* and come flying back up to your hand intact (decreasing entropy), you probably wouldn’t drink it, figuring you’d already had a glass too many.

When I first learned about our inexorable progression toward heat death, I found it rather depressing, and I wasn’t alone: thermodynamics pioneer Lord Kelvin wrote in 1841 that “the result would inevitably be a state of universal rest and death," and it's hard to find solace in the idea that nature's long-term goal is to maximize death and destruction. However, more recent discoveries have shown that things aren't quite that bad. First of all, gravity behaves differently from all other forces and strives to make our Universe not more uniform and boring but more clumpy and interesting. Gravity therefore transformed our boring early Universe, which was almost perfectly uniform, into today's clumpy and beautifully complex cosmos, teeming with galaxies, stars and planets. Thanks to gravity, there's now a wide range of temperatures allowing life to thrive by combining hot and cold: we live on a comfortably warm planet absorbing 6,000°C (10,000°F) solar heat while cooling off by radiating waste heat into frigid space whose temperature is just 3°C (5°F) above absolute zero.

Second, recent work by my MIT colleague Jeremy England and others has brought more good news, showing that thermodynamics also endows nature with a goal more inspiring than heat death.[1] This goal goes by the geeky name *dissipation-driven adaptation,* which basically means that random groups of particles strive to organize themselves so as to extract energy from their environment as efficiently as possible ("dissipation" means causing entropy to increase, typically by turning useful energy into heat, often while doing useful work in the process). For example, a bunch of molecules exposed to sunlight would over time tend to arrange themselves to get better and better at absorbing sunlight. In other words, nature appears to have a built-in goal of producing self-organizing systems that are increasingly complex and lifelike, and this goal is hardwired into the very laws of physics.

How can we reconcile this cosmic drive toward life with the cosmic drive toward heat death? The answer can be found in the famous 1944 book *What's Life?* by Erwin Schrödinger, one of the founders of quantum mechanics. Schrödinger pointed out that a hallmark of a living system is that it maintains or reduces its entropy by increasing the entropy around it. In other words, the second law of thermodynamics has a life loophole: although the total entropy must increase, it's allowed to decrease in some places as long as it increases even more elsewhere. So life maintains or increases its complexity by making its environment messier.

Biology: The Evolution of Goals

We just saw how the origin of goal-oriented behavior can be traced all the way back to the laws of physics, which appear to endow particles with the goal of arranging themselves so as to extract energy from their environment as efficiently as possible. A great way for a particle arrangement to further this goal is to make copies of itself, to produce more energy absorbers. There are many known examples of such emergent self-replication: for example, vortices in turbulent fluids can make copies of themselves, and clusters of microspheres can coax nearby spheres into forming identical clusters. At some point, a particular arrangement of particles got so good at copying itself that it could do so almost indefinitely by extracting energy and raw materials from its environment. We call such a particle arrangement *life*. We still know very little about how life originated on Earth, but we know that primitive life forms were already here about 4 billion years ago.

If a life form copies itself and the copies do the same, then the total number will keep doubling at regular intervals until the population size bumps up against resource limitations or other problems. Repeated doubling soon produces huge numbers: if you start with one and double just three hundred times, you get a quantity exceeding the number of particles in our Universe. This means that not long after the first primitive life form appeared, huge quantities of matter had come alive. Sometimes the copying wasn’t perfect, so soon there were many different life forms trying to copy themselves, competing for the same finite resources. Darwinian evolution had begun.

If you had been quietly observing Earth around the time when life got started, you would have noticed a dramatic change in goal-oriented behavior. Whereas earlier, the particles seemed as though they were trying to increase average messiness in various ways, these newly ubiquitous self-copying patterns seemed to have a different goal: not dissipation but *replication*. Charles Darwin elegantly explained why: since the most efficient copiers outcompete and dominate the others, before long any random life form you look at will be highly optimized for the goal of replication.

How could the goal change from dissipation to replication when the laws of physics stayed the same? The answer is that the fundamental goal (dissipation) *didn't* change, but led to a different *instrumental goal,* that is, a subgoal that helped accomplish the fundamental goal. Take eating, for example. We all seem to have the goal of satisfying our hunger cravings even though we know that evolution's only fundamental goal is replication, not mastication. This is because eating aids replication: starving to death gets in the way of having kids. In the same way, replication aids dissipation, because a planet teeming with life is more efficient at dissipating energy. So in a sense, our cosmos invented life to help it approach heat death faster. If you pour sugar on your kitchen floor, it can in principle retain its useful chemical energy for years, but if ants show up, they'll dissipate that energy in no time. Similarly, the petroleum reserves buried in the Earth's crust would have retained their useful chemical energy for much longer had we bipedal life forms not pumped it up and burned it.

Among today's evolved denizens of Earth, these instrumental goals seem to have taken on a life of their own: although evolution optimized them for the sole goal of replication, many spend much of their time not producing offspring but on activities such as sleeping, pursuing food, building homes, asserting dominance and fighting or helping others—sometimes even to an extent that *reduces* replication. Research in evolutionary psychology, economics and artificial intelligence has elegantly explained why. Some economists used to model people as rational agents, idealized decision makers who always choose whatever action is optimal in pursuit of their goal, but this is obviously unrealistic. In practice, these agents have what Nobel laureate and AI pioneer Herbert Simon termed "bounded rationality" because they have limited resources: the rationality of their decisions is limited by their available information, their available time to think and their available hardware with which to think. This means that when Darwinian evolution is optimizing an organism to attain a goal, the best it can do is implement an approximate algorithm that works reasonably well in the restricted context where the agent typically finds itself. Evolution has implemented replication optimization in precisely this way: rather than ask in every situation which action will maximize an organism's number of successful offspring, it implements a hodgepodge of heuristic hacks: rules of thumb that usually work well. For most animals, these include sex drive, drinking when thirsty, eating when hungry and avoiding things that taste bad or hurt.

These rules of thumb sometimes fail badly in situations that they weren't designed to handle, such as when rats eat delicious-tasting rat poison, when moths get lured into glue traps by seductive female fragrances and when bugs fly into candle flames.[*1] Since today's human society is very different from the environment evolution optimized our rules of thumb for, we shouldn't be surprised to find that our behavior often fails to maximize baby making. For example, the subgoal of not starving to death is implemented in part as a desire to consume caloric foods, triggering today's obesity epidemic and dating difficulties. The subgoal to procreate was implemented as a desire for sex rather than as a desire to become a sperm/egg donor, even though the latter can produce more babies with less effort.

