MLDS Ethics - Part 1

Zak Varty

2022-10-05

Contents

Welcome!

Data-driven decision making is now pervasive and impacts us all. Your data is used by others to make decisions about who you are, how you will behave, and what options should be made available to you. Predictive models are used to decide anything from the promotion that is offered to you by a retailer through to whether your loan application is granted by a bank.

The ways in which these predictive models can fail *mathematically* form a core part of the training for an aspiring statistician, data scientist or machine learning practitioner. In contrast, the potential for *ethical* failures in these same models is rarely covered in-depth during as part of this initial training. As a result, these ethical modes of failure are often not considered until those predictive models have been put into production and are actively causing harm.

To prevent this harm, the ethical impacts of using data to make decisions must be made core to the curriculum of both statistics and data science. This course aims to address that gap.

The course takes a practical and technical approach to identifying these ethical issues. It has a strong mathematical focus and will not not require the authoring of extended essays or moral treaties. Throughout the course, we give actionable ways in which these topics may be integrated into a data science workflow at a range of levels.

Module Description

This module will investigate the ethical implications of the new capabilities offered by Data Science and Artificial Intelligence.

Part 1 will begin by discussing the ethical use of data itself - the raw materials of data science pipelines. It will then discuss sets of principles that tech leaders and international bodies are adopting to promote ethical use of data science and artificial intelligence algorithms, including a discussion of real-world examples of failings and adverse outcomes.

Parts 2 and 3 will then revisit the issues explored in Part 1 in greater technical detail. These parts will introduce data science methodologies that provide novel solutions to ethical problems of old such as explainability, prejudice and bias.

Learning Objectives

On successful completion of this module, you should be able to:

1. Recognise and accept responsibility for the societal impact of data science and machine learning technologies;
2. Participate in the broader debate about the issues surrounding the use of data science and machine learning for prediction, decision making and knowledge generation tasks;
3. Identify common ethical pitfalls of data science and ML algorithms via a mental "check-list" and evaluate the degree to which a given algorithm is likely to conform with ethical best practices;
4. Formally test for common ethical pitfalls of data science and ML algorithms;
5. Implement mitigation measures against the ethical risks posed by the use of data science and ML algorithms;
6. Construct well-founded and evidence-based arguments with which to positively influence the actions of stakeholders and decision-makers;
7. Use a systems perspective to holistically appraise data science projects on their ethical and societal impacts.

Contributors

These notes are structured around a course delivered as part of the the Master's degree in Machine Learning and Data Science at Imperial College London, which was developed by Christoforos Anagnostopoulos and Zak Varty.

These course notes were written by Zak Varty and are still under active development. If you spot any issues or would like to contribute to their development, you may raise an issue or submit a pull request to the associated github repository.

Course Overview

This section is only relevant to students taking the MLDS course in the academic year 2022-23.

0.1 Assessments

Table 1: Assessment schedule for Ethics Part 1 (2022 Cohort)

Assessment Type	Description	% of Ethics Module	Release Date	Due Date
Reading Summaries	Weekly summary of one assigned reading and peer-feedback for two other students.	5	TBC	TBC
Coursework	Individual short report. This will involve a mixture of questions and guided case-studies to assess technical understanding of the course content alongside its implementation and limitations when used in context.	15	TBC	TBC

0.2 Reading summaries

There are a wide range of areas in which the use of AI has lead to harm for an individual or a section of society. In lectures we heard of an example of individual harm, through the advertisement of maternity products to a woman who has miscarried. We also heard an example of group harm, through pretrial assessments that incorrectly predicted that Black defendants were high risk individuals. For this assessed question you will consider another example in which the use of AI has caused harm.

A rhetorical precis is a short summary and analysis of a piece of writing, which considers both the content and the delivery of the piece. A rhetorical precis includes an accurate bibliographic reference to the text, a list of keywords relating to the text, and a highly structured four-sentence paragraph which serves to summarise and analyse the text. Each sentence has a specific role:

1. The first sentence establishes the aim or thesis or the work;
2. The second sentence explains how this aim is addressed or outlines supporting arguments;
3. The third sentence states the purpose of the work (why it is important);
4. The final sentence describes the intended audience and tone of the writing.

An example of a rhetorical precis can be found on the lumenlearning website. Creating such summaries of a paper promotes clarity and precision both when reading and writing. It also provides a useful aid when trying to recall the contents of a paper long after you originally read it.

Select and read an academic article that is *not on the course reading list* and *either}*highlights one harmful instance of AI *or* summarises the potential harms of AI more generally. Write a rhetorical precis for your selected paper, and submit this for assessment. (Note: using a text from the course materials or reading list will result in a 50% reduction in the marks awarded for this question.)

