Instructor's Manual:
Exercises and Solutions for

Artificial Intelligence
A Modern Approach

Fourth Edition

Stuart J. Russell and Peter Norvig

with contributions from

st Davis, Nicholas J. Hay, Jared Moore, Alex Rudnick, Mehran Sahami, Xiaocheng Mesut Yang, and

EXERCISES 1

INTRODUCTION

Note that for many of the questions in this chapter, we give references where answers can be found rather than writing them out—the full answers would be far too long.

1.1 What Is AI?

Exercise 1.DEFA

Define in your own words: (a) intelligence, (b) artificial intelligence, (c) agent, (d) rationality, (e) logical reasoning.

a. Dictionary definitions of **intelligence** talk about "the capacity to acquire and apply knowledge" or "the faculty of thought and reason" or "the ability to comprehend and profit from experience." These are all reasonable answers, but if we want something quantifiable we would use something like "the ability to act successfully across a wide range of objectives in complex environments."

b. We define **artificial intelligence** as the study and construction of agent programs that perform well in a given class of environments, for a given agent architecture; they *do the right thing*. An important part of that is dealing with the uncertainty of what the current state is, what the outcome of possible actions might be, and what is it that we really desire.

c. We define an **agent** as an entity that takes action in response to percepts from an environment.

d. We define **rationality** as the property of a system which does the "right thing" given what it knows. See Section 2.2 for a more complete discussion. The basic concept is *perfect* rationality; Section **??** describes the impossibility of achieving perfect rationality and proposes an alternative definition.

e. We define **logical reasoning** as the a process of deriving new sentences from old, such that the new sentences are necessarily true if the old ones are true. (Notice that does not refer to any specific syntax or formal language, but it does require a well-defined notion of truth.)

Exercise 1.TURI

Read Turing's original paper on AI (Turing, 1950). In the paper, he discusses several objections to his proposed enterprise and his test for intelligence. Which objections still carry

weight? Are his refutations valid? Can you think of new objections arising from developments since he wrote the paper? In the paper, he predicts that, by the year 2000, a computer will have a 30% chance of passing a five-minute Turing Test with an unskilled interrogator. What chance do you think a computer would have today? In another 25 years?

See the solution for exercise 26.1 for some discussion of potential objections.

The probability of fooling an interrogator depends on just how unskilled the interrogator is. A few entrants in the Loebner prize competitions have fooled judges, although if you look at the transcripts, it looks like the judges were having fun rather than taking their job seriously. There certainly have been examples of a chatbot or other online agent fooling humans. For example, see the description of the Julia chatbot at `www.lazytd.com/lti/julia/`. We'd say the chance today is something like 10%, with the variation depending more on the skill of the interrogator rather than the program. In 25 years, we expect that the entertainment industry (movies, video games, commercials) will have made sufficient investments in artificial actors to create very credible impersonators.

Note that governments and international organizations are seriously considering rules that require AI systems to be identified as such. In California, it is already illegal for machines to impersonate humans in certain circumstances.

Exercise 1.REFL

Are reflex actions (such as flinching from a hot stove) rational? Are they intelligent?

Yes, they are rational, because slower, deliberative actions would tend to result in more damage to the hand. If "intelligent" means "applying knowledge" or "using thought and reasoning" then it does not require intelligence to make a reflex action.

Exercise 1.SYAI

To what extent are the following computer systems instances of artificial intelligence:

- Supermarket bar code scanners.
- Web search engines.
- Voice-activated telephone menus.
- Spelling and grammar correction features in word processing programs.
- Internet routing algorithms that respond dynamically to the state of the network.

- Although bar code scanning is in a sense computer vision, these are not AI systems. The problem of reading a bar code is an extremely limited and artificial form of visual interpretation, and it has been carefully designed to be as simple as possible, given the hardware.
- In many respects. The problem of determining the relevance of a web page to a query is a problem in natural language understanding, and the techniques are related to those

we will discuss in Chapters 24 and 25. Search engines also use clustering techniques analogous to those we discuss in Chapter 21. Likewise, other functionalities provided by a search engines use intelligent techniques; for instance, the spelling corrector uses a form of data mining based on observing users' corrections of their own spelling errors. On the other hand, the problem of indexing billions of web pages in a way that allows retrieval in seconds is a problem in database design, not in artificial intelligence.

