ARTIFICIAL INTELLIGENCE: THE SCIENCE OF COGNITIVE AGENTS

CHAPTER ONE: Introduction

This is a book-in-progress, based largely on evolving lecture notes that I have tended to re-prepare each time I teach CMSC421 – Intro to AI – at the University of Maryland. I keep searching for a way to introduce the vast territory of AI in a way that does it some justice, maintains student interest, and also is true to what I see as the excitement and underlying scientific theme to AI. But the field has grown so much that most of my efforts in the past have resulted in more-or-less a grab-bag of topics only very loosely connected.

Back in, say, 1980, it was perhaps just barely possible for one person to have a pretty good grasp of all of AI, and also possible to give an introductory one-semester overview course that gave students a reasonably good sense of the field as a whole.

That was then, this is now.

Specialization has set in, big time. Many individual subfields – machine learning being the current most prominent example – now appear to involve more researchers (and lots more research dollars) than the entire field a few decades ago. As a result, most researchers also specialize. And when we teach an intro course, we tend to emphasize the parts we know best. As a result, CMSC421 is really more like half a dozen courses, depending on who is teaching it. Yet I keep resolving to do something about this, at least when I teach it. But how? Covering – even briefly – all the many subfields would trivialize the course.

The title above gives my solution: to present AI as it was originally envisioned. The scientific aim of AI is to reveal the computational nature of mind. This will not be without controversy; and that is fine. Feel free to take it as my opinion. In any event, a great deal of AI is motivated by that aim, and it even has some names, such as Big AI, Artificial General Intelligence, Strong AI, and the AI Problem.

Doing so, however – at least the way I envision it – requires weaving back and forth among AI, neuroscience, cognitive psychology, linguistics, and philosophy. It would be perilous to attempt to understand the mind while ignoring most of what has been done and thought already. Still, our focus will be on computational insights, and in particular (as emphasized in the next chapter) on ways to try to *build* a cognitive agent. But we should take care to have an open mind here: it often happens that approaches that one might have thought irrelevant end up being right on target; this has happened in particular a number of times in my own work.

But then the problem of an intro course becomes even more baffling: how to cover AI *and* the rest of cognitive science? I confess to being a bit ruthless in choosing what to leave out, and also to playing favorites a bit too.

Why take on the aim to understand the mind? One very worthy reason is to find better treatments for mental illness. But there are other compelling reasons as well. Achieving such an aim – and again this is an opinion – will change the world far far more than will self-driving cars or household robots. How the mind works is, arguably, the most important thing there is for us to know. Everything we engage in – from education to science to art to love and politics and war and peace – depends crucially on how our minds work. And we are very largely ignorant of that, despite the enormous strides made in the allied cognitive neurosciences (of which AI is a prominent part).

Not only that: there may well be other kinds of minds than those of humans and other animals – maybe minds elsewhere in our galaxy, or future human minds we might grow into. We should not assume human minds of the early 21st century are all they can become. So we might discover not only who we are, but who we can be. This could be anything from better memories and better anger management to whole new realms of being or of what is it to be a self in a world.

So here we are about to embark on an introductory tour of what promises to become the greatest revolution humanity has ever seen. "When will that occur?" you might ask. Ah, that's a toughie. We won't know until we get there, as with any scientific adventure. But you can help us do it! And some of us think that the edges and outlines of the solution may be dimly coming into view.

Why is it so hard? Well, it's the mind-body problem! How can material things (like brains, made of molecules) have thoughts? And apart from that (an admittedly philosophical issue), there is this:

The best laid plans o mice n men
Gang aft agley

That is, the world is a complicated place, even at the level of everyday life; and yet somehow our brains usually manage to get us through each day pretty well. Just how complicated everyday life is, will become apparent as we consider how one might try to build a brain (a robot, if you like) that can do what we do. Our plans often go wrong, but even so usually we avoid disaster and tend to make some useful progress. (We will return to Robert Burns' famous poem much later.)

And so in a way I lied: a general-purpose household robot – if we really could build one to perform reasonably on its own – would be an enormous step forward, and would surely show us a great deal about the nature of intelligence.

So we will begin by rather naively considering what it might take to build such a robot. And that will then lead us into some technical depth. A key idea to keep in mind is what I like to call the Method of Enormous Over-Simplification (or MEOS). It is used (without a name, as far as I am aware) in all of science and even everyday life: simplify what you are pondering into something so simple that you can make progress and then step back, see what is missing, and do it again.