- Accurate bibliographic reference and keywords [1]
- Aim or thesis sentence [2]
- How aim is addressed / thesis argued [2]
- Purpose or importance of this work [2]
- Identified target audience [2]
- Selection of a paper not included in course reading list [1].

Chapter 1

Foundations of Ethical AI

Introduction

Welcome to the first week of this unique course.

This will likely be different to any other course you have taken before if you come from a scientific or engineering background. We will often spend a few days without seeing any equation, and we will be asked to think carefully about the real-world implications of our work.

Having said that, rest assured that there will be plenty of technical and mathematical content in the course, and any content that is purely conceptual will be immediately relevant to your life as a professional data scientist.

This is not a course in moral philosophy, but I really hope it will encourage you to take one. And with that, let's get started.

This week, we'll spend our time getting to understand precisely what we mean by harm, what we mean by data science itself, what moral frameworks and codes of conduct are already in place, and finally codify them in a set of five principles that we can use as guardrails for our own work as professional data scientists.

Although the term AI is not in the course title, I will sometimes collectively refer to our subject as ethical AI. That's not necessarily very accurate. There is a tendency these days to label even simple data analysis as AI and that can be misleading. Also, there are flavours of AI that do not use any data or learning. However, the literature and regulation around the topic of professional ethics is increasingly consolidating under the term ethical AI, so I will use that as short cut.

You can safely assume that whenever I use the term AI, I am referring to statistical machine learning and data science, which is the focus of the program you're attending.

All right. Disclaimers aside, this week we will cover a huge amount of conceptual ground.

It will also be the only week in this entire course that does not have mathematical or programming content, so sit back and enjoy.

1.1 Do No Harm

1.1.1 AI for good

1.1.1.1 Go

It is hard to open a news website these days and not come across a headline that celebrates the incredible achievements of AI algorithms. Millions of people held their breath while DeepMind's neural network AlphaGo played a game of Go, the world's most challenging strategy board game, against the world champion, Eventually landing a victory and ushering in the era of AI supremacy in yet another frontier of human intelligence.

What makes this moment dramatic is that once a computer becomes a world champion at a certain game, say chess or go, it means that the basic algorithmic design, allowing it to learn how to play that game is finally in place.

And given year-on-year increases in computational power, the computer can get better and better at a much faster pace than a human can. Put differently, a human will never again manage to out compete state-of-the-art AI in chess or go.

A common criticism of AI used to be that it can only outperform humans in toy domains or games which, though fascinating, does not really help humanity. We also used to criticise AI systems for being very narrow, for example, only able to deal with one task at a time. Now, that era seems to be over.

1.1.1.2 Protein folding

DeepMind recently demonstrated that an algorithm architecturally similar to AlphaGo was able to solve a riddle in biochemistry that holds the promise of revolutionising drug discovery. This is known as protein folding, and it is the task of predicting the 3D shape of a protein on the basis of the sequence of amino acids that it is made of.

Proteins are very complex twisty things and the way they fold depends on a number of spontaneous interactions between different parts of their amino acids. So it is a very hard computational problem to solve exactly.

AI researchers have instead opted for predictive modeling using a large database of known structures as a training data set and then predicting what the likely shape of new proteins will be.

1.1.1.3 Arrhythmia

AI is now reaching superhuman performance not only in games and scientific problems in the lab, but also in everyday real-world tasks that currently require a well-trained professional.

For example, cardiologists can detect arrhythmia by inspecting the echocardiogram of a patient. That's a graph of the electrical activity of the heart.

Recent work by Andrew Ng and others demonstrated that an AI algorithm is able to perform that task with similar performance but can do so in milliseconds everywhere in the world, just as reliably.

The implications for remote health care or care of patients in countries without robust health care systems and absence of specialist doctors are mind-blowing. And yet, increasingly we also hear news stories about AI getting it wrong.

1.1.2 AI can cause harm

1.1.2.1 Self-driving

Self-driving cars are a particularly interesting example of ethical AI.

That is because driving is a relatively easy task for humans; we learn how to do it in a matter of months and most of us can do it reasonably well. Yet, to an Artificial Intelligence system, driving is much harder than chess.

It requires very advanced computer vision, planning and movement forecasting. These are all things that humans and most animals excel at because we evolved over millions of years to move while avoiding objects under various conditions, including rain, low visibility in a range of different physical environments.

Similarly, driving safely requires knowing that a dog is more unpredictable than, say, a pedestrian or that an elderly man typically walks slower than a teenager. We didn't learn any of those facts during driving lessons. To know how to drive a car, we need to understand much more than just cars, roads, and roadsides.

Despite these huge challenges, self-driving cars remain an incredibly valuable business proposition. Therefore, many tech companies and car manufacturers are competing in this space.

