Chapter 2

Matter Turns Intelligent

Hydrogen…, given enough time, turns into people.

Edward Robert Harrison, 1995

One of the most spectacular developments during the 13.8 billion years since our Big Bang is that dumb and lifeless matter has turned intelligent. How could this happen and how much smarter can things get in the future? What does science have to say about the history and fate of intelligence in our cosmos? To help us tackle these questions, let's devote this chapter to exploring the foundations and fundamental building blocks of intelligence. What does it mean to say that a blob of matter is intelligent? What does it mean to say that an object can remember, compute and learn?

What Is Intelligence?

My wife and I recently had the good fortune to attend a symposium on artificial intelligence organized by the Swedish Nobel Foundation, and when a panel of leading AI researchers were asked to define intelligence, they argued at length without reaching consensus. We found this quite funny: there's no agreement

on what intelligence is even among intelligent intelligence researchers! So there's clearly no undisputed "correct" definition of intelligence. Instead, there are many competing ones, including capacity for logic, understanding, planning, emotional knowledge, self-awareness, creativity, problem solving and learning.

In our exploration of the future of intelligence, we want to take a maximally broad and inclusive view, not limited to the sorts of intelligence that exist so far. That's why the definition I gave in the last chapter, and the way I'm going to use the word throughout this book, is very broad:

> intelligence = *ability to accomplish complex goals*

This is broad enough to include all above-mentioned definitions, since understanding, self-awareness, problem solving, learning, etc. are all examples of complex goals that one might have. It's also broad enough to subsume the *Oxford Dictionary* definition—"the ability to acquire and apply knowledge and skills"—since one can have as a goal to apply knowledge and skills.

Because there are many possible goals, there are many possible types of intelligence. By our definition, it therefore makes no sense to quantify intelligence of humans, non-human animals or machines by a single number such as an IQ.[*1] What's more intelligent: a computer program that can only play chess or one that can only play Go? There's no sensible answer to this, since they're good at different things that can't be directly compared. We can, however, say that a third program is more intelligent than both of the others if it's at least as good as them at accomplishing *all* goals, and strictly better at at least one (winning at chess, say).

It also makes little sense to quibble about whether something is or isn't intelligent in borderline cases, since ability comes on a spectrum and isn't necessarily an all-or-nothing trait. What people have the ability to accomplish the goal of speaking? Newborns? No. Radio hosts? Yes. But what about toddlers who

can speak ten words? Or five hundred words? Where would you draw the line? I've used the deliberately vague word "complex" in the definition above, because it's not very interesting to try to draw an artificial line between intelligence and non-intelligence, and it's more useful to simply quantify the degree of ability for accomplishing different goals.