Another comment before we get started: there will be math.

Lots of math creeps into AI here and there. Instead of giving a review at the start, or putting the math in an appendix, we will introduce whatever is needed when it is needed. Just to presage: calculus, linear algebra, mathematical logic, and probability/statistics all will appear. However, if you are rusty on these, or even unfamiliar with some of them, enough will be provided to enable you to keep going (perhaps with some independent review by you if you need it).

A note on organization: Back in the early days of AI, one big success was in the area of AI (heuristic) search; and from then on most intro AI textbooks have started out on that topic. It not only makes some historical sense, but also – since search underlies a great many of the other parts of AI – it makes some conceptual sense. But to me it obscures the essence of AI as the science of cognitive agents. For instance, a major use of search is in automated planning, and planning (in any realistic cognitive sense) requires a great deal of knowledge about the world. Not knowledge in the sense of physics or biochemistry, but in the sense of common sense: things can fall, be close or far, large or small, friend or foe, moving or stationary, be seen or touched or moved, etc. So to me it makes sense to start with commonsense knowledge – both the declarative/factual kind and the procedural/how-to kind. Thus this book is largely organized around getting knowledge (aka learning) and using knowledge (reasoning, planning, deciding, etc), with respect to both kinds. But in addition we will take glimpses of cognitive science, and especially neuroscience, out of which many AI ideas have sprung.

<u>A brief pre-history/sketch of AI</u>

1870s-80s: Golgi and Cajal – the discovery of neurons
1943: McCulloch and Pitts – abstract model of neural processing
1949: Hebb – theory of neural learning
1952: Hodgkin and Huxley – mathematical model of electrochemistry of neural signals

<u>A brief history/sketch of very early AI</u>

1950: Turing – speculations on computing and intelligence
1956: Dartmouth Conference – AI becomes a recognized field of study
1959: McCarthy – paper on “Programs with common sense” (Advice-Taker)

At the Dartmouth Conference were John McCarthy, Marvin Minsky, and eight other luminaries, all of whom largely shaped the course of AI for the many decades.

Ways of surveying the AI scene

We can think of an intelligent system (whatever that might turn out to be) in terms of the following diagram:

Percepts → Agent → Actions

This incredibly simple diagram actually is enormously powerful. (In fact, it can lead in a few short steps to one of the major specialty areas: so-called AI Search. But as noted above, we will not do that here, except for a very brief description a little further on.) The agent in the diagram is supposed to select actions to perform, based on what it perceives and what it already knows; this *knowledge* is hidden inside the agent.

Our diagram prompts questions, such as: what sort of activity is it that the agent does; this then can lead us to classify AI into subspecialties for agent activities, like planning, learning, and natural-language processing.

Or one could ask about the *already-known* stuff, and this prompts further questions:

How does the agent *acquire* what it knows? That is, how can it learn?

How does it *use* what it knows? For instance, how does it make decisions, or carry out effective actions, or come up with plans, solve problems, or even produce more knowledge by inference from current knowledge?

And further: what do we mean by knowledge? Are there different types?

Yes, there are. In fact, one distinction sometimes made in AI, is between systems that are focused on *know-how* and ones focused on *knowledge-that*. This distinction is closely related to a host of other notions that are sometimes clustered into two broad categories that I list here, even though these are rather fuzzy notions and tend to overlap a lot:

Low-level processing knowledge	High-level knowledge
Neural units	Abstract reasoning
Subsymbolic	Symbolic
How-to, skills, abilities	Factual, statement-like

Scruffy (make it work)	Neat (make it logical)
Bottom-up	Top-down
Implicit, distributed, networks	Explicit, rule-based, lists
Procedural	Declarative

The right-hand column tends to make use of a so-called knowledge-base (KB) that can be thought of as a list of statements the agent "believes". But the left-hand column refers to knowledge as well, but of a seeming different sort. I want to lump the two together to some extent, and so I sometimes prefer to refer to the agent's *knowledge-base network* (KBN) for that purpose, referring to the totality of the agent's entire fund of knowledge. There are two reasons for my choice here: Many activities involve both types of knowledge in ways that make it hard to separate out the two types from each other; and – I would argue – knowledge (i.e., its acquisition and use) is the essence of almost all AI research, and hence it deserves to be singled out as such.