Inevitably, a self-driving car will occasionally enter into an accident and some of these will be fatal. This is not a statement about the relative safety of self-driving AI versus human driving. Current statistics suggest that existing self-driving cars are much safer than humans on a scale of number of accidents per mile drive, and that trend is likely to persist.

And yet, when a fatal accident does happen, who is to blame? Could it have been avoided?

Is the data scientist that worked on the algorithm to blame?

Is it the model's fault or the data's fault? If not, then who or what is to blame?

These are hard and important questions that we have never had to ask before. There are also other subtler but equally important ways in which AI can cause harm.

As AI is increasingly used in our daily lives, it will make mistakes and without care, these mistakes may disproportionately affect minorities, vulnerable populations or groups that have been the subject of historic discrimination.

1.1.2.2 Facial recognition

Facial recognition is one particular example where a combination of poor quality data and poor governance have led to headline grabbing failures.

These news articles (and the academic articles on which they were based) draw attention to the fact that these algorithms reliably perform worse on faces of people of colour.

These mistakes can be hurtful and perpetuate historic racism. They can also lead to real-world discrimination in cases where the algorithms are used, for example by the police.

So here, we are at a crossroad of sorts. After decades of broken promises, AI is starting to be powerful enough to solve real-world problems and even outperform humans in a number of valuable tasks. However, as adoption of these technologies increases so does the risk of them doing harm.

Whereas before, AI researchers were playing harmlessly with toy examples in university labs, they are now increasingly holding lives and livelihoods in their hands.

This is not the first time this happens. In fact, it happens with nearly every powerful emerging technology.

1.1.3 Technological Adoption Requires Public Trust

1.1.3.1 Nuclear Power

One particular example of this that we'll come back to is nuclear power.

Before the nuclear bomb, theoretical physicists were completely protected in their academic ivory towers, both unwilling and unable to affect the mess that is the real world.

Suddenly, the right set of equations and a wartime effort to make use of them changed all that. Iconic names like Richard Feynman and Albert Einstein suddenly got thrown into one of the most heated moral debates of the century.

Even putting aside the question of nuclear weapons, nuclear power has struggled to secure public trust even when used for peaceful purposes like energy production. At a time of climate and energy crisis, nuclear power is one of the few truly sustainable sources of energy we have, but it comes with a very considerable safety problem.

Although, nuclear disasters are few, when they do occur, they can be devastating and particularly terrifying, and they do not only happen in struggling economies where failures might be attributed to poor practices. They can even happen in technologically advanced economies like Japan.

Trust is hard to earn and easy to lose, and each such headline makes widespread adoption of nuclear power exponentially harder. A similar future might await AI unless we act now.

So, whose responsibility is it to ensure AI fares better than that? Well, we would expect the physicists and engineers to carry the burden of responsibility of explaining exactly how safe a nuclear reactor is, and how it needs to be maintained for it to remain safe, because they're the only ones that understand it well enough.

1.1.3.2 Taking responsibility

Similarly, data scientists like yourselves are the first line of defence against harmful AI because you are the ones that understand these systems well enough to anticipate what might go wrong with them.

Just like theoretical physicists in the early 20th century, data scientists before the onset of big data did not have to worry about this as their work had limited impact and only in very specific conditions.

Now with data science in AI increasingly becoming a part of the very fabric of society and the economy, we all do carry a burden to demonstrate its safety.

As data scientists, we not alone in this; many other professions carry a similar burden of responsibility and have to abide by strict codes of conduct and are legally liable for the outcomes of their work.

1.1.4 Learning from other professions

Doctors and drug design researchers are one example, which is very prominent in a pandemic era. There are also other,more quotidian, examples that we rarely think about including lawyers and the engineers that build the infrastructure we use on a daily basis like the houses we live in, the power network, the transport network, and the bridges that we hope don't fall.

All of these professions know that lives depend on them getting things right and they hence adopt a safety-first approach to risk-management. This is what we are inviting you to adopt in your own work.

Doctors are a particularly interesting example in that, from antiquity, they have been very explicit about their moral obligations.

1.1.4.1 Hippocratic Oath

As early as 2,500 years ago, doctors would swear by the Hippocratic oath, which among other things included a commitment to medical confidentiality and non-maleficence, or as it is commonly known, “do no harm”.

The idea is that a doctor should not take unnecessary risks when considering treatment options for a patient. This simple command has had tremendous influence throughout the centuries in our understanding of professional ethics. It may even lie at the core of recent corporate mission statements, such as the famous motto, “don’t be evil”, that Google used to have as the opening statement of its internal code of conduct. This was later rephrased to “you can make money without being evil”, but in both cases captures the aspiration that companies should strive to not hurt society or the users of their products, even as they try to make profit.