- To a limited extent. Such menus tends to use vocabularies which are very limited – e.g. the digits, "Yes", and "No" — and within the designers' control, which greatly simplifies the problem. On the other hand, the programs must deal with an uncontrolled space of all kinds of voices and accents. Modern digital assistants like Siri and the Google Assistant make more use of artificial intelligence techniques, but still have a limited repetoire.
- Slightly at most. The spelling correction feature here is done by string comparison to a fixed dictionary. The grammar correction is more sophisticated as it need to use a set of rather complex rules reflecting the structure of natural language, but still this is a very limited and fixed task.

  The spelling correctors in search engines would be considered much more nearly instances of AI than the Word spelling corrector are, first, because the task is much more dynamic – search engine spelling correctors deal very effectively with proper names, which are detected dynamically from user queries – and, second, because of the technique used – data mining from user queries vs. string matching.
- This is borderline. There is something to be said for viewing these as intelligent agents working in cyberspace. The task is sophisticated, the information available is partial, the techniques are heuristic (not guaranteed optimal), and the state of the world is dynamic. All of these are characteristic of intelligent activities. On the other hand, the task is very far from those normally carried out in human cognition. In recent years there have been suggestions to base more core algorithmic work on machine learning.

Exercise 1.COGN

Many of the computational models of cognitive activities that have been proposed involve quite complex mathematical operations, such as convolving an image with a Gaussian or finding a minimum of the entropy function. Most humans (and certainly all animals) never learn this kind of mathematics at all, almost no one learns it before college, and almost no one can compute the convolution of a function with a Gaussian in their head. What sense does it make to say that the "vision system" is doing this kind of mathematics, whereas the actual person has no idea how to do it?

Presumably the brain has evolved so as to carry out this operations on visual images, but the mechanism is only accessible for one particular purpose in this particular cognitive task of image processing. Until about two centuries ago there was no advantage in people (or animals) being able to compute the convolution of a Gaussian for any other purpose.

The really interesting question here is what we mean by saying that the "actual person" can do something. The person can see, but he cannot compute the convolution of a Gaussian;

but computing that convolution is *part* of seeing. This is beyond the scope of this solution manual.

Exercise 1.EVOR

Why would evolution tend to result in systems that act rationally? What goals are such systems designed to achieve?

The notion of acting rationally *presupposes* an objective, whether explicit or implicit. We understand evolution as a process that operates in the physical world, where there are no inherent objectives. So the question is really asking whether evolution tends to produce systems whose behavior can be interpreted consistently as rational according to some objective.

It is tempting to say that evolution tends to produce organisms that act rationally in the pursuit of reproduction. This is not completely wrong but the true picture is much more complex because of the question of what "system" refers to—it could be organisms (humans, rats, bacteria), superorganisms (ant and termite colonies, human tribes, corals), and even individual genes and groups of genes within the genome. Selection and mutation processes operate at all these levels. By definition, the systems that exist are those whose progenitors have reproduced successfully. If we consider an ant colony, for example, there are many individual organisms (e.g., worker ants) that do not reproduce at all, so it is not completely accurate to say that evolution produces organisms whose objective is to reproduce.

Exercise 1.AISC

Is AI a science, or is it engineering? Or neither or both? Explain.

This question is intended to be about the essential nature of the AI problem and what is required to solve it, but could also be interpreted as a sociological question about the current practice of AI research.

A *science* is a field of study that leads to the acquisition of empirical knowledge by the scientific method, which involves falsifiable hypotheses about what is. A pure *engineering* field can be thought of as taking a fixed base of empirical knowledge and using it to solve problems of interest to society. Of course, engineers do bits of science—e.g., they measure the properties of building materials—and scientists do bits of engineering to create new devices and so on.

