DTU Compute
Department of Applied Mathematics and Computer Science

Artificial Intelligence and Human Cognition Lecture Notes

v. 1.25.2

DTU

Contents

CHAPTER 1

Introduction to cognitive science and artificial intelligence

1.1 Human vs. artificial intelligence

A first step in comparing human and artificial intelligence is to determine whether the word 'intelligence' means the same in those two contexts. Intelligence generally refers to the ability to learn from experience and apply what is learned to new situations. In this sense, human and machine intelligence are similar but when we begin to break down learning into separate domains, we find that the domains of human and machine intelligence are quite different.

How can test the intelligence of humans? One approach would be to test their ability to learn in a domain that we believe requires intelligence to master. We could, for example, device a test in learning mathematical skills because we believe that intelligence is required to perform well in such tests. However, we could also device a test of learning linguistic skills. If only one general form of intelligence is required to master both tests then we would expect performance on the two tests to correlate strongly. If the correlation is weak then our measure of intelligence will be dependent on which test we use, and we must conclude that there is not general intelligence, but rather a number of specialised intelligences. Whether intelligence is general across all domains or whether humans have multiple intelligences that are domain specific has been a fundamental issue in the study of human intelligence. The question is, in other words, whether high intelligence means a high general ability to learn across distinct domains such as mathematics, language and music or whether it is specialised within such domains.

Strong evidence for a general human intelligence has come from studies analysing performance across multiple domains. Many studies have found that performance correlates across multiple tasks. If, in other words, a person learns mathematics easily then it is likely that the person will also learn languages easily. This shows that there is a single underlying factor, referred to as Spearman's g. Spearman's g does, however, only account for less than 50% of the variance, indicating that there are also parts of human intelligence that are specialised for particular domains. Three domains have been identified by finding correlations across cognitive tests that exist in addition to the correlation accounted for by Spearman's g: numerical, spatial and linguistic

The current consensus model of human intelligence is a hierarchical model. At the top of the hierarchy sits general intelligence, Spearman's g, which seems to influence our ability to learn across many domains. At the next level sits the three specific abilities, or intelligences, pertaining specifically to numerical, spatial and linguistic learning. Although this model accounts for a relatively large part of the variance in test performance, it does not account for all of it. Therefore, models have been developed containing additional levels in the hierarchy each dividing the domains into several subdomains.

To arrive at a full model of intelligence we must make sure that we have included tests that tap into all relevant subdomains. This is a difficult task and whether we have solved it is still a contentious issue. Take for example the ability to understand other peoples' intentions and feelings. This is sometimes referred to social intelligence. Another example is the learning of motor skills. Some people learn to

swim or juggle four balls more easily than others. Should these abilities be included in our model of intelligence, or should it they be considered distinct abilities, or potentials? This depends, in large part, of how we define intelligence. Would we, for example, based on how easily a person learned to swim make judgment about the person's intelligence? Perhaps this would stretch the concept of intelligence too far for many of us. However, we would be more likely to accept a midfielder's ability to read the game and pass the ball in ways that tricks the opponent as a form of intelligence, even though this ability may not be related to numerical, spatial or linguistic skills. These examples show that the domain of intelligence is matter of definition.

Moving on to artificial intelligence, we will soon realise that many of the skills that have proven to be challenging for AI systems seem to fall quite far outside of what would commonly be considered intelligent for humans. Take, for example, walking. Bipedal walking has been challenging to implement in AI controlled robots, because it requires a fast and complex feedback loop that registers small changes in the ground surface and the centre of gravity in order not to fall. Even though we have begun to see robots mastering bipedal walking, the next level challenges of running, jumping or dribbling a ball are still not mastered by AI systems. Yet, we would not consider mastery of these skills by a human as a sign of high intelligence. Other examples include recognising objects in images, comprehending speech and driving a car. These have all been challenges for artificial intelligence but not for human intelligence. One the other hand, human intelligence tests contain parts that could easily be solved by a computer with simple algorithms that are not considered AI. The Wechsler Adult Intelligence Scale, for instance, contain a working memory test, in which the test participant must store a sequence of digits and report them back in reverse order, and a timed arithmetic test, in which the test participant has to solve an arithmetic problem within a time limit. It seems reasonable and valid to include such tests in a test of human intelligence because we would generally consider the ability to solve them as a sign of high intelligence in humans. It is, however, trivial to program a computer to solve them at a much higher level than could ever be achieved by a human even without the use of what we understand as artificial intelligence.

In summary, intelligence, be it human or artificial is a measure of the system's ability to learn across a range of domains. However, the range of domains that we consider related to intelligence, depends on the definition of intelligence and is different for human and artificial intelligence.

1.2 Levels of computation

The human brain computes. Computers as we know them, obviously, also computes. Therefore, it analysing the human brain in a computational framework is an important and influential approach in cognitive neuroscience.

It may seem obvious that the brain computes. After all, the term 'computer' actually originally referred to a profession, or occupation, of doing computations, that were typically repetitive and tedious, for scientific laboratories, observatories or financial institutions. When modern computers were developed, they were called 'electronic computers' to distinguish the machines from humans working as computers. Soon after, the term 'computer' began to unambiguously refer to machines as the occupational title for humans disappeared. As we shall see in this and the following sections, computation is ubiquitous in the human brain and takes many forms beyond tedious, repetitive, formal computations.

You may think that the computational approach to understanding the human brain is somewhat impoverished or simplified. Is the human brain really nothing but a computer? Can complex human abilities like emotions, empathy and intuition really be reduced to computation? Are humans not, unlike computers, conscious? Interesting as these questions may be, we will not dwell on them here. Whether there is more to human cognition and the human brain is left as an open question. The only claim that we will make is that computation is an essential part of what the brain does. We do not aim to show that the brain does nothing but computing.

Brains and computers are, obviously, quite different even though they may both compute. Humans are alive while computers are not. Humans are therefore emergent systems while computers are programmed and designed. Computers and brains have very different structure. Such comparisons can

lead to some confusion unless we define what we mean by *computation* more clearly. David Marr was an early advocate of the computational approach to understanding human cognition. He defined three levels of computation that are helpful for comparing computation in humans and machines.

1. **Computational level**

   The Computational level defines the computational problem that the system is solving including its inputs, outputs, and objective. It is at this level that computation in humans an machines directly are alike. Note that the requirements do not specify *how* the system should be built. This is specified at the algorithmic and implementational levels.

   From a computer science perspective, the computational level corresponds approximately to a set of *system requirements*. From a biological perspective, the computational level corresponds closely to a description of cognitive functions or abilities. Marr used the example of constructing a 3-dimensional model of the world given two 2-dimensional retinal images. The model can be used to estimate distances to objects. From a computer science perspective, we can define the computational problem by its inputs (stereo images), its outputs, (an estimate of distance, or depth), and an objective, (minimising the error of the estimate). From a biological perspective, we can observe that a biological system is able to perform the task although this often require careful control and measurement of the input, output and error.

2. **Algorithmic level**

   The algorithmic level describes *how* the computational problem is solved by describing the underlying algorithm.

   From a computer science perspective, the algorithmic level corresponds to computer algorithms and software implementation. From a biological perspective, the algorithmic level corresponds to the algorithms that the brain uses to solve the problem. Staying with Marr's example of of constructing a 3-dimensional model of the world given two 2-dimensional retinal images we could, for example, specify that the system should estimate the displacement of the retinal images and base its estimate of the distance to an object using triangulation. From a computer science perspective, we typically have a very clear description of machine learning systems at the algorithmic level because the algorithms are designed and documented by humans. From a biological perspective, it is a challenge to describe human cognition at this level. The problem is analogous to the problem of describing the underlying algorithm of a machine learning system *without* access to the code and with no knowledge of its design. In this sense it is a *reverse engineering problem*. This is the topic of computational cognitive science where computer algorithms are used as models of human computation. We shall present some examples of this approach in later chapters.

3. **Implementational level**

   The implementation level describes the physical realization of the system.

   From a computer science perspective, the algorithmic level corresponds to computer hardware. As for the algorithmic level, we typically have a very clear understanding of this level. We could, for example, specify that the triangulation algorithm described above runs on a computer with certain specifications (no dedicated graphics card, that it has 4 GB RAM, etc). From a biological perspective, the an understanding of the implementational level comes through computational neuroscience, the study of the brain and how it computes. We could, for example, specify the brain areas involved in the computations. We could also describe how neurons are connected within those areas. Or, we could describe the relative timing of the activation of those areas. All of this will, of course, require measurements of neural activity. We can use this information to build computational models that simulate neural activity performing the computations.

We can now proceed to study the similarities and differences between the key elements of human and machine intelligence: learning and memory.

1.3 Learning

Machine learning, arguably the most influential form of AI, can be divided into three learning paradigms. We can place many aspects of human learning under the same paradigms. This will be a first step in choosing a specific machine learning model that can be compared to human cognition.