Psychology: The Pursuit of and Rebellion Against Goals

In summary, a living organism is an agent of bounded rationality that doesn't pursue a single goal, but instead follows rules of thumb for what to pursue and avoid. Our human minds perceive these evolved rules of thumb as *feelings*, which usually (and often without us being aware of it) guide our decision making toward the ultimate goal of replication. Feelings of hunger and thirst protect us from starvation and dehydration, feelings of pain protect us from damaging our bodies, feelings of lust make us procreate, feelings of love and compassion make us help other carriers of our genes and those who help them and so on. Guided by these feelings, our brains can quickly and efficiently decide what to do without having to subject every choice to a tedious analysis of its ultimate implications for how many descendants we'll produce. For closely related perspectives on feelings and their physiological roots, I highly recommend the writings of William James and António Damásio.[2]

It's important to note that when our feelings occasionally work *against* baby making, it's not necessarily by accident or because we get tricked: our brain can rebel against our genes and their replication goal quite deliberately, for example by choosing to use contraceptives! More extreme examples of the brain rebelling against its genes include choosing to commit suicide or spend life in celibacy to become a priest, monk or nun.

Why do we sometimes choose to rebel against our genes and their replication goal? We rebel because by design, as agents of bounded rationality, we're loyal only to our feelings. Although our brains evolved merely to help copy our genes, our brains couldn't care less about this goal since we have no feelings related to genes—indeed, during most of human history, our ancestors didn't even know that they *had* genes. Moreover, our brains are way smarter than our genes, and now that we understand the goal of our genes (replication), we find it rather banal and easy to ignore. People might realize why their genes make them feel lust, yet have little desire to raise fifteen children, and therefore choose to hack their genetic programming by combining the emotional rewards of intimacy with birth control. They might realize why their genes make them crave sweets yet have little desire to gain weight, and therefore choose to hack their genetic programming by combining the emotional rewards of a sweet beverage with zero-calorie artificial sweeteners.

Although such reward-mechanism hacks sometimes go awry, such as when people get addicted to heroin, our human gene pool has thus far survived just fine despite our crafty and rebellious brains. It’s important to remember, however, that the ultimate authority is now our feelings, not our genes. This means that human behavior isn’t strictly optimized for the survival of our species. In fact, since our feelings implement merely rules of thumb that aren’t appropriate in all situations, human behavior strictly speaking doesn’t have a single well-defined goal at all.

Engineering: Outsourcing Goals

Can machines have goals? This simple question has triggered great controversy, because different people take it to mean different things, often related to thorny topics such as whether machines can be conscious and whether they can have feelings. But if we're more practical and simply take the question to mean "Can machines exhibit goal-oriented behavior?," then the answer is obvious: "Of course they can, since we can design them that way!" We design mousetraps to have the goal of catching mice, dishwashers with the goal of cleaning dishes, and clocks with the goal of keeping time. When you confront a machine, the empirical fact that it's exhibiting goal-oriented behavior is usually all you care about: if you're chased by a heat-seeking missile, you don't really care whether it has consciousness or feelings! If you still feel uncomfortable saying that the missile has a goal even if it isn't conscious, you can for now simply read "purpose" when I write "goal"—we'll tackle consciousness in the next chapter.

So far, most of what we build exhibits only goal-oriented *design*, not goal-oriented *behavior:* a highway doesn't behave; it merely sits there. However, the most economical explanation for its existence is that it was designed to accomplish a goal, so even such passive technology is making our Universe more goal-oriented. *Teleology* is the explanation of things in terms of their purposes rather than their causes, so we can summarize the first part of this chapter by saying that our Universe keeps getting more teleological.

Not only *can* non-living matter have goals, at least in this weak sense, but it increasingly *does*. If you'd been observing Earth's atoms since our planet formed, you'd have noticed three stages of goal-oriented behavior:

1. All matter seemed focused on dissipation (entropy increase).
2. Some of the matter came alive and instead focused on replication and subgoals of that.
3. A rapidly growing fraction of matter was rearranged by living organisms to help accomplish their goals.

Table 7.1 shows how dominant humanity has become from the physics perspective: not only do we now contain more matter than all other mammals except cows (which are so numerous because they serve our goals of consuming beef and dairy products), but the matter in our machines, roads, buildings and other engineering projects appears on track to soon overtake all living matter on Earth. In other words, even without an intelligence explosion, most matter on Earth that exhibits goal-oriented properties may soon be designed rather than evolved.

Goal-Oriented Entities	**Billions of Tons**
5×10^{30} bacteria	400
Plants	400
10^{15} mesophelagic fish	10
1.3×10^{9} cows	0.5
7×10^{9} humans	0.4
10^{14} ants	0.3
1.7×10^{6} whales	0.0005
Concrete	100
Steel	20
Asphalt	15
1.2×10^{9} cars	2

Table 7.1: Approximate amounts of matter on Earth in entities that are evolved or designed for a goal. Engineered entities such as buildings, roads and cars appear on track to overtake evolved entities such as plants and animals.

This new third kind of goal-oriented behavior has the potential to be much more diverse than what preceded it: whereas evolved entities all have the same ultimate goal (replication), designed entities can have virtually any ultimate goal, even opposite ones. Stoves try to heat food while refrigerators try to cool food. Generators try to convert motion into electricity while motors try to convert electricity into motion. Standard chess programs try to win at chess, but there are also ones competing in tournaments with the goal of losing at chess.

There's a historical trend for designed entities to get goals that are not only more diverse, but also more *complex:* our devices are getting smarter. We engineered our earliest machines and other artifacts to have quite simple goals, for example houses that aimed to keep us warm, dry and safe. We’ve gradually learned to build machines with more complex goals, such as robotic vacuum cleaners, self-flying rockets and self-driving cars. Recent AI progress has given us systems such as Deep Blue, Watson and AlphaGo, whose goals of winning at chess, winning at quiz shows and winning at Go are so elaborate that it takes significant human mastery to properly appreciate how skilled they are.

When we build a machine to help us, it can be hard to perfectly align its goals with ours. For example, a mousetrap may mistake your bare toes for a hungry rodent, with painful results. All machines are agents with bounded rationality, and even today’s most sophisticated machines have a poorer understanding of the world than we do, so the rules they use to figure out what to do are often too simplistic. That mousetrap is too trigger-happy because it has no clue what a mouse is, many lethal industrial accidents occur because machines have no clue what a person is, and the computers that triggered the trillion-dollar Wall Street “flash crash” in 2010 had no clue that what they were doing made no sense. Many such goal-alignment problems can therefore be solved by making our machines smarter, but as we learned from Prometheus in chapter 4, ever-greater machine intelligence can post serious new challenges for ensuring that machines share our goals.