Here is an example of the former: Learning French. It surely is a skill, in some ways like learning to swim or to do sums. And yet it clearly is also deeply symbolic, factual, declarative, explicit: *chat* means cat. The use and the meaning of French words cannot be separated very well. And similar observations hold for much of our knowledge.

At any rate, this book is largely structured around the notion of knowledge as the central feature that makes an agent an agent. Our main focus early on will be the acquisition of skill-knowledge, then to knowledge by inference, and eventually to knowledge use, commonsense reasoning, planning, and more.

Wason Selection Task

Here is a more complex example, drawn from cognitive psychology. Suppose you are shown four cards lying flat on a table, and told that each card is either green or red on one side and has either an A or a B on the other. Two of the cards are color-side up and two letter-side up, as shown:

RED GREEN A B

You are asked to check whether all the cards satisfy this rule: *If a card is red on one side then it has an A on the other.* You are also told that you should turn over as few cards as possible to do this.

Surprisingly many people – even trained logicians – tend get this wrong, namely they turn over more than just the RED and B cards. Yet when the problem is "rewritten" in terms of more familiar ideas, everyone gets it right. Here is an example of a rewriting: suppose you work in a post office. Instead of cards there are envelopes, and each one already has been verified to have stamps (either 30-cent or 50-cent) on the address side. The flap-side can either be sealed or unsealed. Your job is to make sure this rule is followed: *If an envelope is sealed then it must have a 50-cent stamp on the front.* Here they are:

SEALED UNSEALED 50-cent stamp 30-cent stamp

Now no one has any trouble is seeing that only the SEALED and the 30-cent envelopes need to be turned over. Yet the problems seem to be formally identical. Since the original studies, a vast number of followup work has been done in an attempt to isolate just what is going on. One prominent suggestion is that people are finely tuned social beings who necessarily attend to whether people are meeting their social obligations (such as paying enough postage).

However, the lessons for us are simply that (i) purely abstract reasoning without meaningful connections is hard to do based on general principles alone (and perhaps pointless!), and (ii) a very great deal of information is present that we might not normally be aware of, such as: language, vision, action possibilities (turning cards over), planning, deciding, what one does and does not know, counting, reasoning, and more (including learning – can you see where learning comes into this?). Constructing a robotic system that can manage (with or without human-style mistakes) the Wason Task based on general principles (i.e., not built just for this one task) would be a very large challenge.

Intermission on AI Search

We will however now take a brief "intermission" to see what we are missing in not jumping into the traditional AI Search topic. Our simple diagram is based on the idea that there is a world W in which the agent operates, getting sensory input (or percepts) from it and carrying out action that result in changes to it. Thus at any one time, W is in some state s, and an action will cause W to go into a different state s'. Suppose we have a list of action-specifications that say, for each action a, what has to be true of the current state s in order for a to be performable (these are the preconditions for a) and how the state s' resulting from performing action a in state s will differ from s (these are the postconditions for a). This immediately leads to a tree structure, with a given "start-state" as the root, allowed actions as the edges from the root with associated new states as the children, more actions as edges from each of those, etc. If there is also a specified "goal state" g, then a program can search starting at s_0 until it finds g. Standard tree-search algorithms such as depth-first

and breadth-first search are clearly applicable. Note however several important aspects here:

1. The tree often is infinite, since it is often the case that there are always actions that can be done.
2. The tree is not given as input to the program. What is given are the start state and goal state and specification of actions and what they accomplish (i.e., what changes to W that they bring about). The program has to construct the tree as it considers actions to perform.
3. There need not be any "world" or "actions" at all, just specifications. So complex issues about sensors and effectors (as in robots) are conveniently ignored.

Some typical examples include the 8-puzzle (like the 15-puzzle but simpler); so-called blocks-world problems (re-stack blocks to achieve a certain pattern); and many game-playing algorithms. But search is central to almost all of AI on one form or another.

One of AI's main contributions to tree search is in the use of heuristics (rules of thumb) that can help guide the search so that it is not blindly looking at every node in a brute force manner. We will however set this aside for now, in fact until we have looked more generally at rules of thumb in the context of commonsense reasoning (which itself will also come much later on).