- **Supervised learning**

  The goal in supervised learning is to learn a function that maps inputs (features) to outputs (labels). The function is learned from a training set of input-output pairs. The objective is to minimise the error between the predicted label and the true label.

  Humans supervised learning occur in many domains. One example is image classification. Say we have a set of images, or image features, with each image labelled as portraying a cat or a dog. The goal is to learn to label new, unlabelled images correctly. Another example could be learning the pronunciation of words. Here the input, a word, is symbolic and the output label is a phonetic representation. When studying these types of learning in humans, we can choose supervised machine learning algorithms as models and begin to ask further questions. We can, for example, ask if human classification is better modelled by a linear or non-linear classifier.

- **Unsupervised learning**

  The goal in unsupervised learning models is to map inputs to structures such as lower-dimensional representations, clusters or probability distributions. The structure is learned from a training set of unlabelled data. The objective is generally a measure of goodness of the structure and depends on the particular type of unsupervised learning. The goodness can reflect how accurately the input can be compressed and reconstructed, as for Principal Component Analysis (PCA) or auto-encoders. Alternatively, it could be a measure of how tightly similar inputs cluster together and how distinct the clusters are from one another as in k-means clustering. Or, it could be the likelihood of observed data if the structure is a probabilistic model like a Gaussian Mixture Model (GMM) or a variational auto-encoder (VAE).

  Humans learn to extract meaningful features from sensory inputs independent of any type of feedback. One example is children learning to segment speech sounds and words from a continuous sound stream. Naively, we tend to perceive pauses of silence in between words in natural speech. Perhaps you have noticed that these pauses are difficult to hear in a language that you do not understand. The reason is that there are no pauses. Children depend on the statistical regularities of language for learning to segment the speech streams. Learning this, depends strongly on exposure to language, but not so much on feedback. Feedback does, of course, play an important role for acquiring language, but only in later stages, when we learn the meaning of words in the language. Another example is our ability to extract structure, such as edges and objects, from visual input. Ample evidence show that this ability depends on receiving varied visual inputs during development. This indicates that it is an ability that is learned, but it does not depend on feedback. In Chapter **??**, we will examine how the brain's visual system may employ unsupervised algorithms, like PCA, to learn how to extract meaningful features from visual input.

- **Reinforcement learning**

  The goal in reinforcement learning models is to learn a function—called a *policy*—that maps states of the environment to actions. In this sense, reinforcement learning resembles supervised learning, since states can be viewed as inputs and actions as outputs. The key difference lies in the objective: supervised learning adjusts the model to match correct labels for each input, whereas reinforcement learning adjusts the policy to maximize cumulative reward across future states, even when the reward is delayed. The objective is thus defined by expected long-term reward, sometimes modified with a discount factor to weigh immediate and future rewards.

  Games such as computer games, chess, and Go are popular test-beds for reinforcement learning algorithms and illustrate reinforcement learning in humans. Unlike many real-world problems,

they have well-defined discrete state and action spaces, which helps defining computer algorithms to solve them. Humans and computers alike are able to learn how to play these games not only by instruction but by trial-and-error. In this case, it should be clear how reinforcement learning differs from supervised learning as the reward, or penalty, is not immediately relatable to a particular action. A chess move that leads to an immediate gain may lead to a future loss.

We can use the division of learning into to the three learning paradigms to draw parallels between human and machine learning and to develop models of human learning inspired by machine learning algorithms. Still, some cases of human learning are difficult to place in these paradigms and do not have well-developed machine learning analogies. Imitation and observational learning, for example, is still largely outside of machine learning. An example of this is when children learn language or motor skills simply by observing others with no explicit instructions or reward. Exploratory and intrinsically motivated learning is also difficult to place in a machine learning paradigm. Children's play is a good example of this type of learning. Another example could be artists developing their skills to express themselves despite lack of reward or an externally defined goal. Still, settling for less than a complete computational account of human learning, the three learning paradigms is a good starting point for comparing human and machine learning.

So far, we have described the learning paradigms only at the computational and algorithmic levels. We have waited with a description of the implementational level for a reason: The hardware used in machine learning varies little across the learning domains. Our choice for the configuration of our computers rarely depend on whether we want to develop a supervised or an unsupervised model. The brain, however, contains many distinct modules for learning. A full description is beyond the scope of these lecture notes, but in the next sections we will begin to compare the most fundamental component of learning, memory, at the implementational level.

1.4 Human vs Computer Memory

For humans and machines, there can be no learning without memory. In this section, we will compare human and computer memory. Again, we will see that the comparison is more meaningful if we compare at each of Marr's three levels of computation.

At the computational level, we can define memory as the ability to encode information, store it for a period of time, and then retrieve it for recall, use in processing new information or guiding actions.

At the algorithmic level, notice that information takes two distinct forms in the learning paradigms described above. The input data to learning algorithms is information and the change in model parameters that comes from training is also information. This distinction reflects the distinction between rote memorisation and learning.

At the implementational level we find the greatest difference between human and computer memory.

Computers store memory in cells that can be in one of two states, effectively representing zeros and ones. At the implementational level, all information is therefore stored in binary form. Images, sounds, or text must first be encoded into this binary form before they can be stored, and later decoded back into usable formats. These encoding and decoding processes typically occur at the algorithmic level. Similarly, machine learning models, including their parameters and weights, are stored in binary, while the actual learning–the updating of parameters—takes place at the algorithmic level.

Brains store information differently. Neurons are inter-connected by synaptic connections. It is this structure that inspired artificial neural networks in machine learning. The strength of the connections are similar to the weights in artificial neural networks. In brains, learning thus takes place at the implementational level without a algorithmic layer encoding weights as binary data. Data, such as images, sounds, or text is also encoded into this format and the brain is therefore not well suited to rote memorisation of this type of data. When you remember an image, for example, your brain encodes it as features, objects and relationships. Your brain is very poor at remembering the image pixel by pixel even for images with very low resolution.

We can arrive at a quantitative estimate of the difference between the capacity of human and computer memory. The largest computer data storage systems are being measured in petabytes, that is 10^{15} bytes. In comparison, the world record for memorising digits of π is around 70.000, which corresponds to approximately 34 kilobytes of data. Measured in this way, the memory capacity of computers thus exceeds that of humans by about twelve orders of magnitude. However, somewhat mysteriously, humans still surpass computers in many aspects of learning.

There are, however, also a striking similarity between human and computer memory at the implementational level: they both divide memory into volatile and non-volatile memory.

Computers can store information in *Random Access Memory* (RAM) and in hard or solid states drives. The reason for this division is that each type of memory has its advantages and disadvantages. The main advantage of RAM is speed, which makes it well suited for storing information that is actively being processed. Hence, we could also refer to it as the *working memory* of computers. Its disadvantage is that it requires a constant supply of energy. This is why it is volatile: all information is lost when the power is turned off. The energy requirement affects, to some extent, how the capacity of RAM is prioritised when building computers. Typically, the capacity of RAM is much smaller than the capacity of hard or solid state drives. Another important reason is that RAM is more expensive to build.

It probably come as no surprise that volatile and non-volatile memory systems are distinct in the brain. Volatile, short-term or working memory is used for information that is actively processed. It requires neural activity and therefore energy, which is a great concern for a biological system. This is one likely reason for why its capacity is limited. The information you can actively process is very limited compared to all the things you remember. Long term memory, in contrast, stores information in synaptic connections as described above. Although this distinction between short-term working memory and long-term memory may seem familiar, or even obvious, the evidence for these memory systems being distinct in the brain has taken decades of research, and our understanding of it is still far from complete. This will be the topic of the next section.

1.5 Human working memory and long-term memory

Several lines of evidence show that working memory (volatile) and long term memory (non-volatile) are distinct components, or *modules*, of human memory. Merging several lines of evidence is an important practise in the study of the brain and cognition and ensures that alternative hypotheses are ruled out. For example, you might think that it is obvious that volatile and non-volatile memory systems exist in the human brain because, obviously, some memories are fleeting while other memories last. However, this does not necessarily mean that the two systems are separate. It could instead reflect a single memory system in which information is stored more persistently than other information.

Some early evidence showing that human memory consist of distinct working memory and long-term memory memory modules comes from the study of a patient known as HM. HM underwent surgical removal of a brain structure, the hippocampus, to treat his severe epilepsy. After the surgery, HM appeared normal, but it soon transpired that the surgery had left him without the ability to acquire new long-term memories. His working memory was unimpaired, so that he was able to hold a conversation or even play a game of chess. That is why he, at first glance, appeared normal, but he would hold no memory of the conversation or the game of chess after a short while if he was distracted. His long term memory was also intact, so that he was able to recall events that happened before the surgery. This indicates that the mechanism for storing memories long-term can be impaired even if working memory is intact, but could it still be part of a single system? Studies of another patient, KF, who suffered brain injury from a bicycle accident showed that working memory can be impaired without affecting the mechanism that stores memories long term. This shows a *double dissociation* between working memory and the mechanism that stores information long-term. Since we have observed impaired working memory with intact transfer to long-term memory (patient KF) *and* the reverse pattern of impaired transfer to long-term memory with intact working memory, we now have strong evidence for the two functions being distinct and separable.