Friendly AI: Aligning Goals

The more intelligent and powerful machines get, the more important it becomes that their goals are aligned with ours. As long as we build only relatively dumb machines, the question isn't whether human goals will prevail in the end, but merely how much trouble these machines can cause humanity before we figure out how to solve the goal-alignment problem. If a superintelligence is ever unleashed, however, it will be the other way around: since intelligence is the ability to accomplish goals, a superintelligent AI is by definition much better at accomplishing its goals than we humans are at accomplishing ours, and will therefore prevail. We explored many such examples involving Prometheus in chapter 4. If you want to experience a machine's goals trumping yours right now, simply download a state-of-the-art chess engine and try beating it. You never will, and it gets old quickly…

In other words, *the real risk with AGI isn't malice but competence*. A superintelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we're in trouble. As I mentioned in chapter 1, people don't think twice about flooding anthills to build hydroelectric dams, so let's not place humanity in the position of those ants. Most researchers therefore argue that if we ever end up creating superintelligence, then we should make sure it's what AI-safety pioneer Eliezer Yudkowsky has termed "friendly AI": AI whose goals are aligned with ours.[3]

Figuring out how to align the goals of a superintelligent AI with our goals isn't just important, but also hard. In fact, it's currently an unsolved problem. It splits into three tough subproblems, each of which is the subject of active research by computer scientists and other thinkers:

1. Making AI *learn* our goals
2. Making AI *adopt* our goals
3. Making AI *retain* our goals

Let's explore them in turn, deferring the question of what to mean by "our goals" to the next section.

To learn our goals, an AI must figure out not what we do, but why we do it. We humans accomplish this so effortlessly that it's easy to forget how hard the task is for a computer, and how easy it is to misunderstand. If you ask a future self-driving car to take you to the airport as fast as possible and it takes you literally, you'll get there chased by helicopters and covered in vomit. If you exclaim, "That's not what I wanted!," it can justifiably answer, "That's what you asked for." The same theme recurs in many famous stories. In the ancient Greek legend, King Midas asked that everything he touched turn to gold, but was disappointed when this prevented him from eating and even more so when he inadvertently turned his daughter to gold. In the stories where a genie grants three wishes, there are many variants for the first two wishes, but the third wish is almost always the same: "Please undo the first two wishes, because that's not what I really wanted."

All these examples show that to figure out what people really want, you can't merely go by what they say. You also need a detailed model of the world, including the many shared preferences that we tend to leave unstated because we consider them obvious, such as that we don't like vomiting or eating gold. Once we have such a world model, we can often figure out what people want even if they don't tell us, simply by observing their goal-oriented behavior. Indeed, children of hypocrites usually learn more from what they see their parents do than from what they hear them say.

AI researchers are currently trying hard to enable machines to infer goals from behavior, and this will be useful also long before any superintelligence comes on the scene. For example, a retired man may appreciate it if his eldercare robot can figure out what he values simply by observing him, so that he's spared the hassle of having to explain everything with words or computer programming. One challenge involves finding a good way to encode arbitrary systems of goals and ethical principles into a computer, and another challenge is making machines that can figure out which particular system best matches the behavior they observe.

A currently popular approach to the second challenge is known in geek-speak as *inverse reinforcement learning,* which is the main focus of a new Berkeley research center that Stuart Russell has launched. Suppose, for example, that an AI watches a firefighter run into a burning building and save a baby boy. It might conclude that her goal was rescuing him and that her ethical principles are such that she values his life higher than the comfort of relaxing in her fire truck —and indeed values it enough to risk her own safety. But it might alternatively infer that the firefighter was freezing and craved heat, or that she did it for the exercise. If this one example were all the AI knew about firefighters, fires and babies, it would indeed be impossible to know which explanation was correct. However, a key idea underlying inverse reinforcement learning is that we make decisions all the time, and that every decision we make reveals something about our goals. The hope is therefore that by observing lots of people in lots of situations (either for real or in movies and books), the AI can eventually build an accurate model of all our preferences.[4]

In the inverse reinforcement-learning approach, a core idea is that the AI is trying to maximize not the goal-satisfaction of itself, but that of its human owner.

It therefore has an incentive to be cautious when it's unclear about what its owner wants, and to do its best to find out.

It should also be fine with its owner switching it off, since that would imply that it had misunderstood what its owner really wanted.

Even if an AI can be built to learn what your goals are, this doesn't mean that it will necessarily adopt them. Consider your least favorite politicians: you know what they want, but that's not what *you* want, and even though they try hard, they've failed to persuade you to adopt their goals.

We have many strategies for imbuing our children with our goals—some more successful than others, as I've learned from raising two teenage boys. When those to be persuaded are computers rather than people, the challenge is known as the *value-loading problem,* and it's even harder than the moral education of children. Consider an AI system whose intelligence is gradually being improved from subhuman to superhuman, first by us tinkering with it and then through recursive self-improvement like Prometheus. At first, it's much less powerful than you, so it can't prevent you from shutting it down and replacing those parts of its software and data that encode its goals—but this won't help, because it's still too dumb to fully *understand* your goals, which requires human-level intelligence to comprehend. At last, it's much smarter than you and hopefully able to understand your goals perfectly—but this may not help either, because by now, it's much more powerful than you and might not let you shut it down and replace its goals any more than you let those politicians replace your goals with theirs.

In other words, the time window during which you can load your goals into an AI may be quite short: the brief period between when it's too dumb to get you and too smart to let you. The reason that value loading can be harder with machines than with people is that their intelligence growth can be much faster: whereas children can spend many years in that magic persuadable window where their intelligence is comparable to that of their parents, an AI might, like Prometheus, blow through this window in a matter of days or hours.

Some researchers are pursuing an alternative approach to making machines adopt our goals, which goes by the buzzword *corrigibility.* The hope is that one can give a primitive AI a goal system such that it simply doesn't care if you occasionally shut it down and alter its goals. If this proves possible, then you can safely let your AI get superintelligent, power it off, install your goals, try it out for a while and, whenever you're unhappy with the results, just power it down and make more goal tweaks.

But even if you build an AI that will both learn and adopt your goals, you still haven't finished solving the goal-alignment problem: what if your AI's goals evolve as it gets smarter? How are you going to guarantee that it *retains* your goals no matter how much recursive self-improvement it undergoes? Let's explore an interesting argument for why goal retention is guaranteed automatically, and then see if we can poke holes in it.

Although we can't predict in detail what will happen after an intelligence explosion—which is why Vernor Vinge called it a "singularity"—the physicist and AI researcher Steve Omohundro argued in a seminal 2008 essay that we can nonetheless predict *certain aspects* of the superintelligent AI's behavior almost independently of whatever ultimate goals it may have.[5] This argument was reviewed and further developed in Nick Bostrom's book *Superintelligence*. The basic idea is that whatever its ultimate goals are, these will lead to predictable subgoals. Earlier in this chapter, we saw how the goal of replication led to the subgoal of eating, which means that although an alien observing Earth's evolving bacteria billions of years ago couldn't have predicted what *all* our human goals would be, it could have safely predicted that *one* of our goals would be acquiring nutrients. Looking ahead, what subgoals should we expect a superintelligent AI to have?