Getting Knowledge, overview

Returning to our diagram

Percepts → Agent → Actions

and asking once more how the agent can learn (get knowledge), several things can easily be noted. The agent can:

1. observe patterns in the world (rain falls, birds fly, water drips, stones are heavy, tomatoes are red, etc)
2. observe agents accomplish results and try to imitate them
3. be informed in language (by reading, communication, etc)
4. infer (new pieces of knowledge from old)
5. train/practice

The last item above is called *machine learning* even though all five really are forms of learning. We'll start with this last one, which is a huge AI topic on its own; we will take only a cursory look at it in general and then focus a bit on particular methods

closely linked to studies of the brain, especially low-level processing. (Alas, very little is know about high-level processing in the brain.) Items 1, 2, and 3 above we will treat very lightly.

FOOD FOR THOUGHT: Consider ways in which items 1 and 2 overlap with item 5.

Machine Learning (ML)

Machine learning has over the past decade or so become a vast subfield of AI, and itself involves many subspecialties. There are various ways to break these down, and we will give a quick overview now before singling in on one of them.

Supervised learning: Here training examples in the form of pairs (x,y) are used to get the "system" (or agent, but there is as yet little agent-like in this) to predict the y's from x's. That is, the rough aim is to learn a function y = f(x); but usually it is unrealistic to expect an actual function f to be found. Instead, the system is considered successful if it gives close approximations to the y's in a high percentage of cases. A customary procedure is to first train on the given example pairs (of which there could be vast numbers, even millions) and then once training seems successful to "test" the system on new pairs. This general approach is called "supervised" due to the training pairs, as if a teacher is there giving examples to be practiced with before a test. Sometimes in addition there are negative training pairs, e.g., that a given y is NOT to be matched with some x. Learning to recognize handwritten characters is a standard example. Notice that each x can be very complex; for instance in the case of handwritten characters (where each x is such a character and the correct associated y might be the ascii code for the intended character), the x would probably be an image, i.e., a 2-dimensional array of pixels, with possibly hundreds or thousands of bytes for just one such x. Frequently this would be represented as a vector or possibly a matrix or even a tensor. NIST (the National Institute for Standards and Technology) has created MNIST which is an enormous set of training pairs (x,y) of handwritten digit-images x and their associated digits y (0 through 9), together with another large set of test pairs.

Unsupervised learning: Here the system simply looks for natural clusters in the data. In that regard this overlaps with item 1 above. The famous Google-Brain study in which YouTube video images were the input x values, and there were no y's given (since this is unsupervised). The system (Google Brain) eventually "learned" on its own to group the x's into various categories, which the scientists then recognized as being: ones with cats, ones with people, and so on. This was an example of so-called Deep Learning, which we will discuss briefly later on. But note that not all Deep Learning is unsupervised.

Semi-supervised learning: This is what it sounds like. Consider again the character-recognition example. Suppose (unlike NIST) that we do not have the resources to build a very large training set that reasonable covers most ways people might write certain characters. So instead we might create a modest training set of, say, 100

pairs (x,y), train the system on these, and then let it continue to train itself in unsupervised mode to cluster new x's as best it can with respect to what it already has learned.

Reinforcement learning: This is based on the idea from cognitive psychology that an agent can be trained to exhibit a certain behavior if there are suitable rewards and punishments. Here instead of learning pairs or clusters, one learns which actions lead to the best results in given circumstances. Sometimes this is illustrated via a 2-dimensional array of cells among which the agent can move about (left, right, up, down) and note in each cell what reward is there (e.g., a real number, with larger being higher reward, and negative being the opposite).

Having completed this all too brief overview of types of machine learning, in the next chapter we will look a little more closely at supervised learning, and in particular a form of it based on ideas from neuroscience.

Chapter closing:
So we want to know (in some computational sense) what intelligence is, or how a mind can work. Instead of (or in addition to) making philosophical arguments or studying the brain or getting data on human performance, we can try to build a system to do something, and hope that along the way insights – very possibly including how humans do that same something (call it X) – will come along.

We may also find other ways of doing X as well, perhaps faster or more accurately than humans; and also perhaps some Xs that humans can't do at all. In such options lies a great deal of the practical or engineering or technology side of AI. The two sides – scientific and technological – are closely linked, and it is often the case that one (either one) pushes the other along. And while our focus will be more on the scientific side of AI, it is inevitable that both will be much in evidence.

So AI at its best does not ignore the larger picture of cognitive science, but we also try to let our imaginations run free rather than merely imitate existing intelligent being like ourselves. Indeed, we may well come up with new ideas that later are found to be matched in the brain.