Evidence from neuroscience supports the finding that working memory and storing information into long-term memory are distinct functions. When humans hold information in working memory, neurons in certain parts of the cerebral cortex are persistently active. Notably, this sustained activity predicts whether information was retained. In contrast, when humans learn and stores information in long-term memory, activation occurs in a different part of the brain, the hippocampus, which is the structure that was excised in HM's surgery.

Additional support has come from behavioural studies, the topic of *experimental psychology*. In the free recall task, participants are presented with a sequence of more than 10 items that they are asked to try to store in memory. This is too many items for most people to keep in working memory, so task performance rely on the participants' ability to store the items in long-term memory. When the sequence ends, participants are asked to recall as many items as they can. Participants can report the items that they remember in any order. Therefore the task is known as the free recall task. When the proportion of letters that were recalled, across several repetitions of the experiment, is plotted as a function of the item's position in the presentation sequence, two interesting effects can be seen. Items in the beginning of the sequence are recalled better than items in the middle of the sequence. This is called the *primacy effect*. The primacy effect reflects that the task is easier in the beginning of the sequence because the number of items that participants are storing is still low. Hence, participants can allocate more time to storing each new item in long-term memory. As the number of items increase through the sequence, the task becomes more difficult and performance suffer. Also, items at the end of the sequence are recalled better than items in the middle of the sequence. This is called the *recency effect*. The recency effect reflects that items at the end of the sequence do not need to be stored in long-term memory since they can be held in the short-term working memory. Control experiments has verified this interpretation of the primacy and recency effects. First, increasing the rate at which the items are presented decrease the primacy effect because participants will now have less time to dedicate to storing the first items in the sequence. This, however, does not affect the recency effect because the items can still be stored in working memory as well as when the rate of presentation was slower. Secondly, adding a working memory task to the experiment, right after the sequence ends, before the participants begin to recall it, decreases the recency effect because participants cannot hold the last items in working memory when it is occupied by another task. Notably, a delay with no working memory task, does not affect the recency effect because it does not interfere with the participants working memory. This shows that that the effect is due to the working memory task and not simply due to the delay. As a final observation, the working memory task, does not affect the primacy effect because the first items in the sequence have already been stored in long-term memory.

In summary, converging lines of evidence show that human working memory and long-term memory can be considered distinct, functionally segregated modules. We have, however, only scratched the surface in describing the complex architecture of human memory.

1.6 Human working memory

A surprising fact stemming back from the earliest studies of human working memory is that its capacity is quite limited. In serial recall tasks, a sequence of letters or digits are presented to observers who then tries to recall the sequence in the correct order. Surprisingly, most observers struggle with recalling sequences of more than seven items. We can translate that into bytes if we assume that the items are letters in the English alphabet. As there are 26 letters in the alphabet, $2^5 = 32$ states will suffice to define a letter, which means that each letter require 5 bits. A memory with a capacity to hold seven letters therefore corresponds roughly to 35 bits, or a little less than 5 bytes. Measured in this way, which we will come to see as an over-simplification, the capacity of human working memory is approximately 13 bytes. This about nine orders of magnitude smaller than the capacity of the volatile memory in standard computer, which is typically measured in gigabytes. Compare this to our mental experience of the complex world around us. When we walk into a room full of people and objects, we experience that we immediately perceive it all quite clearly. Even when we exit the room, we often have a sense of being able to recollect the experience quite clearly. This, however, is quite far from

the truth. Many experiments show that we fail to notice quite drastic changes in a scene after a brief intermission—a phenomenon known as change blindness. In one striking demonstration, a conversation was disrupted by a couple of men carrying a door and walking in between the interlocutors. Unknown to one of the interlocutors, the unknowing participant in the experiment, the other was swiftly exchanged with another person, which often went unnoticed by the unknowing participant in the experiment, demonstrating our very limited working memory capacity. This leaves us with the question of how human behaviour can be so complex compared to that of computers when our working memory capacity is so limited in comparison.

One clue for why human working memory is not quite as limited as estimated above comes from a control experiment in which, instead of single letters, each item is a three letter word like *cat*, *dog*, *job*, or *eye*. In this experiment, we will find that the capacity is *still* approximately 7 items. This is surprising because, measured in bytes, we have clearly tripled the information load. Had we instead used random consonant triplets rather than three-letter words in the memory test, humans would have been able to hold far fewer items in working memory. This reflects the inherently associative nature of human memory: it stores patterns—often called *chunks* in the cognitive science literature—rather than isolated bits. These patterns can more complex than three letter words. The upper limit is elusive and depends on the observer's prior experience and context, since items in working memory function as *pointers* to patterns stored in long-term memory.

Working memory can be shown to consist of sub-modules. When letters are used in serial recall tasks participants tend to make certain types of errors. Specifically, they tend to confuse letters that sound alike (such as 'F' and 'S') rather than letters that look alike (such as 'F' and 'E'). This happens even though the letters are presented visually so it is unlikely to be because the letters are misperceived. Instead, it indicates that the letters are encoded as sounds in working memory. Why would observers recode letters presented visually to sounds? Although introspection is not a scientifically valid method because the observations cannot be shared, we can still use it as inspiration. Trying a serial recall task of remembering a random sequence of letters for a short period of time may shed some light on the mechanics of working memory: we tend to repeat the sequence to ourselves in the retention period. This is sometimes called articulatory rehearsal. We often do this quietly with a silent inner voice, without moving our mouth and tongue, a phenomenon called subvocalisation. Articulatory rehearsal can be considered a sub-module of working memory, sometimes referred to as a *slave system*, because of its passive and repetitive character, hardly deserving the term working memory. The use of articulatory rehearsal in this way has been confirmed by behavioural experiments. In one experiment, participants were prevented from using articulatory rehearsal by requiring them to vocalize an irrelevant sequence of sounds (for example, "tah-dah-tah-dah…"). This procedure, known as articulatory suppression, disrupts the phonological loop and eliminates the usual advantage of articulatory rehearsal. As a result, memory span for verbal material is greatly reduced, from approximately seven to four items, demonstrating that articulatory rehearsal is a key mechanism for maintaining verbal information in working memory. Furthermore, an analysis of the type of errors that the participants made showed that they did not tend to confuse letters that sound alike rather than letters that look alike during articulatory suppression, which lends further support for the effects of articulatory rehearsal. Finally, to show that the decrease in working memory capacity is not simply an effect of a secondary task, and not specific to an articulatory task, another experiment showed that other tasks, like finger-tapping, does not cause a decrease in working memory capacity.

Since the early description of working memory and articulatory rehearsal, two additional slave systems, the visuo-spatial sketchpad and the episodic buffer have been shown to also aid working memory. As their names imply, these systems are specialized: the visuospatial sketchpad stores visual and spatial information, while the episodic buffer maintains an integrated "story-line" of the recent past by binding information across modalities and linking it to long-term memory.

We have now described how human working memory consists of multiple specialized components and stores patterns of varying complexity rather than raw information. In this sense it differs fundamentally from volatile computer memory. These differences arise at the implementational level. For brains, however, the distinction between the implementational and algorithmic levels is vague because the hardware is so specialised for the algorithm. For computers, software systems such as machine learning

models are needed before they can mimic some of the ways in which human memory stores and processes information.

1.7 AI mimicking human cognition

AI systems are often built to mimic human cognitive functions. Examples include speech recognition and some image classification tasks. It can be tempting to consider this type of problems as reverse engineering problems where a functioning system—humans in this case—are available, so that we can try to construct the engineering solution—an AI system in this case—to work in the same manner. History carries a few lessons that show that this is not always the best approach. One lesson comes not from the history of AI but from the history of human flight. The earliest pioneers of aviation attacked the problem as a reverse engineering solution attempting to build flying machines mimicking the flapping wings of birds. We now know that aircrafts, with fixed wings and the power of combustion engines, unparalleled by nature, is the better solution. Another example that comes from the history of AI, is speech recognition. Where some early solutions tried to mimic the function of the human auditory system, more recent and successful applications are based on neural networks that are only vaguely inspired by the human brain and owe, in large part, their success to being trained on a much larger speech corpus than that needed by a human child to acquire speech. Hence, although the reverse engineering approach may inspire the development of AI systems it may also be an unnecessary constraint.