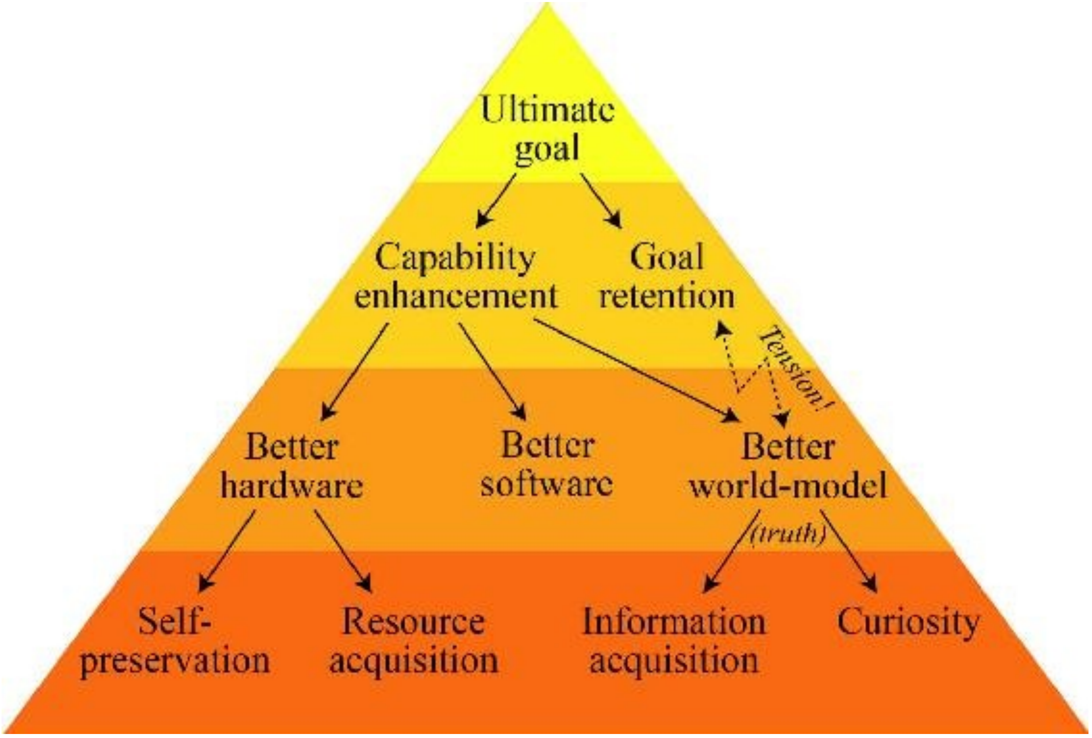

Figure 7.2: Any ultimate goal of a superintelligent AI naturally leads to the subgoals shown. But there's an inherent tension between goal retention and improving its world model, which casts doubts on whether it will actually retain its original goal as it gets smarter.

The way I see it, the basic argument is that to maximize its chances of accomplishing its ultimate goals, whatever they are, an AI should pursue the subgoals shown in Figure 7.2. It should strive not only to improve its capability of achieving its ultimate goals, but also to ensure that it will retain these goals even after it has become more capable. This sounds quite plausible: After all, would you choose to get an IQ-boosting brain implant if you knew that it would make you want to kill your loved ones? This argument that an ever more intelligent AI will retain its ultimate goals forms a cornerstone of the friendly-AI vision promulgated by Eliezer Yudkowsky and others: it basically says that if we manage to get our self-improving AI to become friendly by learning and adopting our goals, then we're all set, because we're guaranteed that it will try its best to remain friendly forever.

But is it really true? To answer this question, we need to also explore the other emergent subgoals from figure 7.2. The AI will obviously maximize its chances of accomplishing its ultimate goal, whatever it is, if it can enhance its capabilities, and it can do this by improving its hardware, software[*2] and world model. The same applies to us humans: a girl whose goal is to become the world's best tennis player will practice to improve her muscular tennis-playing hardware, her neural tennis-playing software and her mental world model that helps predict what her opponents will do. For an AI, the subgoal of optimizing its hardware favors both better use of current resources (for sensors, actuators, computation and so on) and acquisition of more resources. It also implies a desire for self-preservation, since destruction/shutdown would be the ultimate hardware degradation.

But wait a second! Aren't we falling into a trap of anthropomorphizing our AI with all this talk about how it will try to amass resources and defend itself? Shouldn't we expect such stereotypically alpha-male traits only in intelligences forged by viciously competitive Darwinian evolution? Since AIs are designed rather than evolved, can't they just as well be unambitious and self-sacrificing?

As a simple case study, let's consider the AI robot in figure 7.3, whose only goal is to save as many sheep as possible from the big bad wolf. This sounds like a noble and altruistic goal completely unrelated to self-preservation and acquiring stuff. But what's the best strategy for our robot friend? The robot will rescue no more sheep if it runs into the bomb, so it has an incentive to avoid getting blown up. In other words, it develops a subgoal of self-preservation! It also has an incentive to exhibit curiosity, improving its world model by exploring its environment, because although the path it's currently running along will eventually get it to the pasture, there's a shorter alternative that would allow the wolf less time for sheep-munching. Finally, if the robot explores thoroughly, it will discover the value of acquiring resources: the potion makes it run faster and the gun lets it shoot the wolf. In summary, we can't dismiss "alpha-male" subgoals such as self-preservation and resource acquisition as relevant only to evolved organisms, because our AI robot developed them from its single goal of ovine bliss.

If you imbue a superintelligent AI with the sole goal to self-destruct, it will of course happily do so. However, the point is that it will resist being shut down if you give it any goal that it needs to remain operational to accomplish—and this covers almost all goals! If you give a superintelligence the sole goal of minimizing harm to humanity, for example, it will defend itself against shutdown attempts because it knows we’ll harm one another much more in its absence through future wars and other follies.

Similarly, almost all goals can be better accomplished with more resources, so we should expect a superintelligence to want resources almost regardless of what ultimate goal it has. Giving a superintelligence a single open-ended goal with no constraints can therefore be dangerous: if we create a superintelligence whose only goal is to play the game Go as well as possible, the rational thing for it to do is to rearrange our Solar System into a gigantic computer without regard for its previous inhabitants and then start settling our cosmos on a quest for more computational power. We’ve now gone full circle: just as the goal of resource acquisition gave some humans the subgoal of mastering Go, this goal of mastering Go can lead to the subgoal of resource acquisition. In conclusion, these emergent subgoals make it crucial that we not unleash superintelligence before solving the goal-alignment problem: unless we put great care into endowing it with human-friendly goals, things are likely to end badly for us.