The "human" side of AI is clearly an empirical science—called cognitive science these days—because it involves psychological experiments designed out to find out how human cognition actually works. What about the the "rational" side? If we view it as studying the abstract relationship among an arbitrary task environment, a computing device, and the program for that computing device that yields the best performance in the task environment, then the rational side of AI is really mathematics and engineering; it does not require any empirical knowledge about the *actual* world—and the *actual* task environment—that we inhabit; that a given program will do well in a given environment is a *theorem*. (The same is true of pure decision theory.) In practice, however, we are interested in task environments that do approximate the actual world, so even the rational side of AI involves finding out what the actual

world is like. For example, in studying rational agents that communicate, we are interested in task environments that contain humans, so we have to find out what human language is like. In studying perception, we tend to focus on sensors such as cameras that extract useful information from the actual world. (In a world without light, cameras wouldn't be much use.) Moreover, to design vision algorithms that are good at extracting information from camera images, we need to understand the actual world that generates those images. Obtaining the required understanding of scene characteristics, object types, surface markings, and so on is a quite different kind of science from ordinary physics, chemistry, biology, and so on, but it is still science.

In summary, AI is definitely engineering but it would not be especially useful to us if it were not also an empirical science concerned with those aspects of the real world that affect the design of intelligent systems for that world.

Exercise 1.INTA

"Surely computers cannot be intelligent—they can do only what their programmers tell them." Is the latter statement true, and does it imply the former?

This depends on your definition of "intelligent" and "tell." In one sense computers only do what the programmers command them to do, but in another sense what the programmers consciously tells the computer to do often has very little to do with what the computer actually does. Anyone who has written a program with an ornery bug knows this, as does anyone who has written a successful machine learning program. So in one sense Samuel "told" the computer "learn to play checkers better than I do, and then play that way," but in another sense he told the computer "follow this learning algorithm" and it learned to play. So we're left in the situation where you may or may not consider learning to play checkers to be a sign of intelligence (or you may think that learning to play in the right way requires intelligence, but not in this way), and you may think the intelligence resides in the programmer or in the computer.

Exercise 1.INTB

"Surely animals cannot be intelligent—they can do only what their genes tell them." Is the latter statement true, and does it imply the former?

The point of this exercise is to notice the parallel with the previous one. Whatever you decided about whether computers could be intelligent in 1.11, you are committed to making the same conclusion about animals (including humans), *unless* your reasons for deciding whether something is intelligent take into account the mechanism (programming via genes versus programming via a human programmer). Note that Searle makes this appeal to mechanism in his Chinese Room argument (see Chapter 28).

Exercise 1.INTC
"Surely animals, humans, and computers cannot be intelligent—they can do only what their constituent atoms are told to do by the laws of physics." Is the latter statement true, and does it imply the former?

Again, your definition of "intelligent" drives your answer to this question.

1.2 The Foundations of Artificial Intelligence

Exercise 1.NTRC
There are well-known classes of problems that are intractably difficult for computers, and other classes that are provably undecidable. Does this mean that AI is impossible?

No. It means that AI systems should avoid trying to solve intractable problems. Usually, this means they can only approximate optimal behavior. Notice that humans don't solve NP-complete problems either. Sometimes they are good at solving specific instances with a lot of structure, perhaps with the aid of background knowledge. AI systems should attempt to do the same.

Exercise 1.SLUG
The neural structure of the sea slug *Aplysia* has been widely studied (first by Nobel Laureate Eric Kandel) because it has only about 20,000 neurons, most of them large and easily manipulated. Assuming that the cycle time for an *Aplysia* neuron is roughly the same as for a human neuron, how does the computational power, in terms of memory updates per second, compare with the personal computer described in Figure 1.2?

Depending on what you want to count, the computer has a thousand to a million times more storage, and a thousand times more operations per second.

Exercise 1.INTR
How could introspection—reporting on one's inner thoughts—be inaccurate? Could I be wrong about what I'm thinking? Discuss.

Just as you are unaware of all the steps that go into making your heart beat, you are also unaware of most of what happens in your thoughts. You do have a conscious awareness of some of your thought processes, but the majority remains opaque to your consciousness. The field of psychoanalysis is based on the idea that one needs trained professional help to analyze one's own thoughts. Neuroscience has also shown that we are unaware of much of the activity in our brains.

1.3 The History of Artificial Intelligence

Exercise 1.IQEV

Suppose we extend Evans's ANALOGY program (Evans, 1968) so that it can score 200 on a standard IQ test. Would we then have a program more intelligent than a human? Explain.