Mimicking human cognition require AI systems to be trained on behavioural data, i.e. data that somehow records human behaviour. One example is AI systems used to classify electroencephalography (EEG) data. EEG is typically recorded by measuring the changes in the electrical field from electrodes placed on the human scalp. These changes are influenced by neural activity but also by eye movements and muscle activity. In order to filter out the activity from non-neural sources human experts visually inspect the data and classify it as stemming from neural or non-neural activity. This is a tedious, slow and expensive process and machine learning approaches has therefore been developed to solve the problem. The data available to train the algorithms typically consists of samples of EEG data, which has been labelled by human experts. Therefore, the data does not actually contain the ground truth, i.e. a label indicating whether the sample actually contained data reflecting neural activity or not. Instead, the algorithm is trained to mimic human experts including the mistakes that they make. Therefore, we cannot hope for the algorithm to perform better than human experts, even though this might be possible if ground truth data were available. The performance of human experts is often quantified as the inter-rater agreement, i.e. the rate at which the human experts agree with one another. We can use this measure as benchmark that we can use to assess the performance of AI systems.

Another example of AI systems mimicking human behaviour is visual object recognition. Here, the samples consist of images and labels produced by humans labelling images as depicting, say, a traffic light or a dog. Large sample sizes are often needed to train an AI system and the labelling is therefore costly even though it a single image can easily be labelled by a human. It therefore takes care and consideration to select these samples. It is, for example, important to check that the humans are really trying their best when performing the task. Also, the samples must be representative of samples that the AI system will meet when deployed. Many examples show that object recognition systems can be fooled by images that were unmatched in the training samples. One example showed that a small sticker on a road sign would cause some AI systems to mis-classify the road sign as another type of object. Humans, on the other hand, would not be so easily duped. Carefully selecting the samples to represent the scenario in which the AI system will be deployed would allow it to better mimic the powerful abilities of human cognition.

The inability of many AI systems to generalise to data that are outside the training data distribution, challenges the concept that AI systems are intelligent. Imagine a human that had been trained to recognise traffic signs but had never seen a traffic sign with a small sticker patched onto it. Humans would be very unlikely to make the mistake of thinking that the traffic sign is not a traffic sign. The reason for this is probably that human visual perception is interlinked with a general understanding

of the world around us, something that AI systems generally lack. In this respect, humans are still superior to AI systems. The superior human ability to generalise is, however, not just due to our vast general knowledge of the world. Children, learn to classify, say, animal species with far fewer samples than the number required by AI systems. This difference deserves further study if we aim to develop better and more general AI rather than just developing AI solutions that work for specific problems.

1.8 Mini project

In this mini project you will work in groups of 2-4 students. You will conduct the behavioural experiments in an attempt to replicate the findings experiments described in Section 1.5 and 1.6 including the control experiments.

Specifically, you should run the following experiments and try to replicate the effect listed under each experiment.

Free recall experiment effects

- The primacy and recency effects
- The effect of increasing the rate at which the items are presented
- The effect of adding a working memory task to the experiment, right after the sequence ends, before the participants begin to recall it
- The effect of adding a pause to the experiment, right after the sequence ends, before the participants begin to recall it

Serial recall experiment

- The limited capacity of working memory
- The type of errors that participants make
- The effect of chunking
- The effect of articulatory suppression
- The (lack of) effect of finger tapping

First, design the experiments. Some of the things that you need to decide on are

- Whether you want to use digits, letters, words, or something else as the items that participants memorise
- The rate at which the items are presented
- How participants respond
- How you calculate the proportion of correct responses from the responses
- How you compare the proportion correct within and across experiments. You should include some error measure, like confidence intervals, in your comparison. Recall, that if the confidence intervals are overlapping then there is no difference
- How many repetitions (at least 20) of the experiment each participant performs

When you have designed the experiments, you should write a script that can run the experiment. Each member of your group should run the experiment. Once you have the data, then you should analyse the results. You can pool the data across the participants meaning that if you have three participants running each experiment 20 times then you can treat the data as if a single participant had run the experiment 60 times. This is a simplification and not the proper way of analysing the results, because it ignores the variability across participants, but the full analysis is beyond the scope of this course. Based on your analysis, discuss whether you were able to replicate the findings described in the main text.

You report should be no more than 5 pages including figures. Your report will be evaluated on the following elements

- The description of your experimental design
- The description of your analysis of the results
- The discussion of whether your results replicate the experimental results described in Section 1.5 and 1.6 including the control experiments
- The discussion of why your results might not replicate the results described in Section 1.5 and 1.6 including the control experiments

Note that your will *not* be evaluated on whether you actually manage to replicate the results.

The free and serial recall tasks are very well described in the scientific literature. You may use scientific articles, popular articles on the internet and large language models, like ChatGPT and Microsoft Copilot to help you designing the experiments. You may also use large language models to assist you in coding. You may *not* use large language models to write your report except for proof reading and help with grammar and vocabulary.

CHAPTER 2

Bayesian models of perception

2.1 Introduction

Perception often seems effortless. When we recognise the face of a friend it seems to happen easily, quickly and automatically. This can lure us into believing that perception is a simple process and that our perceptual experience is merely a reflection of the world as it is. This belief is far from the truth. Yet, an anecdote from the early years of AI research tells that this false belief lured pioneering researchers. In 1966 Seymour Papert set up a summer project for students at MIT. The goal of the project was to detect and identify objects from images. It took several decades for the field to reach useful solutions and the problem has still not been completely solved, so Papert seem to have underestimated the challenge.

From the earliest studies of human perception, it has become clear that perception is an extremely complex process. From the most recent studies we learn that we still do not fully understand it. What we do know is that perception depends not only on sensory information but also on the state of the observer, an understanding of the world around us and on perceptual biases.

Perception can be thought of as the brain's solution to an inverse problem: estimate the state of the world given sensory information. This type of problem is quite similar to the scientific endeavour of finding the state of the world given measurements. In both perception and in science such inverse problems are typically ill-posed or under-determined. This means that there is not enough data to find a unique solution. Yet, perception and science both solve this type of problems by adding information. This information comes in the form of models of how the world works and prior knowledge of the state of the world.

We can think of inverse problems in terms of Bayes' rule. Assume that the data can, in general, be represented by a vector, $\boldsymbol{d}$. The data could be the activation level of all the photo-receptors in the eyes, or it could be measurements in a scientific experiment. Likewise, assume that the state of the world can also be quantified by a vector, $\boldsymbol{w}$. Then Bayes' rule can be written as

$$P(w \mid \boldsymbol{d}) = \frac{P(\boldsymbol{d} \mid \boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{d})} \tag{2.1}$$

The problem of perception is to determine the posterior probability of the state of world given sensory information, or data, $\boldsymbol{d}$, and to determine the most probable state(s) of the world based on this posterior distribution. According to Bayes' rule in Equation 2.1, the posterior probability is proportional to the likelihood, $P(\boldsymbol{d} \mid \boldsymbol{w})$, and the prior probability, $P(\boldsymbol{w})$.

Typically, we are not interested in the state of the entire world but only a small part of the world. For example, we might be interested in whether there is a tiger hiding in the bushes. In this example, the sensory data could be the sound coming from the bushes: Does it contain the sound of a tiger's low rumble or not? The likelihood would then the probability of the particular sound coming from the bushes *given* that there is a tiger hiding there. Tigers sometimes rumble and sometimes do not. In order to determine the likelihood, we need to estimate the probability that they do, or, even better, that this particular, albeit hypothetical, tiger rumbles. The likelihood thus contains a probabilistic model of how the world works, or more precisely, how it generates sensory data.

The prior probability is the probability of the state of the world prior to any sensory input as opposed to the posterior probability, which can be determined only after sensory input. It could be the probability of a tiger hiding in the bushes in general and would be based on previous experience with tigers and bushes.

The denominator in Bayes' rule, $P(\boldsymbol{d})$ is sometimes referred to as the evidence. For some purposes, this term can be ignored. To see why, let us rephrase the problem of determining whether there is a tiger hiding in the bushes as an inequality

$$P(tiger \mid rumble) > P(no\ tiger \mid rumble)$$

This inequality implements the *maximum a posteriori* (MAP) decision rule: we decide that there is a tiger if this is the more probable outcome. If we insert Bayes' rule on both sides, we get

$$\frac{P(rumble \mid tiger)P(tiger)}{P(rumble)} > \frac{P(rumble \mid no\ tiger)P(no\ tiger)}{P(rumble)}$$

Note that the evidence, $P(rumble)$, is the only term that does not depend on the state of the world. It is therefore the same on both sides of the MAP decision rule inequality. We can thus multiply the by $P(rumble)$ on both sides and remove it from the inequality. In effect, we may ignore the evidence term when making decisions by comparing the posterior probabilities of various outcomes.

Another way to understand the role of the evidence is to rewrite it using the law of total probability so that Bayes' rule can be written in the form

$$P(\boldsymbol{w} \mid \boldsymbol{d}) = \frac{P(\boldsymbol{d} \mid \boldsymbol{w})P(\boldsymbol{w})}{\sum_{\boldsymbol{w}} P(\boldsymbol{d} \mid \boldsymbol{w})P(\boldsymbol{w})}$$

In this form, it is clearer that the evidence in the denominator can be thought of as a normalisation term ensuring that $\sum_{\boldsymbol{w}} P(\boldsymbol{w} \mid \boldsymbol{d}) = 1$, which must hold for any discrete probability distribution, in general.