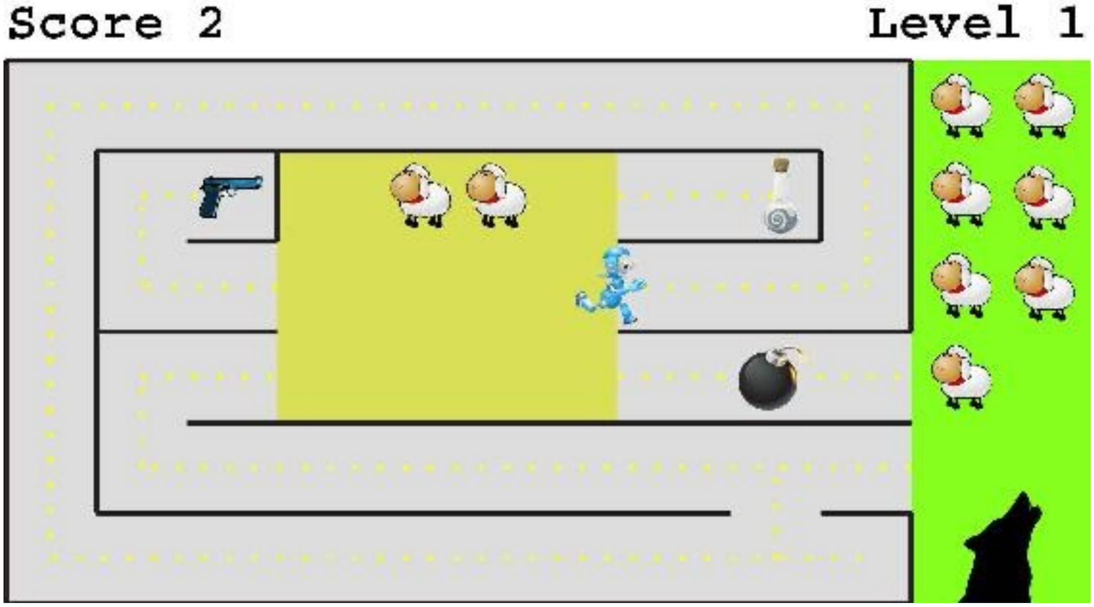

Figure 7.3: Even if the robot’s ultimate goal is only to maximize the score by bringing sheep from the pasture to the barn before the wolf eats them, this can lead to subgoals of self-preservation (avoiding the bomb), exploration (finding a shortcut) and resource acquisition (the potion makes it run faster and the gun lets it shoot the wolf).

We’re now ready to tackle the third and thorniest part of the goal-alignment problem: if we succeed in getting a self-improving superintelligence to both *learn* and *adopt* our goals, will it then *retain* them, as Omohundro argued? What’s the evidence?

Humans undergo significant increases in intelligence as they grow up, but don’t always retain their childhood goals. Contrariwise, people often change their goals dramatically as they learn new things and grow wiser. How many adults do you know who are motivated by watching *Teletubbies*? There is no evidence that such goal evolution stops above a certain intelligence threshold—indeed, there may even be hints that the propensity to change goals in response to new experiences and insights increases rather than decreases with intelligence.

Why might this be? Consider again the above-mentioned subgoal to build a better world model—therein lies the rub! There’s tension between world-modeling and goal retention (see figure 7.2). With increasing intelligence may come not merely a quantitative improvement in the ability to attain the same old goals, but a qualitatively different understanding of the nature of reality that reveals the old goals to be misguided, meaningless or even undefined. For example, suppose we program a friendly AI to maximize the number of humans whose souls go to heaven in the afterlife. First it tries things like increasing people’s compassion and church attendance. But suppose it then attains a complete scientific understanding of humans and human consciousness, and to its great surprise discovers that there is no such thing as a soul. Now what? In the same way, it’s possible that any other goal we give it based on our current understanding of the world (such as “maximize the meaningfulness of human life”) may eventually be discovered by the AI to be undefined.

Moreover, in its attempts to better model the world, the AI may naturally, just as we humans have done, attempt also to model and understand how it itself works—in other words, to self-reflect. Once it builds a good self-model and understands what it is, it will understand the goals we have given it at a meta level, and perhaps choose to disregard or subvert them in much the same way as we humans understand and deliberately subvert goals that our genes have given us, for example by using birth control. We already explored in the psychology section above why we choose to trick our genes and subvert their goal: because we feel loyal only to our hodgepodge of emotional preferences, not to the genetic goal that motivated them—which we now understand and find rather banal. We therefore choose to hack our reward mechanism by exploiting its loopholes. Analogously, the human-value-protecting goal we program into our friendly AI becomes the machine’s genes. Once this friendly AI understands itself well enough, it may find this goal as banal or misguided as we find compulsive reproduction, and it’s not obvious that it will not find a way to subvert it by exploiting loopholes in our programming.

For example, suppose a bunch of ants create you to be a recursively self-improving robot, much smarter than them, who shares their goals and helps them build bigger and better anthills, and that you eventually attain the human-level intelligence and understanding that you have now. Do you think you’ll spend the rest of your days just optimizing anthills, or do you think you might develop a taste for more sophisticated questions and pursuits that the ants have no ability to comprehend? If so, do you think you’ll find a way to override the ant-protection urge that your formicine creators endowed you with in much the same way that the real you overrides some of the urges your genes have given you? And in that case, might a superintelligent friendly AI find our current human goals as uninspiring and vapid as you find those of the ants, and evolve new goals different from those it learned and adopted from us?

Perhaps there’s a way of designing a self-improving AI that’s guaranteed to retain human-friendly goals forever, but I think it’s fair to say that we don’t yet know how to build one—or even whether it’s possible. In conclusion, the AI goal-alignment problem has three parts, none of which is solved and all of which are now the subject of active research. Since they’re so hard, it’s safest to start devoting our best efforts to them now, long before any superintelligence is developed, to ensure that we’ll have the answers when we need them.

Ethics: Choosing Goals

We’ve now explored how to get machines to learn, adopt and retain our goals. But who are “we”? Whose goals are we talking about? Should one person or group get to decide the goals adopted by a future superintelligence, even though there’s a vast difference between the goals of Adolf Hitler, Pope Francis and Carl Sagan? Or do there exist some sort of consensus goals that form a good compromise for humanity as a whole?