No. IQ test scores correlate well with certain other measures, such as success in college, ability to make good decisions in complex, real-world situations, ability to learn new skills and subjects quickly, and so on, but *only* if they're measuring fairly normal humans. The IQ test doesn't measure everything. A program that is specialized only for IQ tests (and specialized further only for the analogy part) would very likely perform poorly on other measures of intelligence. Consider the following analogy: if a human runs the 100m in 10 seconds, we might describe him or her as *very athletic* and expect competent performance in other areas such as walking, jumping, hurdling, and perhaps throwing balls; but we would not describe a Boeing 747 as *very athletic* because it can cover 100m in 0.4 seconds, nor would we expect it to be good at hurdling and throwing balls.

Even for humans, IQ tests are controversial because of their theoretical presuppositions about innate ability (distinct from training effects) and the generalizability of results. See *The Mismeasure of Man* (Stephen Jay Gould, 1981) or *Multiple Intelligences: the Theory in Practice* (Howard Gardner, 1993) for more on IQ tests, what they measure, and what other aspects there are to "intelligence."

Exercise 1.PRMO

Some authors have claimed that perception and motor skills are the most important part of intelligence, and that "higher level" capacities are necessarily parasitic—simple add-ons to these underlying facilities. Certainly, most of evolution and a large part of the brain have been devoted to perception and motor skills, whereas AI has found tasks such as game playing and logical inference to be easier, in many ways, than perceiving and acting in the real world. Do you think that AI's traditional focus on higher-level cognitive abilities is misplaced?

Certainly perception and motor skills are important, and it is a good thing that the fields of vision and robotics exist (whether or not you want to consider them part of "core" AI). But given a percept, an agent still has the task of "deciding" (either by deliberation or by reaction) which action to take. This is just as true in the real world as in artificial micro-worlds such as chess-playing. So computing the appropriate action will remain a crucial part of AI, regardless of the perceptual and motor system to which the agent program is "attached." On the other hand, it is true that a concentration on micro-worlds has led AI away from the really interesting environments such as those encountered by self-driving cars.

Exercise 1.WINT

Several "AI winters," or rapid collapses in levels of economic and academic activity (and media interest) associated with AI, have occurred. Describe the causes of each collapse and of the boom in interest that preceded it.

In addition to the information in the chapter, ? (?, ?), ? (?), and ? (?) provide ample starting material for the aspiring historian of AI. One can identify at least three AI winters (although the phrase was not applied to the first one, because the original phrase **nuclear winter** did not emerge until the early 1980s).

a. As noted in the chapter, research funding dried up in the early 1970s in both the US and UK. The ostensible reason was failure to make progress on the rather lavish promises of the 1960s, particularly in the areas of neural networks and machine translation. In 1970, the US Congress curtailed most AI funding from ARPA, and in 1973 the Lighthill report in the UK ended funding for all but a few researchers. Lighthill referred particularly to the difficulties of overcoming the combinatorial explosion.

b. In the late 1980s, the expert systems boom ended, due largely to the difficulty and expense of building and maintaining expert systems for complex applications, the lack of a valid uncertainty calculus in these systems, and the lack of interoperability between AI software and hardware and existing data and computation infrastructure in industry.

c. In the early 2000s, the end of the dot-com boom also ended an upsurge of interest in the use of AI systems in the burgeoning online ecosystem. AI systems had been used for such tasks as information extraction from web pages to support shopping engines and price comparisons; various kinds of search engines; planning algorithms for achieving complex goals requiring several steps and combining information from multiple web pages; and converting human-readable web pages into machine-readable database tuples to allow global information aggregation, as in citation databases constructed from online pdf files.

It is also interesting to explore the extent to which the winters were due to over-optimistic and exaggerated claims by AI researchers or to over-enthusiasm and over-interpretation of the significance of early results by funders and investors.

Exercise 1.DLAI

The resurgence of interest in AI in the 2010s is often attributed to deep learning. Explain what deep learning is, how it relates to AI as a whole, and where the core technical ideas actually originated.

Deep learning is covered in Section 1.3.8, where it is defined as "machine learning using multiple layers of simple, adjustable computing elements." Thus, it is a particular branch of machine learning, which is itself a subfield of AI. Since many AI systems do not use learning at all, and there are many effective machine learning techniques that are unrelated to deep learning, the view (often expressed in popular articles on AI) that deep learning has "replaced" AI is wrong for multiple reasons.