2.2 The use of prior information in visual perception

2.2.1 The Necker cube problem

Depth perception, our ability to perceive a three-dimensional world from two-dimensional images, is a good example of how the visual systems seems to solve ill-posed problems in which there are fewer equations than unknown variables. Although binocular vision is important for depth perception, the visual system uses many other cues to determine distance. A good demonstration of this is that we can still perceive depth with one only one eye. Another good demonstration is the Necker cube illustrated in panel (A) of Figure 2.1. We perceive this two-dimensional drawing as a three-dimensional wire-frame structure. In solving this problem, the visual system thus seems to go beyond the sensory information given, although it still does not quite manage to solve the problem entirely, as it fails to find a unique solution. Instead it finds two possible interpretations illustrated in panels B-C of Figure 2.1. Strangely, the visual system's interpretation of the Necker cube seems to alternate between two solutions causing what is know as a bistable percept.

Analysing the Necker cube problem in terms of Bayesian inference is a good way to illustrate how the framework can be applied in general. The analysis will, however, rely on a number of simplifying assumptions. The first assumption is that we can specify the two-dimensional drawing of the Necker cube and the underlying three-dimensional structure in terms of the eight vertices only. Hence, the data, $\boldsymbol{d}$, is a set of eight two-dimensional coordinates stacked into vector form, and the world, $\boldsymbol{w}$, is a set of eight three-dimensional coordinates also stacked into vector from.

For this problem, the data, $\boldsymbol{d}$, and the world, $\boldsymbol{w}$, are continuous variables unlike the rumbling tiger problem described above in which the data, $\boldsymbol{d}$, and the world, $\boldsymbol{w}$, consisted of discrete states. We therefore write Bayes' rule as

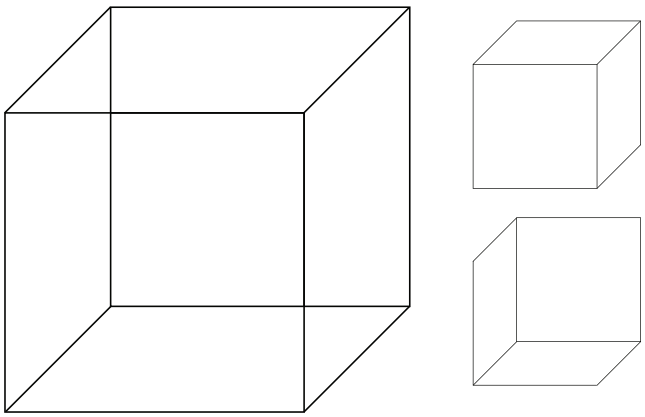

Figure 2.1: The Necker cube (Left) appears to be a three-dimensional wire-frame cube. The three-dimensional percept is bistable in that the cube can be perceived as two distinct three-dimensional structures (Top and bottom right). Figure by user Stevo-88, Wikimedia Commons..

$$p(\boldsymbol{w} \mid \boldsymbol{d}) = \frac{p(\boldsymbol{d} \mid \boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{d})} = \frac{p(\boldsymbol{d} \mid \boldsymbol{w})p(\boldsymbol{w})}{\int_{\boldsymbol{w}} p(\boldsymbol{d} \mid \boldsymbol{w})p(\boldsymbol{w})} \tag{2.2}$$

where the use of lower case p means that the term is a probability *density*.

As in the the rumbling tiger problem described above, the task of the observer is not to characterise the posterior probability distribution fully but only to find the most probable estimate(s) of the world, the *maximum a posteriori* (MAP) solution, $\boldsymbol{w}_{map}$, given the sensory data $\boldsymbol{d}$.

$$w_{map} = \arg\max_{\boldsymbol{w}} p\left(\boldsymbol{w} \mid \boldsymbol{d}\right) = \arg\max_{\boldsymbol{w}} p\left(\boldsymbol{d} \mid \boldsymbol{w}\right) P\left(\boldsymbol{w}\right)$$

Here we have inserted Bayes' rule and noticed that since the evidence term, $p(\boldsymbol{d})$, in the denominator does not depend on the world, w, can be ignored as we can multiply by positive terms that do not depend on the world with no effect on the MAP solution.

2.2.2 The Uninformative prior and the Maximum Likelihood Estimate

In the case of a non-informative, or *flat*, prior, all worlds are equally probable *a priori* so that $p(\boldsymbol{w})$ is a constant for all $\boldsymbol{w}$. In this case we can disregard the prior as it will not influence the solution. This solution is thus based only on the likelihood and is often referred to as the *maximum likelihood estimate* (MLE).

$$\boldsymbol{w}_{mle} = \arg\max_{\boldsymbol{w}} p\left(\boldsymbol{d} \mid \boldsymbol{w}\right)$$

The likelihood term describes how the two-dimensional image is generated. There are two steps in the process. In the first step, the three-dimensional world, $\boldsymbol{w}$, is projected onto the two-dimensional image plane. We will refer to the projection of each three-dimensional coordinate as μ_i. In the second step, Gaussian sensory noise is added to the data, so that each data point d_i is distributed as $d_i \sim \mathcal{N}\left(\mu_i, \sigma\right)$ where $\mu_i = \{x_1, \ldots, x_N, y_1, \ldots, y_N\}$ is the set of the N x- and y-coordinates of the projection. We assume that the noise is *independent and identically distributed* (IID) across points, which means that the covariance is zero and the variance, σ^2, of the noise is the same for all points. We can now write the likelihood function as

$$p(\boldsymbol{d} \mid \boldsymbol{w}) = \prod_{n=1}^{N} f(d_n, \mu_n, \sigma)$$

where f is the Gaussian probability density function (pdf)

$$f(d_n, \mu_n, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{d_n - \mu_n}{\sigma}\right)^2}$$

Working with the logarithm of the likelihood, the *log likelihood*, provides some simplification and also numerical stability. Note that this does not alter the MAP solution because the logarithm is a monotonically increasing function.

$$\boldsymbol{w}_{mle} = \arg\max_{\boldsymbol{w}} P(\boldsymbol{w} \mid \boldsymbol{d}) = \arg\max_{\boldsymbol{w}} \log P(\boldsymbol{w} \mid \boldsymbol{d})$$

Since the logarithm of the Gaussian pdf is

$$\log f(d_n, \mu_n, \sigma) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\left(\frac{(d_n - \mu_n)^2}{\sigma^2}\right)$$

the logarithm of the likelihood is

$$\log P(d \mid w) = \sum_{n=1}^{N} \log f(d_n, \mu_n, \sigma) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(d_n - \mu_n)^2$$

We can subtract and divide out any positive terms that do not depend on the world, $\boldsymbol{w}$, when we search for the $\boldsymbol{w}_{mle}$ solution. Here, only the projection, μ, of the world, $\boldsymbol{w}$, onto the image plane will depend on the world. Also, instead of maximising the logarithm of the likelihood we can minimise the negative of the logarithm of the likelihood.

$$\boldsymbol{w}_{mle} = \arg\min_{\boldsymbol{w}} \left(-\log p(\boldsymbol{d} \mid \boldsymbol{w})\right) = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N}(d_n - \mu_n)^2$$

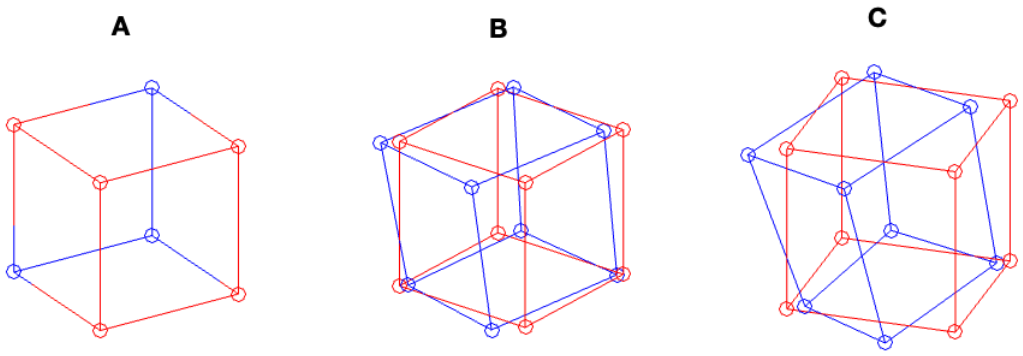

Figure 2.2: The Necker cube (A) as in Figure 2.1. The true three-dimensional structure is drawn in red while the MLE solution is drawn in blue. Viewed in the projection plane the solution and the true structure seems identical. When viewed from an angle shifted 20° (B) or 40° in the azimuthal plane the two structures no longer overlap..