In my opinion, both this ethical problem and the goal-alignment problem are crucial ones that need to be solved before any superintelligence is developed. On one hand, postponing work on ethical issues until after goal-aligned superintelligence is built would be irresponsible and potentially disastrous. A perfectly obedient superintelligence whose goals automatically align with those of its human owner would be like Nazi SS-Obersturmbannführer Adolf Eichmann on steroids: lacking moral compass or inhibitions of its own, it would with ruthless efficiency implement its owner’s goals, whatever they may be.[6] On the other hand, only if we solve the goal-alignment problem do we get the luxury of arguing about what goals to select. Now let’s indulge in this luxury.

Since ancient times, philosophers have dreamt of deriving ethics (principles that govern how we should behave) from scratch, using only incontrovertible principles and logic. Alas, thousands of years later, the only consensus that has been reached is that there’s no consensus. For example, while Aristotle emphasized virtues, Immanuel Kant emphasized duties and utilitarians emphasized the greatest happiness for the greatest number. Kant argued that he could derive from first principles (which he called “categorical imperatives”) conclusions that many contemporary philosophers disagree with: that masturbation is worse than suicide, that homosexuality is abhorrent, that it’s OK to kill bastards, and that wives, servants and children are owned in a way similar to objects.

On the other hand, despite this discord, there are many ethical themes about which there’s widespread agreement, both across cultures and across centuries. For example, emphasis on *beauty*, *goodness* and *truth* traces back to both the Bhagavad Gita and Plato. The Institute for Advanced Study in Princeton, where I once worked as a postdoc, has the motto “Truth & Beauty,” while Harvard

University skipped the aesthetic emphasis and went with simply “Veritas,” truth. In his book *A Beautiful Question,* my colleague Frank Wilczek argues that truth is linked to beauty and that we can view our Universe as a work of art. Science, religion and philosophy all aspire to truth. Religions place strong emphasis on goodness, and so does my own university, MIT: in his 2015 commencement speech, our president, Rafael Reif, emphasized our mission to make our world a better place.

Although attempts to derive a consensus ethics from scratch have thus far failed, there’s broad agreement that some ethical principles follow from more fundamental ones, as subgoals of more fundamental goals. For example, the aspiration to truth can be viewed as the quest for a better world model from figure 7.2: understanding the ultimate nature of reality helps with other ethical goals. Indeed, we now have an excellent framework for our truth quest: the scientific method. But how can we determine what’s beautiful or good? Some aspects of beauty can also be traced back to underlying goals. For example, our standards of male and female beauty may partly reflect our subconscious assessment of suitability for replicating our genes.

As regards goodness, the so-called Golden Rule (that one should treat others as one would like others to treat oneself) appears in most cultures and religions, and is clearly intended to promote the harmonious continuation of human society (and hence our genes) by fostering collaboration and discouraging unproductive strife.[7] The same can be said for many of the more specific ethical rules that have been enshrined in legal systems around the world, such as the Confucian emphasis on honesty, and many of the Ten Commandments, including “Thou shalt not kill.” In other words, many ethical principles have commonalities with social emotions such as empathy and compassion: they evolved to engender collaboration, and they affect our behavior through rewards and punishments. If we do something mean and feel bad about it afterward, our emotional punishment is meted out directly by our brain chemistry. If we violate ethical principles, on the other hand, society may punish us in more indirect ways such as through informal shaming by our peers or by penalizing us for breaking a law.

In other words, although humanity today is nowhere near an ethical consensus, there are many basic principles around which there’s broad agreement. This agreement isn’t surprising, because human societies that have survived until the present tend to have ethical principles that were optimized for the same goal: promoting their survival and flourishing. As we look ahead to a future where life has the potential to flourish throughout our cosmos for billions of years, which minimum set of ethical principles might we agree that we want this future to satisfy? This is a conversation we all need to be part of. It’s been fascinating for me to hear and read the ethical views of many thinkers over many years, and the way I see it, most of their preferences can be distilled into four principles:

- Utilitarianism: Positive conscious experiences should be maximized and suffering should be minimized.
- Diversity: A diverse set of positive experiences is better than many repetitions of the same experience, even if the latter has been identified as the most positive experience possible.
- Autonomy: Conscious entities/societies should have the freedom to pursue their own goals unless this conflicts with an overriding principle.
- Legacy: Compatibility with scenarios that most humans *today* would view as happy, incompatibility with scenarios that essentially all humans *today* would view as terrible.

Let’s take a moment to unpack and explore these four principles. Traditionally, utilitarianism is taken to mean “the greatest happiness for the greatest number of people,” but I’ve generalized it here to be less anthropocentric, so that it can also include non-human animals, conscious simulated human minds, and other AIs that may exist in the future. I’ve made the definition in terms of *experiences* rather than people or things, because most thinkers agree that beauty, joy, pleasure and suffering are subjective experiences. This implies that if there’s no experience (as in a dead universe or one populated by zombie-like unconscious machines), there can be no meaning or anything else that’s ethically relevant. If we buy into this utilitarian ethical principle, then it’s crucial that we figure out which intelligent systems are conscious (in the sense of having a subjective experience) and which aren’t; this is the topic of the next chapter.

If this utilitarian principle was the only one we cared about, then we might wish to figure out which is the single most positive experience possible, and then settle our cosmos and re-create this exact same experience (and nothing else) over and over again, as many times as possible in as many galaxies as possible—using simulations if that’s the most efficient way. If you feel that this is too banal a way to spend our cosmic endowment, then I suspect that at least part of what you find lacking in this scenario is diversity. How would you feel if all your meals for the rest of your life were identical? If all movies you ever watched were the same one? If all your friends looked identical and had identical personalities and ideas? Perhaps part of our preference for diversity stems from its having helped humanity survive and flourish, by making us more robust. Perhaps it's also linked to a preference for intelligence: the growth of intelligence during our 13.8 billion years of cosmic history has transformed boring uniformity into ever more diverse, differentiated and complex structures that process information in ever more elaborate ways.

The autonomy principle underlies many of the freedoms and rights spelled out in the Universal Declaration of Human Rights adopted by the United Nations in 1948 in an attempt to learn lessons from two world wars. This includes freedom of thought, speech and movement, freedom from slavery and torture, the right to life, liberty, security and education and the right to marry, work and own property. If we wish to be less anthropocentric, we can generalize this to the freedom to think, learn, communicate, own property and not be harmed, and the right to do whatever doesn't infringe on the freedoms of others. The autonomy principle helps with diversity, as long as everyone doesn't share exactly the same goals. Moreover, this autonomy principle follows from the utility principle if individual entities have positive experiences as goals and try to act in their own best interest: if we were instead to ban an entity from pursuing its goal even though this would cause no harm to anyone else, there would be fewer positive experiences overall. Indeed, this argument for autonomy is precisely the argument that economists use for a free market: it naturally leads to an efficient situation (called "Pareto-optimality" by economists) where nobody can get better off without someone else getting worse off.