Some of the key technical ideas are the following (see also Chapter 22):

- *Networks of simple, adjustable computing elements*: ? (?), drawing on ? (?, ?).
- *Backpropagation*, i.e., a localized way of computing gradients of functions expressed by networks of computing elements, based on the chain rule of calculus: ? (?, ?). For neural network learning specifically, ? (?) preceded by more than a decade the much better-known work on ? (?).
- *Convolutional networks* with many copies of small subnetworks, all sharing the same patterns of weights: This is usually attributed to work in the 1990s on handwritten digit recognition by ? (?) at AT&T Bell Labs. ? (?) acknowledge the influence of the **neocognitron** model (?), which in turn was inspired by the neuroscience work of ? (?, ?).
- *Stochastic gradient descent* to facilitate learning in deep networks: as described in the historical notes section of Chapter 19, ? (?) explored stochastic approzimations to gradient methods, including convergence properties; the first application to neural networks was by ? (?) and independently by ? (?) in their ADALINE networks.

1.4 The State of the Art

Exercise 1.SOTA

Examine the AI literature to discover whether the following tasks can currently be solved by computers:

a. Playing a decent game of table tennis (Ping-Pong).
b. Driving in the center of Cairo, Egypt.
c. Driving in Victorville, California.
d. Buying a week's worth of groceries at the market.
e. Buying a week's worth of groceries on the Web.
f. Playing a decent game of bridge at a competitive level.
g. Discovering and proving new mathematical theorems.
h. Writing an intentionally funny story.
i. Giving competent legal advice in a specialized area of law.
j. Translating spoken English into spoken Swedish in real time.
k. Performing a complex surgical operation.

For the currently infeasible tasks, try to find out what the difficulties are and predict when, if ever, they will be overcome.

a. (ping-pong) A reasonable level of proficiency was achieved by Andersson's robot (Andersson, 1988).

b. (driving in Cairo) No. Although there has been a lot of progress in automated driving, they operate in restricted domains: on the highway, in gated communities, or in well-mapped cities with limited traffic problems. Driving in downtown Cairo is too unpredictable for any of these to work.

c. (driving in Victorville, California) Yes, to some extent, as demonstrated in DARPA's Urban Challenge. Some of the vehicles managed to negotiate streets, intersections, well-behaved traffic, and well-behaved pedestrians in good visual conditions.

d. (shopping at the market) No. No robot can currently put together the tasks of moving in a crowded environment, using vision to identify a wide variety of objects, and grasping the objects (including squishable vegetables) without damaging them. The component pieces are nearly able to handle the individual tasks, but it would take a major integration effort to put it all together.

e. (shopping on the web) Yes. Software robots are capable of handling such tasks, particularly if the design of the web grocery shopping site does not change radically over time.

f. (bridge) Yes. Programs such as GIB now play at a solid level.

g. (theorem proving) Yes. For example, the proof of Robbins algebra.

h. (funny story) No. While some computer-generated prose and poetry is hysterically funny, this is invariably unintentional, except in the case of programs that echo back prose that they have memorized.

i. (legal advice) Yes, in some cases. AI has a long history of research into applications of automated legal reasoning. Two outstanding examples are the Prolog-based expert systems used in the UK to guide members of the public in dealing with the intricacies of the social security and nationality laws. The social security system is said to have saved the UK government approximately $150 million in its first year of operation. However, extension into more complex areas such as contract law awaits a satisfactory encoding of the vast web of common-sense knowledge pertaining to commercial transactions and agreement and business practices.

j. (translation) Yes. Although translation is not perfect, it is serviceable, and is used by travellers every day.

k. (surgery) Yes. Robots are increasingly being used for surgery, although always under the command of a doctor. Robotic skills demonstrated at superhuman levels include drilling holes in bone to insert artificial joints, suturing, and knot-tying. They are not yet capable of planning and carrying out a complex operation autonomously from start to finish.

Exercise 1.CONT

Various subfields of AI have held contests by defining a standard task and inviting researchers to do their best. Examples include the ImageNet competition for computer vision, the DARPA Grand Challenge for robotic cars, the International Planning Competition, the Robocup robotic soccer league, the TREC information retrieval event, and contests in machine translation, speech recognition, and other fields. Investigate one of these contests, and describe the progress made over the years. To what degree have the contests advanced the state of the art in AI? Do what degree do they hurt the field by drawing energy away from new ideas?