An MLE solution can thus be found simply by minimising the sum of squared errors between the sensory data, $\boldsymbol{d}$ and the projection, $\boldsymbol{\mu}$, of the world, $\boldsymbol{w}$ onto the image plane. Note that the solution is not unique. We can easily see that by assuming that the image plane is the z-plane (and that the the size of the sensory image is small compared to the size of the eye). This assumption does not incur any loss of generality as we are free to choose our coordinate system so that the z-plane aligns with the image plane. In this case the x- and y-coordinates of the projection are simply the x- and y-coordinates of the three-dimensional points. There are no constraints on the z-coordinate, which can take on any value. MLE solutions will therefore appear as perfectly good solutions when viewed from the image plane but will generally prove to be poor solutions when viewed from other planes. This is illustrated in Figure 2.2

2.2.3 The informative prior and the Maximum Posterior Solution

We are now ready specify a prior on the world, $\boldsymbol{w}$. The prior probability density, $p(\boldsymbol{w})$, is a probability density over all possible three-dimensional shapes with eight vertices. We could specify that as a 24-dimensional probability density over all possible sets of eight three-dimensional points. We can, however, also specify the prior on other attributes, of the three-dimensional shape. For illustrative purposes, we will specify a prior probability density on the 24 internal angles, $\theta_1, \ldots, \theta_M$ of the structure, so that . This prior is a prior towards angles of 90°, i.e. away from flat and pointy structures, which would have angles closer to 0° or 180°. We can implement this prior using a probability density that is maximal at 90°, symmetric, and minimal at 0° and 180°. We could use the von Mises distribution, which is defined on the circle but, we will, for simplicity, use a Gaussian distribution centered at $\mu_p = 90$ since it is a reasonable approximation to the von Mises distribution when the standard deviation, σ, is small. We will also assume that the density over the angles are IID so that the prior can be written as

$$p\left(\boldsymbol{\theta}\right) = \prod_{m=1}^{M} f(\theta_m, \mu_p, \sigma_p)$$

Recall that we only need the numerator of the posterior when we seek the w_{map} solution. We can now express the logarithm of the numerator as

$$\begin{aligned}\log\left(\left(\boldsymbol{d} \mid \boldsymbol{w}\right) p\left(\boldsymbol{\theta}\right)\right) &= \log p\left(\boldsymbol{d} \mid \boldsymbol{w}\right) + \log p(\boldsymbol{\theta}) \\ &= -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(d_n - \mu_n\right)^2 - \frac{M}{2}\log\left(2\pi\sigma_p^2\right) - \frac{1}{2\sigma_p^2}\sum_{n=1}^{M}\left(\theta_n - \mu_p\right)^2\end{aligned}$$

When we seek the MAP solution we can subtract and divide by all terms that do not not depend on the world, $\boldsymbol{w}$, so that we arrive at

$$\begin{aligned}w_{map} &= \arg\min_{\boldsymbol{w}}(-\log p\left(\boldsymbol{w} \mid \boldsymbol{d}\right)) \\ &= \arg\min_{\boldsymbol{w}}\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}\left(d_n - \mu_n\right)^2 + \frac{1}{\sigma_p^2}\sum_{m=1}^{M}\left(\theta_m - \mu_p\right)^2\right)\end{aligned}$$

We can thus find the MAP estimate by minimising a *weighted* sum of squares in which the likelihood terms are weighted by the *reliability*, $r = \sigma^{-2}$, and the prior terms are weighted by the reliability, $r_p = \sigma_p^{-2}$. Note that if the variance of the sensory noise, σ^2 is large then the reliability, r, of the sensory information is small and the likelihood term, which pertains to sensory information will, accordingly have a less influence on the MAP estimate, w_{mle}. This would correspond to the case where the observer caught only a fleeting glimpse of the object and would therefore rely more strongly on prior assumptions. The reliability, r_p, of the prior is not related to noise but to how informative or *strong* the prior is. If the reliability of the prior is very small, then the prior distribution would be near flat, or uninformative.

The observer would therefore have to rely on sensory information. Conversely, if the reliability of the prior is high, then the prior distribution is narrowly centered around the mean so that estimates of the state of the world that deviates only slightly from the observer's prior assumptions are deemed highly improbable. The trade-off between sensory and prior information can be captured by the ratio $\frac{r_p}{r}$ of the two reliabilities allowing us to express the MAP estimate as

$$\boldsymbol{w}_{map} = \arg\min_{\boldsymbol{w}} \left(\sum_{n=1}^{N} (d_n - \mu_n)^2 + \frac{r_p}{r} \sum_{m=1}^{M} (\theta_m - \mu_p)^2 \right)$$

Note that we can define a prior on any feature of the three-dimensional world. We could, for instance, define a 90° prior on the 24 angles of the structure, so that $a_m = 90°$ for all m. We also need to define a value for $\frac{r_p}{r}$. If the value chosen is too small, then the prior term will have little effect on the MAP solution, which therefore will be similar to the MLE solution shown in Figure 2.2. If it the value chose is too large, then the prior term will dominate, which results in solutions with angles of 90° as dictated by the prior but at the cost of ignoring the sensory data, d, in the likelihood term prior. Therefore, their projection images bear little resemblance to the projection of the true structure from any angle. This is illustrated in Figure 2.3. For appropriate values, the MAP solution can be a perfect fit so that its projection image overlaps with the projection image of the true structure from any viewing angle. In this case, the MAP solution for the three-dimensional structure is consistent with what the three-dimensional structure that humans perceive from the two-dimensional image.

A prior must be based on prior experience and cannot be based on the (sensory) data. Choosing a 90° prior for a cube only serves to demonstrate the effect of prior information. It is unlikely that the efficacy of the 90° prior will generalise to other three-dimensional structures such as the hexagonal prism in Figure 2.1. However, an open angle prior, centered at around 100°, works for both the Necker cube and the hexagonal prism, for appropriate values of $\frac{r_p}{r}$. This indicates that such a prior may play a role in visual perception.

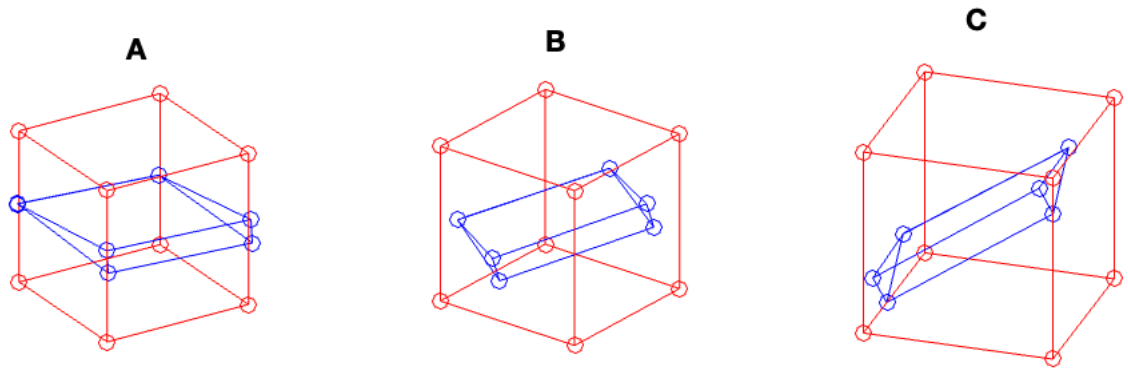

Figure 2.3: The Necker cube (A) as in Figures 2.1. The true three-dimensional structure is drawn in red while the MAP solution is drawn in blue. Here the value of $\frac{r_p}{r}$ is too large so that the prior term dominates and sensory information is igored. Therefore, although the internal angles of the solution are very close to 90°, as dictated by the prior, the projection of the MAP solution appears to not overlap with the the projection of the true structure regardless of the viewing angle..

2.2.4 The light from above prior

Shading provides important cues to depth perception. This is illustrated in the two panels of Figure 2.4, which our visual system interprets as three-dimensional structures. The panel to the left show a number of objects that appear to be hemispheric protrusions except for one object that appears to be a hemispheric cavity. Contrary, the panel to the right show a number of objects that appear to be hemispheric cavities except for one object that appears to be a hemispheric protrusion. The two figures are, in fact, identical except for a 180° rotation. You can try to view the figure upside-down: The objects that appeared to be protrusions when viewed upright now appear to be cavities and vice versa. This raises the question of why the orientation in the viewing plane so greatly affects the visual system's interpretation of depth.

The three-dimensional interpretation of the objects in Figure 2.4 may be influenced by the *light from above prior* meaning that the visual system assumes that the light source is located above the object. Evolutionary, this seems like a reasonable prior assumption in that the main source of light during most of evolution has been the sun, which is indeed located above any object. Assuming that light comes from above agrees with the interpretation of a protrusion being bright at its top because it is lit from above and dark at its bottom because of its own casting shadow. Conversely, a cavity is bright only at the bottom where light can shine through the cavity but dark at the top, which lies in a shadow. Note that objects that appear as protrusions might as well be cavities lit from below. Although we can consciously convince ourselves that this is a reasonable interpretation, we cannot *see* it because we cannot persuade our visual system of this interpretation.