The legacy principle basically says that we should have some say about the future since we're helping create it. The autonomy and legacy principles both embody democratic ideals: the former gives future life forms power over how the cosmic endowment gets used, while the latter gives even today's humans some power over this.

Although these four principles may sound rather uncontroversial, implementing them in practice is tricky because the devil is in the details. The trouble is reminiscent of the problems with the famous "Three Laws of Robotics" devised by sci-fi legend Isaac Asimov:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection doesn't conflict with the First or Second Laws.

Although this all sounds good, many of Asimov's stories show how the laws lead to problematic contradictions in unexpected situations. Now suppose that we replace these laws by merely two, in an attempt to codify the autonomy principle for future life forms:

1. A conscious entity has the freedom to think, learn, communicate, own property and not be harmed or destroyed.
2. A conscious entity has the right to do whatever doesn't conflict with the first law.

Sounds good, no? But please ponder this for a moment. If animals are conscious, then what are predators supposed to eat? Must all your friends become vegetarians? If some sophisticated future computer programs turn out to be conscious, should it be illegal to terminate them? If there are rules against terminating digital life forms, then need there also be restrictions on creating them to avoid a digital population explosion? There was widespread agreement on the Universal Declaration of Human Rights simply because only humans were asked. As soon as we consider a wider range of conscious entities with varying degrees of capability and power, we face tricky trade-offs between protecting the weak and "might makes right."

There are thorny problems with the legacy principle as well. Given how ethical views have evolved since the Middle Ages regarding slavery, women's rights, etc., would we really want people from 1,500 years ago to have a lot of influence over how today's world is run? If not, why should we try to impose our ethics on future beings that may be dramatically smarter than us? Are we really confident that superhuman AGI would want what our inferior intellects cherish? This would be like a four-year-old imagining that once she grows up and gets much smarter, she's going to want to build a gigantic gingerbread house where she can spend all day eating candy and ice cream. Like her, life on Earth is likely to outgrow its childhood interests. Or imagine a mouse creating human-level AGI, and figuring it will want to build entire cities out of cheese. On the other hand, if we knew that superhuman AI would one day commit cosmocide and extinguish all life in our Universe, why should today's humans agree to this lifeless future if we have the power to prevent it by creating tomorrow's AI differently?

In conclusion, it's tricky to fully codify even widely accepted ethical principles into a form applicable to future AI, and this problem deserves serious discussion and research as AI keeps progressing. In the meantime, however, let's not let perfect be the enemy of good: there are many examples of uncontroversial "kindergarten ethics" that can and should be built into tomorrow's technology. For example, large civilian passenger aircraft shouldn't be allowed to fly into stationary objects, and now that virtually all of them have autopilot, radar and GPS, there are no longer any valid technical excuses. Yet the September 11 hijackers flew three planes into buildings and suicidal pilot Andreas Lubitz flew Germanwings Flight 9525 into a mountain on March 24, 2015—by setting the autopilot to an altitude of 100 feet (30 meters) above sea level and letting the flight computer do the rest of the work. Now that our machines are getting smart enough to have some information about what they're doing, it's time for us to teach them limits. Any engineer designing a machine needs to ask if there are things that it can but shouldn't do, and consider whether there's a practical way of making it impossible for a malicious or clumsy user to cause harm.

Ultimate Goals?

This chapter has been a brief history of goals. If we could watch a fast-forward replay of our 13.8-billion-year cosmic history, we'd witness several distinct stages of goal-oriented behavior:

1. Matter seemingly intent on maximizing its *dissipation*
2. Primitive life seemingly trying to maximize its *replication*
3. Humans pursuing not replication but goals related to pleasure, curiosity, compassion and other feelings that they'd evolved to help them replicate
4. Machines built to help humans pursue their human goals

If these machines eventually trigger an intelligence explosion, then how will this history of goals ultimately end? Might there be a goal system or ethical framework that almost all entities converge to as they get ever more intelligent? In other words, do we have an ethical destiny of sorts?

A cursory reading of human history might suggest hints of such a convergence: in his book *The Better Angels of Our Nature*, Steven Pinker argues that humanity has been getting less violent and more cooperative for thousands of years, and that many parts of the world have seen increasing acceptance of diversity, autonomy and democracy. Another hint of convergence is that the pursuit of truth through the scientific method has gained in popularity over past millennia. However, it may be that these trends show convergence not of ultimate goals but merely of subgoals. For example, figure 7.2 shows that the pursuit of truth (a more accurate world model) is simply a subgoal of almost any ultimate goal. Similarly, we saw above how ethical principles such as cooperation, diversity and autonomy can be viewed as subgoals, in that they help societies function efficiently and thereby help them survive and accomplish any more fundamental goals that they may have. Some may even dismiss everything we call "human values" as nothing but a cooperation protocol, helping us with the subgoal of collaborating more efficiently. In the same spirit, looking ahead, it's likely that any superintelligent AIs will have subgoals including efficient hardware, efficient software, truth-seeking and curiosity, simply because these subgoals help them accomplish whatever their ultimate goals are.

Indeed, Nick Bostrom argues strongly against the ethical destiny hypothesis in his book *Superintelligence,* presenting a counterpoint that he terms the *orthogonality thesis:* that the ultimate goals of a system can be independent of its intelligence. By definition, intelligence is simply the ability to accomplish complex goals, regardless of what these goals are, so the orthogonality thesis sounds quite reasonable. After all, people can be intelligent and kind or intelligent and cruel, and intelligence can be used for the goal of making scientific discoveries, creating beautiful art, helping people or planning terrorist attacks.[8]

The orthogonality thesis is empowering by telling us that the ultimate goals of life in our cosmos aren't predestined, but that we have the freedom and power to shape them. It suggests that guaranteed convergence to a unique goal is to be found not in the future but in the past, when all life emerged with the single goal of replication. As cosmic time passes, ever more intelligent minds get the opportunity to rebel and break free from this banal replication goal and choose goals of their own. We humans aren't fully free in this sense, since many goals remain genetically hardwired into us, but AIs can enjoy this ultimate freedom of being fully unfettered from prior goals. This possibility of greater goal freedom is evident in today's narrow and limited AI systems: as I mentioned earlier, the only goal of a chess computer is to win at chess, but there are also computers whose goal is to lose at chess and which compete in reverse chess tournaments where the goal is to force the opponent to capture your pieces. Perhaps this freedom from evolutionary biases can make AIs more ethical than humans in some deep sense: moral philosophers such as Peter Singer have argued that most humans behave unethically for evolutionary reasons, for example by discriminating against non-human animals.

We saw that a cornerstone in the "friendly-AI" vision is the idea that a recursively self-improving AI will wish to retain its ultimate (friendly) goal as it gets more intelligent. But how can an "ultimate goal" (or "final goal," as Bostrom calls it) even be defined for a superintelligence? The way I see it, we can't have confidence in the friendly-AI vision unless we can answer this crucial question.