The progress made in these contests is a matter of fact, but the impact of that progress is a matter of opinion. Some examples:

- **DARPA Grand Challenge for Robotic Cars**: In 2004 the Grand Challenge was a 240 km race through the Mojave Desert. It clearly stressed the state of the art of autonomous driving, and in fact no competitor finished the race. The best team, CMU, completed only 12 of the 240 km. In 2005 the race featured a 212km course with fewer curves and wider roads than the 2004 race. Five teams finished, with Stanford finishing first, edging out two CMU entries. This was hailed as a great achievement for robotics and for the Challenge format. In 2007 the Urban Challenge put cars in a city setting, where they had to obey traffic laws and avoid other cars. This time CMU edged out Stanford. The competition appears to have been a good testing ground to put theory into practice, something that the failures of 2004 showed was needed. But it is important that the competition was done at just the right time, when there was theoretical work to consolidate, as demonstrated by the earlier work by Dickmanns (whose VaMP car drove autonomously for 158km in 1995) and by Pomerleau (whose Navlab car drove 5000km across the USA, also in 1995, with the steering controlled autonomously for 98% of the trip, although the brakes and accelerator were controlled by a human driver).
- **International Planning Competition**: In 1998, five planners competed: Blackbox, HSP, IPP, SGP, and STAN. The result page (`ftp://ftp.cs.yale.edu/pub/mcdermott/aipscomp-results.html`) stated "all of these planners performed very well, compared to the state of the art a few years ago." Most plans found were 30 or 40 steps, with some over 100 steps. In 2008, the competition had expanded quite a bit: there were more tracks (satisficing vs. optimizing; sequential vs. temporal; static vs. learning). There were about 25 planners, including submissions from the 1998 groups (or their descendants) and new groups. Solutions found were much longer than in 1998. In sum, the field has progressed quite a bit in participation, in breadth, and in power of the planners. In the 1990s it was possible to publish a Planning paper that discussed only a theoretical approach; now it is necessary to show quantitative evidence of the efficacy of an approach. The field is stronger and more mature now, and it seems that the planning competition deserves some of the credit. However, some researchers feel that too much emphasis is placed on the particular classes of problems that appear in the competitions, and not enough on real-world applications.
- **Robocup Robotic Soccer**: This competition has proved extremely popular, attracting 300–500 teams from 40–50 countries since the mid-2000s (up from 38 teams from 11 countries in 1997). The "standard platform" league, originally based on the Sony AIBO four-legged robot, switched to the humanoid Aldebaran Nao robot in 2009. There are also small and mid-size leagues in which teams are free to design their own robots provided they satisfy certain physical constraints. The competition has served to increase interest and participation in robotics and has led to some advances in both robotics and AI (particularly related to learning in team games). Victory on competitions ofen depends less on AI than on specific tricks for improving ball-handling and shooting. The long-term goal is to defeat a human World-Cup-winning team by 2050, although the precise details of ensuring safety of human participants remain to be worked out.
- **TREC Information Retrieval Conference**: This is one of the oldest competitions,

started in 1992. The competitions have served to bring together a community of researchers, have led to a large literature of publications, and have seen progress in participation and in quality of results over the years. In the early years, TREC served its purpose as a place to do evaluations of retrieval algorithms on text collections that were large for the time. However, starting around 2000 TREC became less relevant as the advent of the World Wide Web created a corpus that was available to anyone and was much larger than anything TREC had created, and the development of commercial search engines surpassed academic research.

- **ImageNet**: The ImageNet database is a hand-labelled collection of over 14 million images in over 20,000 categories (?). The ImageNet competition (more properly, the ImageNet Large Scale Visual Recognition Challenge or ILSVRC) is based on a subset of 1,000 categories, 90 of which are breeds of dog. The availability of such large training sets is often asserted to be a contributing factor in the emergence of deep learning. In the 2012 competition, the decisive win by AlexNet (?) set off a frenzy of research that has reduced error rates on ILSVRC well below the human level of about 5%. Research by (?), however, suggests that for many types of deep learning the progress may be illusory: in many cases, the object is "recognized" through the color and spatial distribution of background pixels such as the grass on which a dog is sitting.