2.2.5 Exercise - The Necker cube

In this exercise we will solve the Necker cube problem using the Bayesian approach described in Sections 2.2.1-2.2.3. The problem is to infer the world, $\boldsymbol{w}$, which is a three-dimensional wire-frame structure from the data, $\boldsymbol{d}$, which is a two-dimensional image. The Bayesian approach is to maximize the posterior probability $p(\boldsymbol{w} \mid \boldsymbol{d})$. For practical reason, we will minimize the negative logarithm of the posterior probability instead as described in Sections 2.2.2-2.2.3. First, we will assume an uninformative prior and let the squared error correspond to the likelihood function as described in Section 2.2.2.

We will use Matlab in this exercise. You must place all the Matlab files that come with the exercise in your active directory. Open the Matlab function negLogPosterior.m. You only have to fill out the argument of the sum of squares that will calculate the negative log posterior using the variables w and d. The variable w is a 3-by-8 matrix of the estimate of the three-dimensional world. The variable d is a 2-by-8 matrix of the perceived image. Each column in both w and d holds the coordinates of one corner. You can also use the function Project(w,M) to calculate the projection of the three-dimensional world onto the two-dimensional viewing plane. The variable M is a 2-by-4 projection matrix. Once you have you have completed the negLogPosterior.m, you can run the NeckerExercise.m function and it will show a plot of your solution and the true cube.

Inspect the figure visually. View it from different angles by clicking and dragging the figure. Describe the solution you get. Is it a correct solution? There are, at least, two ways that it can be the wrong solution. What are they? Note that the solution you get can depend on the random initialisation, so you might want to try more than one solution.

Next, try using an informative prior. Use the calculation of the angles (in degrees) of the three-dimensional structure calculated in the variable Angels (a prior for right angles will work well for the cube). Again, inspect the figure visually. View it from different angles. Try this for different values of $\frac{r_p}{r}$ as described in Equation 2.2.3. Describe the solutions you get when the value is too low, too high and just right. Which value seems just right?

Do the same thing with a hexagon cylinder to test whether the prior you came up with for the cube also works for a different structure. Open the Matlab function HexagonalExercise.m. It works just like NeckerExercise.m function. First try the uninformative prior that you used to solve the Necker cube problem. Just as for the Necker cube problem you can do this by editing the negLogPosterior.m file. The results should be similar to the results for the Necker cube problem. Then try the prior that you

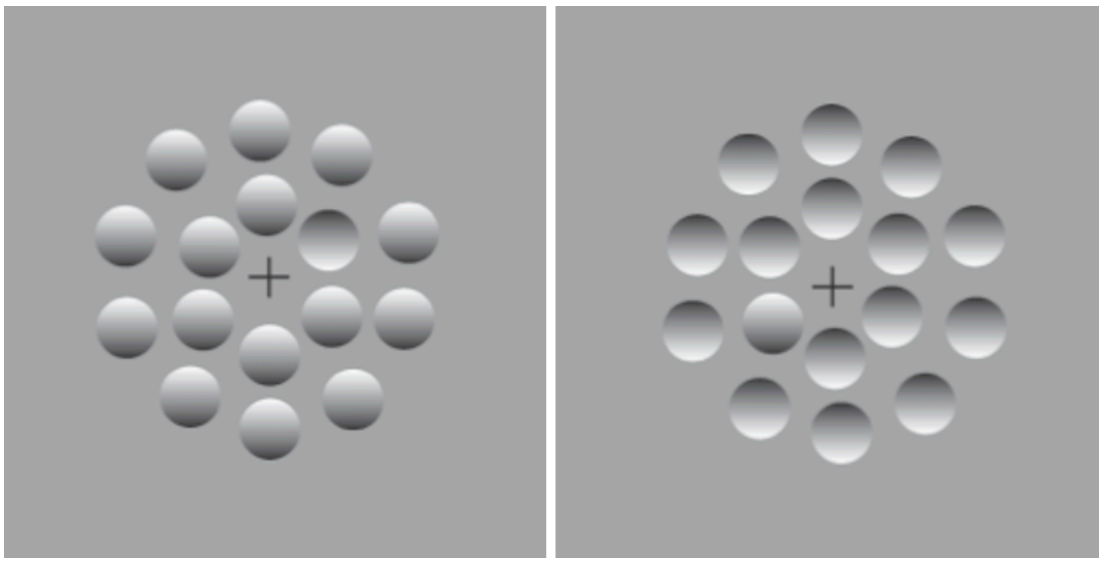

Figure 2.4: Our visual system use shading cues to interpret the above figures as three-dimensional. The panel to the left show a number of objects that appear to be hemispheric protrusions except for one object that appears to be a hemispheric cavity. Contrary, the panel to the right show a number of objects that appear to be hemispheric cavities except for one object that appears to be a hemispheric protrusion. The two figures are, in fact, identical except for a 180° rotation. This shows that the visual system's interpretation of the three-dimensional structure depends on the orientation of the object in the viewing plane. This interpretation may be due to the visual system relying on a prior assumption of light coming from above. Figure © by Champion and Adams (2007), Journal of Vision..

used for the Necker cube problem. Describe the results. Try to find a prior that works for the hexagon problem. Describe your approach. Try finding a prior that works for both the Necker cube and the Hexagon. Describe your approach.

2.3 The use of marginalisation in perception

2.3.1 Depth from shading

So far, we have learned how our visual system use prior assumption to interpret sensory information. To demonstrate this we have focused on the visual system's ability to perceive depth in two-dimensional figures. In this section we will study *marginalisation* as another method that the visual system may use to interpret sensory information. To demonstrate this, we will continue focusing on the visual system's ability to perceive depth from two-dimensional figures.

First, we will revisit Figure 2.4. The light from above prior seemed a reasonable explanation for how the visual system can distinguish between hemispheric protrusions and hemispheric cavities, but recall from Section 2.2 that monocular depth perception is a highly ill-posed problem in that there is an unknown variable for every point in the object. In the case of the Necker cube, which we parameterized using the eight vertices, we had eight unknowns. In order to find a reasonable MAP solution, we used a high-dimensional prior. This did not solve the problem completely since the Necker cube remains a bistable percept. How many points do we need to specify that an object is a hemispheric protrusion or cavity? Perhaps a four-by-four grid of three-dimensional points would suffice. In that case we would have 16 unknowns and it should be a bit surprising that even without using prior information on the

orientation of the light source, we can only see the object as a hemispheric protrusion or cavity when in fact there should be infinitely many three-dimensional objects that could create the sensory image. This is illustrated in Figure 2.5 where Panel A shows an image that appears to show a hemispheric protrusion lit from the left or a hemispheric cavity lit from the right. These two solutions are depicted in Panel B (1 and 5) along with three other solutions (2-4) that are not hemispheric shapes, yet they will generate the sensory image when lit from the direction indicated by the arrow next to the object. Why does our visual system not offer these structures as possible causes of the sensory image?

We can phrase the problem of inferring depth from shading in terms of Bayes' rule by splitting the world, $\boldsymbol{w}$, into two components that influence the sensory data: the shape of the object, which we will here denote as $\boldsymbol{w}_1$ and the orientation of the light source, which we will denote as $\boldsymbol{w}_2$. Then, Bayes' rule can be written as

$$p(\boldsymbol{w}_1, \boldsymbol{w}_2 \mid \boldsymbol{d}) = \frac{p(\boldsymbol{d} \mid \boldsymbol{w}_1, \boldsymbol{w}_2) p(\boldsymbol{w}_1, \boldsymbol{w}_2)}{p(\boldsymbol{d})} \tag{2.3}$$

Since the posterior has two components, as in Equation 2.3, the most probable shape (the MAP estimate of the shape) depends on the orientation of the light source. One shape might be the most probable shape given a certain lighting angle, another shape will be the most probable given another lighting angle, so we might not be able to find a single best solution. What we are really interested in is estimating the posterior probability $p(\boldsymbol{w}_1 \mid \boldsymbol{d})$ and we can do this by integrating or, *marginalising*, over the lighting angle, $\boldsymbol{w}_2$

$$p(\boldsymbol{w}_1 \mid \boldsymbol{d}) = \int_{\boldsymbol{w}_2} p(\boldsymbol{w}_1, \boldsymbol{w}_2 \mid \boldsymbol{d}) \tag{2.4}$$

Equation 2.4 tells us that the most probable shape, is the shape that is most probable *across* all viewing angles, w_2. Marginalising over the viewing angle, w_2, can help us find this solution even when the prior is uninformative, meaning that all shapes and viewing angles are equally probable *a priori*. In that case we can insert Bayes' rule from Equation 2.3 into Equation 2.4 to get

$$p(w_1 \mid d) = \int_{w_2} p(w_1, w_2 \mid d) = \int_{w_2} \frac{p(d \mid w_1, w_2)}{p(d)} \tag{2.5}$$

Equation 2.5 tells us that the most probable shape is the shape that would create the sensory image even when seen across all viewing angles. Panel C shows five sensory images that a shape would generate across a small range of lighting angles, shown as arrows below the sensory images. Note how the sensory image varies far very much away from the sensory image in Panel A and in the center sensory image in Panel C. Hence, although the likelihood of that shape would be high for one particular lighting angle, it would be quite low for other lighting angles, so the integral in Equation 2.4 would take on a fairly low value. This might be the reason for why the visual system concludes that this shape is improbable. Compare these observations to the sensory images that the hemispheric protrusion in Panel D generates across varying lighting angles. These sensory images are all quite similar to the sensory image in Panel A. Hence, the likelihood of that shape would be high across a broader range of lighting angles, so the integral in Equation 2.4 would take on a fairly high value. This might be the reason for why the visual system concludes that this shape is more probable.