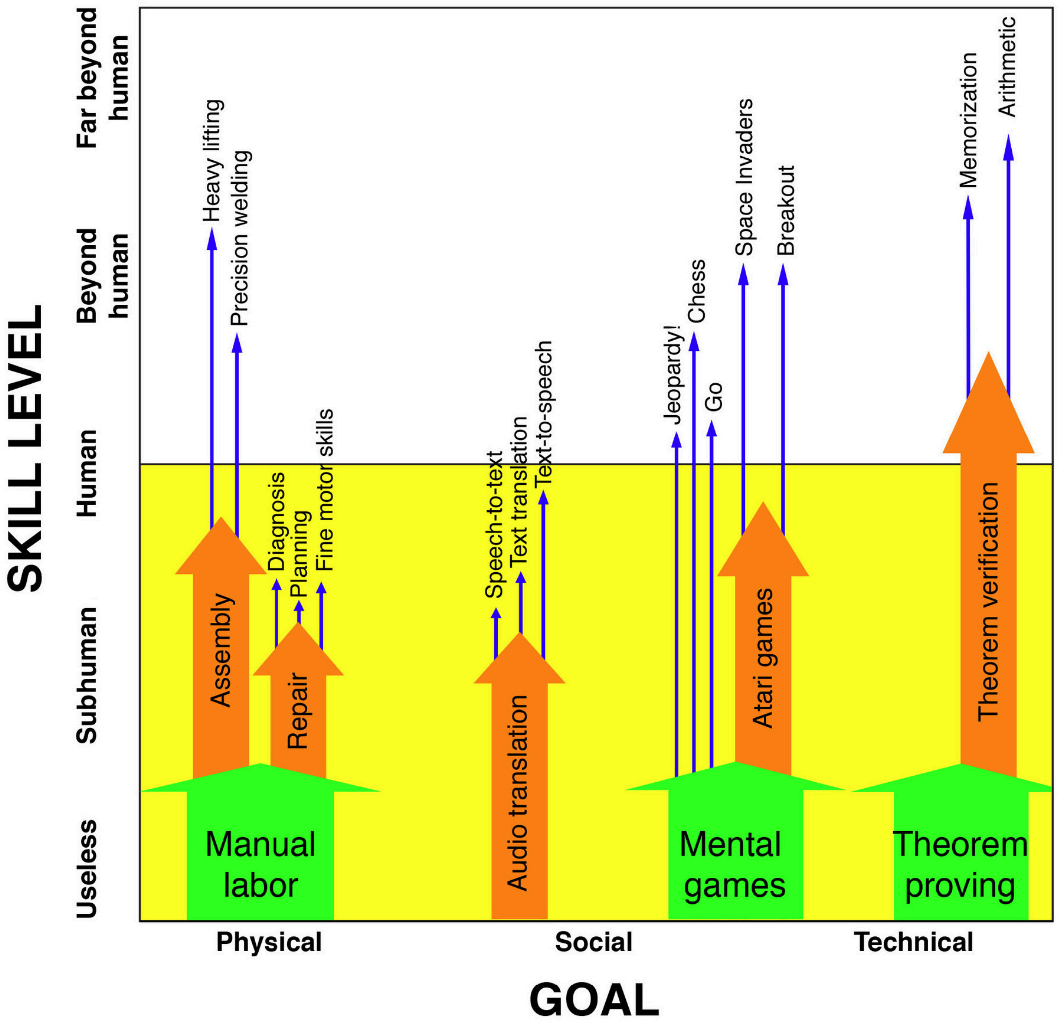

Figure 2.1: Intelligence, defined as ability to accomplish complex goals, can't be measured by a single IQ, only by an ability spectrum across all goals. Each arrow indicates how skilled today's best AI systems are at accomplishing various goals, illustrating that today's artificial intelligence tends to be *narrow,* with each system able to accomplish only very specific goals. In contrast, human intelligence is remarkably broad: a healthy child can learn to get better at almost anything.

To classify different intelligences into a taxonomy, another crucial distinction is that between *narrow* and *broad* intelligence. IBM's Deep Blue chess computer, which dethroned chess champion Garry Kasparov in 1997, was only able

to accomplish the very narrow task of playing chess—despite its impressive hardware and software, it couldn't even beat a four-year-old at tic-tac-toe. The DQN AI system of Google DeepMind can accomplish a slightly broader range of goals: it can play dozens of different vintage Atari computer games at human level or better. In contrast, human intelligence is thus far uniquely broad, able to master a dazzling panoply of skills. A healthy child given enough training time can get fairly good not only at *any* game, but also at any language, sport or vocation. Comparing the intelligence of humans and machines today, we humans win hands-down on breadth, while machines outperform us in a small but growing number of narrow domains, as illustrated in figure 2.1. The holy grail of AI research is to build "general AI" (better known as *artificial general intelligence,* AGI) that is maximally broad: able to accomplish virtually any goal, including learning. We'll explore this in detail in chapter 4. The term "AGI" was popularized by the AI researchers Shane Legg, Mark Gubrud and Ben Goertzel to more specifically mean *human-level* artificial general intelligence: the ability to accomplish any goal at least as well as humans.[1] I'll stick with their definition, so unless I explicitly qualify the acronym (by writing "superhuman AGI," for example), I'll use "AGI" as shorthand for "human-level AGI."[*2]

Although the word "intelligence" tends to have positive connotations, it's important to note that we're using it in a completely value-neutral way: as ability to accomplish complex goals regardless of whether these goals are considered good or bad. Thus an intelligent person may be very good at helping people or very good at hurting people. We'll explore the issue of goals in chapter 7. Regarding goals, we also need to clear up the subtlety of whose goals we're referring to. Suppose your future brand-new robotic personal assistant has no goals whatsoever of its own, but will do whatever you ask it to do, and you ask it to cook the perfect Italian dinner. If it goes online and researches Italian dinner recipes, how to get to the closest supermarket, how to strain pasta and so on, and then successfully buys the ingredients and prepares a succulent meal, you'll presumably consider it intelligent even though the original goal was yours. In fact, it adopted your goal once you'd made your request, and then broke it into a

hierarchy of subgoals of its own, from paying the cashier to grating the Parmesan. In this sense, intelligent behavior is inexorably linked to goal attainment.