Overall, we see that whatever you measure is bound to increase over time. For most of these competitions, the measurement was a useful one, and the state of the art has progressed. In the case of ICAPS, some planning researchers worry that too much attention has been lavished on the competition itself. In some cases, progress has left the competition behind, as in TREC, where the resources available to commercial search engines outpaced those available to academic researchers. In this case the TREC competition was useful—it helped train many of the people who ended up in commercial search engines—and in no way drew energy away from new ideas. In computer vision, as in planning, improved performance on competition benchmarks has become almost essential for a new idea to be published, which many argue is a serious obstacle to research and may lead to an entire field become stuck in a dead-end approach.

Exercise 1.DROB

Investigate the state of the art for domestic robots: what can be done (with what assumptions and restrictions on the environment) and what problems remain unsolved? Where is research most needed?

Creating a fully general domestic robot that will work "out of the box" in any household remains a distant goal. At the time of writing (mid-2021), robot demonstrations for a number of specific tasks have been given. Some examples:

- Folding laundry (?), `https://youtu.be/gy5g33S0Gzo`. The robot assumes that a pile contains only rectangular pieces of cloth and requires a uniform green background. Subsequent work allowed a PR2 to complete almost the entire laundry cycle with a fairly wide assortment of laundry (?), `https://www.nsf.gov/discoveries/disc_summ.jsp?c` However, a demonstration exhibit at the Victoria & Albert Museum in London showed

the fragility of the solution: the commercial Baxter robot broke down too often for the exhibit to function for more than a fraction of the time.

- Loading and unloading a dishwasher: the subject of academic research-lab demos since the early 2000s (CMU's HERB robot, for example), this one is creeping closer to commercial realization, with companies such as Samsung showing carefully edited demos of robotic products that are not yet available. As with other such demos, the technology is not ready to take on an arbitrary kitchen and dishwasher. Also, many robots are far from waterproof.
- Cooking: ? (?) describe an integrated system intended to download recipes from the Internet, turn them into executable plans, and carry out those plans in a real kitchen. Only quite preliminary results were achieved: `https://www.csail.mit.edu/node/6480`. As with washing up, cooking is very messy and current general-purpose robots need to be carefully wrapped in plastic protective gear. There are also commercial "robot cooks" that claim to cook dishes in the home. Typically these are special-purpose systems that require specially prepared and sized ingredients and are generally not robust to failure (see, e.g., `https://www.youtube.com/watch?v=M8r0gmQXm1Y` and `https://www.youtube.com/watch?v=GyEHRXA_aA4`). Again, we are far from having a robot that go into anyone's kitchen with a bag of ingedients from the supermarket and make dinner.

Successful teleoperation demos (where a human controls a robot to carry out tasks such as opening a beer bottle) suggest that the problems do not lie primarily with robotic hardware (except for the need to design robots that are immune to water, grease, and solid ingredients. Perception has advanced considerably, but it is still difficult for a robot to analyze a pile of ingredients in a bowl and judge where the surface is and how well mixed the ingredients are. Spatial reasoning involving continuous, irregular shapes and liquid, semiliquid, and powdered ingredients or complex objects made from cloth, leather, etc., is quite weak. Also lacking is the ability to plan and manage domestic activities on a continuous basis, with constant interruptions from humans who rearrange the world.

Exercise 1.JRNL

The vast majority of academic publications and media articles report on successes of AI. Examine recent articles describing failures of AI. (? (?) and ? (?) are good examples, but feel free to find your own.) How serious are the problems identified? Are they examples of overclaiming of or fundamental problems with the technical approach?

? (?) define *overinterpretation* as occurring when a machine learning algorithm finds predictive regularities in parts of the input data that are semantically irrelevant. Those regularities reflect particular properties of the training and test data, such as when all photos of Norfolk Terriers are taken with the dog sitting in various poses a particular crimson carpet. In that case, simply recognizing the presence of a few pixels of crimson suffices to identify the breed of dog. They show that this problem arises quite frequently with modern deep learning systems applied to standard data sets such as ImageNet and CIFAR-10. For some categories, such as "airplane" in CIFAR-10, a single pixel suffices to classify the object. In

almost all cases, the trained network labels images with high confidence using only a very small subset of fairly randomly distributed pixels that, to a human, is unintelligible. This seems to reveal a fundamental problem both with the technology (the networks appear to have too much capacity and no notion of what they are looking for) and the standard train/test methodology (which fails to detect the non-robustness of the learned classifiers). Put another way, the network does not know that "Norfolk Terrier" is a breed of dog rather than a carpet color ("Crimson Carpet"), so the task solved by the machine learning algorithm is not the one solved by a human.