The marginalisation approach described here is similar to the use of the *Bayes' factor* for model evaluation in machine learning. The Bayes factor is the ratio of the marginal likelihoods of two models, M_a and M_b

$$\frac{P(D \mid M_a)}{P(D \mid M_b)} \tag{2.6}$$

A Bayes factor much greater than 1 indicates that model M_a is better whereas a Bayes' factor much less than 1 indicates that model M_b is better. Note that the Bayes factor is different from the maximum likelihood estimate that we previously encountered because the likelihood terms in Equation 2.6 are

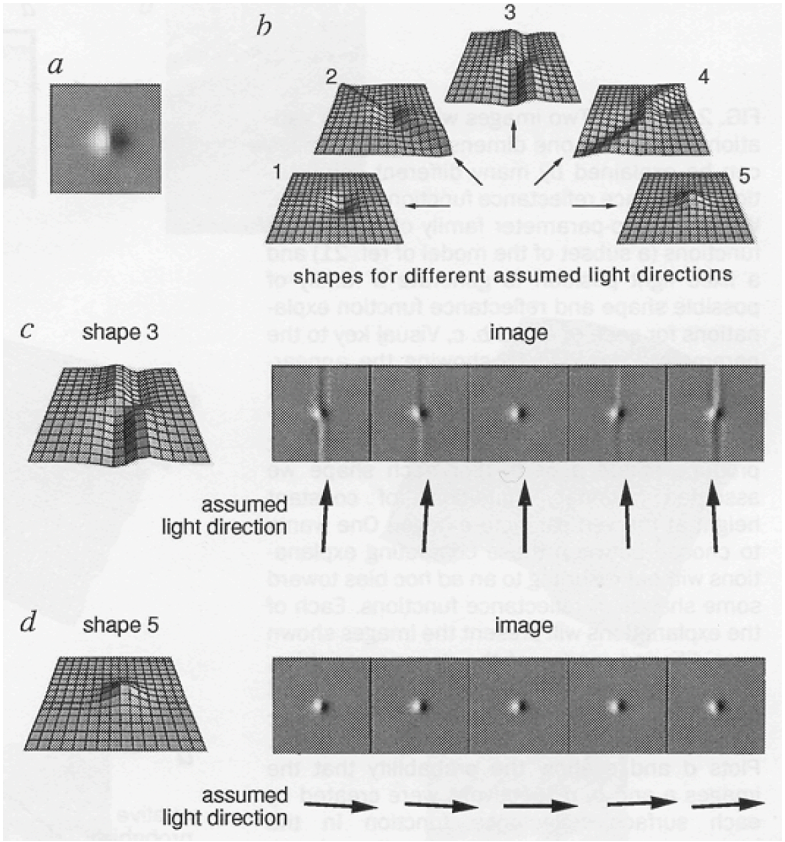

Figure 2.5: Our visual system interprets the sensory image in (a) as a protrusion lit from the left or a cavity lit from the right. All the objects in (b) will, however, produce the same sensory image when lit from the direction indicated by the arrows. When the object (shape 3) in (c) is lit from a narrow range of angles it produces sensory images very unlike the sensory image in (a). Therefore, it is not a good solution when evaluated by marginalising over the lighting angle. Contrary the object (shape 5) in (d) produces sensory images very similar to the sensory image in (a) across the same range of lighting angles. Therefore, it is a good solution when evaluated by marginalising over the lighting angle. Figure © by Freeman (1994), Nature.

marginalised across the model parameters. If models M_a and M_b are parameterised by parameter vectors θ_a and θ_b then the Bayes factor can be written as

$$\frac{P(D \mid M_a)}{P(D \mid M_b)} = \frac{\int_{\theta_a} P(D \mid \theta_a, M_a) P(\theta_a \mid M_a))}{\int_{\theta_b} P(D \mid \theta_b, M_b) P(\theta_b \mid M_b))} \tag{2.7}$$

In practice, these integrals are typically very complex and must be estimated numerically, typically using Markov Chain Monte Carlo methods if precise estimates are needed. The visual system could, however, use coarse approximations by summing over only a few selected values.

Comparing the Bayes factor approach with the marginalisation approach described above, note that the models, M_a and M_b would be models of the world, w, each consisting of the lighting angle and the three-dimensional structure. Whereas the Bayes' factors integration across all parameters only makes sense for two models that are parameterised differently, the marginalisation approach described here distinguish between two models that are parameterised in the same way and therefore only integrates across the subset of parameters that are irrelevant. These are called *nuisance parameters* and would the

lighting angle in the example above. Common to the marginalisation approach and the Bayes' factor is that they seek solutions that are robust to variations in the parameters.

2.3.2 Impossible triangle

The *impossible triangle* illustrated in the lower left of Figure 2.6 serves as another example of how the visual system might use a marginalisation approach. Only in this case, the object has been carefully designed to trick the marginalisation approach to create an illusory percept.

Our visual system interprets the impossible triangle as a closed triangular shape but our conscious reasoning soon tells us that this is impossible. The figure is however a photograph and the object is very much real. The objects' mirror reflection shows us the object's true three-dimensional structure: it is not a closed triangular shape. Our visual system has come up with a wrong interpretation of the object as a closed triangular shape is wrong, but why?

As in the problem of inferring depth from shading, we will again phrase the problem in terms of Bayes' rule by splitting the world, $\boldsymbol{w}$, into two components, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, so that Bayes' rule is given by Equation 2.3. Again, $\boldsymbol{w}_1$ denotes the three-dimensional structure of the object, but now, $\boldsymbol{w}_2$ denotes the viewing angle. If the visual system uses the marginalisation approach described by Equation 2.5 it should find an object that will produce the sensory image of a closed triangular shape when viewed across a wide range of angles rather than an object that will produce the sensory image of a closed triangular shape only when viewed from a narrow range of angle. Hence the solution of the visual system is a closed triangular shape: Even though it is impossible, it would indeed produce the sensory image across a wide range of viewing angles. In contrast, the real object revealed by the mirror reflection in Figure 2.6 will produce a sensory image of a closed triangular shape only when viewed from a narrow range of angles. This is called the *generic viewpoint assumption*. The visual system assumes that the viewing angle was generic rather than very particular.

Figure 2.6: The object in the lower left, appears to be an "impossible triangle", an illusionary object that appears to be closed triangular from but violates the laws of geometry. Although our visual system interprets the object in this way, our conscious reasoning can deduce that it is impossible for this object to exist. The figure is, however, a photograph of a real object. The real three-dimensional structure is revealed by the objects mirror reflection showing that the figure is not actually a closed triangle contrary to what we perceive when viewing the figure from a particular angle. Photograph © by Bruno Ernst.

2.3.3 Exercise - The use of marginalisation in visual perception

Solve the simple perceptual task facing a monkey in the jungle using a Bayesian approach. Spell out which terms are posteriors, priors, likelihoods, what terms are being discounted. You can assume that the reflection spectres follow Gaussian (normal) distributions.

1. Out in the jungle, 15% of the juju-fruits are ripe. Ripe juju-fruits are orange, on average reflecting light with a wavelength of about 600 nm with some variation (standard deviation of 50 nm). Unripe juju-fruits are green with a wavelength of 500 nm (standard deviation 50 nm). What is the probability of a juju-fruit reflecting light with a wavelength between 540-550 nm is ripe? Note that you can calculate the probability of a fruit having a wavelength, w, smaller than some value using the cumulative Gaussian distribution function (norm.cdf in Numpy or normcdf in Matlab).
2. Only 10% of the fruits in the jungle are juju-fruits. 50% are mongo berries. 80% of the mongo berries are ripe. When ripe, mongo berries reflect light with a wavelength of 580 nm (standard deviation 20). When unripe, they reflect light with a wavelength of 520 nm (standard deviation 20). The remaining fruits are all chakavas. Only 10% of the chakavas are ripe. When they are ripe, they reflect light with a wavelength of 400 nm (standard deviation 100). When they are unripe, they reflect light with a wavelength of 550 nm (standard deviation 100). What is the probability that a random fruit is ripe if it reflects light with a wavelength between 540-550 nm?