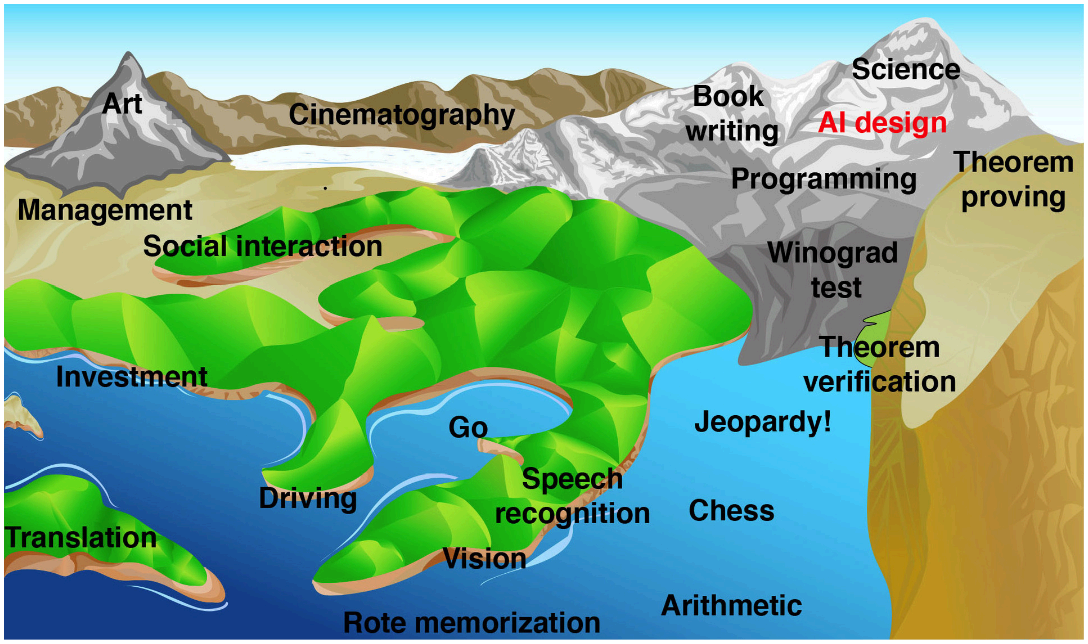

Figure 2.2: Illustration of Hans Moravec's "landscape of human competence," where elevation represents difficulty for computers, and the rising sea level represents what computers are able to do.

It's natural for us to rate the difficulty of tasks relative to how hard it is for us humans to perform them, as in figure 2.1. But this can give a misleading picture of how hard they are for computers. It feels much harder to multiply 314,159 by 271,828 than to recognize a friend in a photo, yet computers creamed us at arithmetic long before I was born, while human-level image recognition has only recently become possible. This fact that low-level sensorimotor tasks seem easy despite requiring enormous computational resources is known as Moravec's paradox, and is explained by the fact that our brain makes such tasks feel easy by dedicating massive amounts of customized hardware to them—more than a quarter of our brains, in fact.

I love this metaphor from Hans Moravec, and have taken the liberty to illustrate it in figure 2.2:

> Computers are universal machines, their potential extends uniformly over a boundless expanse of tasks. Human potentials, on the other hand, are strong in areas long important for survival, but weak in things far removed. Imagine a "landscape of human competence," having lowlands with labels like "arithmetic" and "rote memorization," foothills like "theorem proving" and "chess playing," and high mountain peaks labeled "locomotion," "hand-eye coordination" and "social interaction." Advancing computer performance is like water slowly flooding the landscape. A half century ago it began to drown the lowlands, driving out human calculators and record clerks, but leaving most of us dry. Now the flood has reached the foothills, and our outposts there are contemplating retreat. We feel safe on our peaks, but, at the present rate, those too will be submerged within another half century. I propose that we build Arks as that day nears, and adopt a seafaring life![2]

During the decades since he wrote those passages, the sea level has kept rising relentlessly, as he predicted, like global warming on steroids, and some of his foothills (including chess) have long since been submerged. What comes next and what we should do about it is the topic of the rest of this book.

As the sea level keeps rising, it may one day reach a tipping point, triggering dramatic change. This critical sea level is the one corresponding to machines becoming able to perform AI design. Before this tipping point is reached, the sea-level rise is caused by *humans* improving machines; afterward, the rise can be driven by *machines* improving machines, potentially much faster than humans could have done, rapidly submerging all land. This is the fascinating and controversial idea of the *singularity,* which we'll have fun exploring in chapter 4.

Computer pioneer Alan Turing famously proved that if a computer can perform a certain bare minimum set of operations, then, given enough time and memory, it can be programmed to do anything that *any* other computer can do. Machines exceeding this critical threshold are called *universal computers* (aka Turing-universal computers); all of today's smartphones and laptops are universal in this

sense. Analogously, I like to think of the critical intelligence threshold required for AI design as the threshold for *universal intelligence:* given enough time and resources, it can make itself able to accomplish any goal as well as *any* other intelligent entity. For example, if it decides that it wants better social skills, forecasting skills or AI-design skills, it can acquire them. If it decides to figure out how to build a robot factory, then it can do so. In other words, universal intelligence has the potential to develop into Life 3.0.

The conventional wisdom among artificial intelligence researchers is that intelligence is ultimately all about information and computation, not about flesh, blood or carbon atoms. This means that there's no fundamental reason why machines can't one day be at least as intelligent as us.

But what are information and computation really, given that physics has taught us that, at a fundamental level, everything is simply matter and energy moving around? How can something as abstract, intangible and ethereal as information and computation be embodied by tangible physical stuff? In particular, how can a bunch of dumb particles moving around according to the laws of physics exhibit behavior that we'd call intelligent?

If you feel that the answer to this question is obvious and consider it plausible that machines might get as intelligent as humans this century—for example because you're an AI researcher—please skip the rest of this chapter and jump straight to chapter 3. Otherwise, you'll be pleased to know that I've written the next three sections specially for you.

What Is Memory?

If we say that an atlas contains *information* about the world, we mean that there's a relation between the state of the book (in particular, the positions of certain molecules that give the letters and images their colors) and the state of the world (for example, the locations of continents). If the continents were in different places, then those molecules would be in different places as well. We humans use a panoply of different devices for storing information, from books and brains to