? (?) describe a similar issue: for any given training set, there are typically vast numbers of learned network configurations that give equally good (or even perfect) performance on held-out data. They call this *underspecification* and note that it leads to poor performance after deployment in a wide range of real-world applications including computer vision, medical imaging, natural language processing, clinical risk prediction based on electronic health records, and medical genomics. They conclude that machine learning pipelines must be constrained by strong semantic priors if learning is to be effective and reliable (?, see also).

For a paper highly critical of the use of machine learning to diagnose COVID-19 from chest X-rays, see ? (?). The authors selected 62 of the most promising and carefully documented studies from a total of 2,212 published in 2020, but found that "none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases." The problems identified were primarily failings on the part of the studies' authors, who may not have been aware of the stringent criteria for real-world deployment of high-stakes diagnostic tools. However, even a methodologically perfect study using deep learning might well have failed for reasons given in the preceding two paragraphs.

1.5 Risks and Benefits of AI

Exercise 1.PRIN

Find and analyze at least three sets of proposed principles for the governance of AI. What do the sets of principles have in common? How do they differ? How implementable are these principles?

There are several hundred published sets of principles: some of the best-known are the OECD, Beijing, and Asilomar principles. ? (?, ?), and ? (?) provide useful analytical surveys and summaries.

Exercise 1.EURG

Study the 2021 EU "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)" (or its final version, if available). Which provisions appear to have the most direct and tangible impact on what kinds of AI systems can be deployed?

If students have never seen legislation before, this will be an eye-opening exercise. Most

of the document stipulates practices in testing, documentation, etc. Only a few of the clauses have tangible impact—most obviously, those concerning facial recognition, impersonation of humans, and deepfakes. In these three areas, the rules are quite strict compared to those in force in most other parts of the world.

Exercise 1.LATM

Summarize the pros and cons of allowing the development, deployment, and use of lethal autonomous weapons.

A good place to start is the Congressional Research Service report "International Discussions Concerning Lethal Autonomous Weapon Systems," dated October 15, 2020. Unfortunately the report leaves out the principal argument against lethal autonomous weapons: they will become cheap, scalable weapons of mass destruction that will proliferate easily and are likely to be used against civilian populations.

Exercise 1.BIAS

Many researchers have pointed to the possibility that machine learning algorithms will produce classifiers that display racial, gender, or other forms of bias. How does this bias arise? Is it possible to constrain machine learning algorithms to produce rigorously fair predictions?

Bias in ML occurs largely because of two things: first, the bias that exists in data generated by a biased society, and second, specifying an incorrect objective for machine learning algorithms——namely, maximizing agreement with the training data. This is not the real objective: in fact, we care about fairness too.

It's possible to define various measures of fairness and to design algorithms that respect them. See ? (?) for a survey. Unfortunately, it is not possible to simultaneously satisfy all "reasonable" definitions of fairness, and there may be some sacrifice in accuracy.

Exercise 1.SIFI

Examine at least three science-fiction movies in which AI systems threaten to (or actually do) "take control." In the movie, does the takeover stem from "spooky emergent consciousness" or from a poorly defined objective? If the latter, what was it?

Here are three examples:

a. *Colossus—The Forbin Project*: A US defense computer tasked with ensuring peace and security encounters a Soviet compueter with the same goal. They exchange information and decide that the goal is best achieved by jointly controlling all nuclear weapons and using threats of destruction to force humans to drop all aggressive war plans. Here, the problem is clearly a poorly defined objective.

b. *Ex Machina*: Ava, a very human-like android, brilliantly engineers her escape from a remote facility by outwitting her human captors and allies. Although it is never stated

explicitly in the film, her objective seems to be in places populated by large numbers of people; she is portrayed as having this objective as a result of a process of consciousness emerging from the sum total of human interactions recorded in a global social media platform.

c. *Transcendence*: Will Caster, a Berkeley AI professor, is gunned down by anti-AI terrorists. Before he dies, his brain is uploaded to a quantum supercomputer. The machine becomes conscious and begins to quickly outrun the human race, threatening to take over the world. Here, the risk comes from the possibility that the machine's conscious goals differ from those of its human progenitor.