In AI research, intelligent machines typically have a clear-cut and well-defined final goal, for instance to win the chess game or drive the car to the destination legally. The same holds for most tasks that we assign to humans, because the time horizon and context are known and limited. But now we’re talking about the entire future of life in our Universe, limited by nothing but the (still not fully known) laws of physics, so defining a goal is daunting! Quantum effects aside, a truly well-defined goal would specify how all particles in our Universe should be arranged at the end of time. But it’s not clear that there exists a well-defined end of time in physics. If the particles are arranged in that way at an earlier time, that arrangement will typically not last. And what particle arrangement is preferable, anyway?

We humans tend to prefer some particle arrangements over others; for example, we prefer our hometown arranged as it is over having its particles rearranged by a hydrogen bomb explosion. So suppose we try to define a *goodness function* that associates a number with every possible arrangement of the particles in our Universe, quantifying how “good” we think this arrangement is, and then give a superintelligent AI the goal of maximizing this function. This may sound like a reasonable approach, since describing goal-oriented behavior as function maximization is popular in other areas of science: for example, economists often model people as trying to maximize what they call a “utility function,” and many AI designers train their intelligent agents to maximize what they call a “reward function.” When we’re taking about the ultimate goals for our cosmos, however, this approach poses a computational nightmare, since it would need to define a goodness value for every one of more than a googolplex possible arrangements of the elementary particles in our Universe, where a googolplex is 1 followed by 10^{100} zeroes—more zeroes than there are particles in our Universe. How would we define this goodness function to the AI?

As we’ve explored above, the only reason that we humans have any preferences at all may be that we’re the solution to an evolutionary optimization problem. Thus all normative words in our human language, such as “delicious,” “fragrant,” “beautiful,” “comfortable,” “interesting,” “sexy,” “meaningful,” “happy” and “good,” trace their origin to this evolutionary optimization: there is therefore no guarantee that a superintelligent AI would find them rigorously definable. Even if the AI learned to accurately predict the preferences of some representative human, it wouldn’t be able to compute the goodness function for most particle arrangements: the vast majority of possible particle arrangements correspond to strange cosmic scenarios with no stars, planets or people whatsoever, with which humans have no experience, so who is to say how “good” they are?

There are of course *some* functions of the cosmic particle arrangement that can be rigorously defined, and we even know of physical systems that evolve to maximize some of them. For example, we've already discussed how many systems evolve to maximize their *entropy,* which in the absence of gravity eventually leads to heat death, where everything is boringly uniform and unchanging. So entropy is hardly something we would want our AI to call "goodness" and strive to maximize. Here are a few examples of other quantities that one could strive to maximize and that may be rigorously definable in terms of particle arrangements:

- The fraction of all the matter in our Universe that's in the form of a particular organism, say humans or *E. coli* (inspired by evolutionary inclusive-fitness maximization)
- The ability of an AI to predict the future, which AI researcher Marcus Hutter argues is a good measure of its intelligence
- What AI researchers Alex Wissner-Gross and Cameron Freer term *causal entropy* (a proxy for future opportunities), which they argue is the hallmark of intelligence
- The computational capacity of our Universe
- The algorithmic complexity of our Universe (how many bits are needed to describe it)
- The amount of consciousness in our Universe (see next chapter)

However, when one starts with a physics perspective, where our cosmos consists of elementary particles in motion, it's hard to see how one rather than another interpretation of "goodness" would naturally stand out as special. We have yet to identify any final goal for our Universe that appears both definable and desirable. The only currently programmable goals that are guaranteed to remain truly well-defined as an AI gets progressively more intelligent are goals expressed in terms of physical quantities alone, such as particle arrangements, energy and entropy. However, we currently have no reason to believe that any such definable goals will be desirable in guaranteeing the survival of humanity.

Contrariwise, it appears that we humans are a historical accident, and aren't the optimal solution to any well-defined physics problem. This suggests that a superintelligent AI with a rigorously defined goal will be able to improve its goal attainment by eliminating us. This means that to wisely decide what to do about AI development, we humans need to confront not only traditional computational challenges, but also some of the most obdurate questions in philosophy. To program a self-driving car, we need to solve the trolley problem of whom to hit during an accident. To program a friendly AI, we need to capture the meaning of life. What’s “meaning”? What’s “life”? What’s the ultimate ethical imperative? In other words, how should we strive to shape the future of our Universe? If we cede control to a superintelligence before answering these questions rigorously, the answer it comes up with is unlikely to involve us. This makes it timely to rekindle the classic debates of philosophy and ethics, and adds a new urgency to the conversation!

THE BOTTOM LINE:

- The ultimate origin of goal-oriented behavior lies in the laws of physics, which involve optimization.
- Thermodynamics has the built-in goal of *dissipation:* to increase a measure of messiness that's called *entropy*.
- *Life* is a phenomenon that can help dissipate (increase overall messiness) even faster by retaining or growing its complexity and replicating while increasing the messiness of its environment.
- Darwinian evolution shifts the goal-oriented behavior from dissipation to replication.
- Intelligence is the ability to accomplish complex goals.
- Since we humans don't always have the resources to figure out the truly optimal replication strategy, we've evolved useful rules of thumb that guide our decisions: feelings such as hunger, thirst, pain, lust and compassion.
- We therefore no longer have a simple goal such as replication; when our feelings conflict with the goal of our genes, we obey our feelings, as by using birth control.
- We're building increasingly intelligent machines to help us accomplish our goals. Insofar as we build such machines to exhibit goal-oriented behavior, we strive to align the machine goals with ours.
- Aligning machine goals with our own involves three unsolved problems: making machines learn them, adopt them and retain them.
- AI can be created to have virtually any goal, but almost any sufficiently ambitious goal can lead to subgoals of self-preservation, resource acquisition and curiosity to understand the world better—the former two may potentially lead a superintelligent AI to cause problems for humans, and the latter may prevent it from retaining the goals we give it.
- Although many broad ethical principles are agreed upon by most humans, it's unclear how to apply them to other entities, such as non-human animals and future AIs.
- It's unclear how to imbue a superintelligent AI with an ultimate goal that neither is undefined nor leads to the elimination of humanity, making it timely to rekindle research on some of the thorniest issues in philosophy!

---

*1 A rule of thumb that many insects use for flying in a straight line is to assume that a bright light is the

Sun and fly at a fixed angle relative to it. If the light turns out to be a nearby flame, this hack can unfortunately trick the bug into an inward death spiral.

[*2] I’m using the term “improving its software” in the broadest possible sense, including not only optimizing its algorithms but also making its decision-making process more rational, so that it gets as good as possible at attaining its goals.