This dual allegiance is a theme we will return to in this course and a core objective of professional codes of conduct. What we are seeking is something like a Hippocratic oath for data scientists.

1.1.4.2 Hippocratic Oath for Data Scientists

Doctors start thinking about medical ethics very early in their training and literally stand in front of a crowd taking an oath not to hurt their patients to protect their privacy, amongst other things.

A profession with as broad an impact as data science ought to have a similar process, as Cathy O’Neil argued in her 2016 book *Weapons of Math Destruction*.

I haven’t yet quite decided whether we should all take an oath together when the course is completed, but that is definitely the idea.

So I’ve invited you to think about using AI for good and ensuring it is safe. But how easy is that, really?

1.1.5 Doing the right thing is neither obvious nor is it easy

It turns out that doing the right thing is neither obvious nor easy.

We will often lack context or lack an understanding of the group which is most at risk. Unless advocates representing that group are consulted, our ignorance might be dangerous.Moreover, even when we have the necessary facts, humans are often hampered by cognitive biases and bad habits. Processes need to be in place to overcome such sources of error.

Even in cases where we know what the right thing is, it might be difficult to do it unless the right incentive structures are in place. If it is not safe to raise a certain concern in the company or if it is likely to negatively impact your career progression, then fewer people will have the courage and determination to do it.

We should build organisations in which you don't have to be a hero to do the right thing. In certain situations, even when both the knowledge and the will to do the right thing are present, we will be faced with moral dilemmas where one harm must be balanced against another. Given all this, it is easy to give up but just because the problem is too big to fix immediately, that does not mean you can't make progress.

Finally, despite our best efforts, our technology can sometimes have unanticipated consequences. A humble attitude where we learn from our mistakes is needed to at least ensure that no accident happens twice.

Having said all that, we should clarify that we won't attempt to define what is right in this course. That's too hard, and this is not a course in moral philosophy. After all, your instructors are both mathematicians, not social scientists or philosophers. We will mostly take certain values as a given, as you will see later in this week.

You can perhaps think of this course as an attempt to change your default perspective on your work. Where you might be used to focusing on success stories and the likely positive impact of your work, you will now start to think about near-misses: things that could have gone wrong and anticipate harm they might have caused.

Where before you might focus on how to access more data, you will now start to think about whether you have permission to use the data in a certain way.

Where before, you might have been committed to do good and not evil, but this making it emotionally difficult to admit that some of your work will necessarily entail risk of harm. Now, you might start to humbly accept that risk and commit to doing the best you can to mitigate against it.

Where before, you might not even have thought it was part of your job as a data scientist to worry about such things. Going forward, as a specialist in the field, you will see it as your responsibility to raise the bar.

Where you would previously see unsolvable moral dilemmas, now you will seek to explicitly quantify these trade-offs so as to have an informed conversation about the right balance.

Where you might have hoped that eventually technology will solve its own problems, and we should just focus on doing more research; now, you will approach the problem more pragmatically, acknowledging that not all problems can be solved with technology and in any case we cannot afford to wait.

Whereas before you'd be obsessed with your model's performance, you will now start to monitor other things that we care about, such as the degree to which it protects privacy or treats people fairly.

And finally, where before you might have preferred to only discuss these things with your fellow engineers and scientists, you will now feel confident to engage in conversation with broader society.

1.1.6 Conclusion

To sum up, we have established that with power comes responsibility and that you, as a trained professional, are best placed to be a driver for positive change.

We have argued at the core of safe and benevolent AI is the ability to anticipate harm, to minimise it (even if we can't eliminate it altogether) and to communicate it transparently to the public.

We have also showed how it is important to think about the humans who are impacted by a technology, putting the human in the centre of the design and not just talking about the technology itself.

In what follows, we will pause to better define what data scientists actually build, so that we can be specific in tying potential harm to different components of a data science pipeline.

We will then discuss the progress that has been made to date by way of codifying the ethics of data science and AI, and conclude the week by offering our own codification of this problem area into five major principles.

These five principles will also serve as the backbone of the rest of the course. Welcome aboard!

1.2 Data Science Pipelines

In the previous section, we covered in detail why we should consider the ethical dimensions of our work as data scientists. To do this we relied on analogies with other professions such as doctors and engineers.

In this section, we will take a break from ethics to get on the same page as to what exactly it is that data scientists and AI engineers build. Is it a model? Is it a product? Or is it a business process? We'll try to come up with a description that fits the most common use cases out there. If you're a practitioner already, then you will recognize much of what follows. You might you refer to some of it with different names so another reason for this discussion to align on the terminology that we will use within this course.

When you ask the question what do data scientists and AI engineers build, you are likely to get one of two popular answers. Many technically minded faults will say they build models, be that a regression model or a neural network. According to these people, the output of a data scientist's job is a trained machine learning model that can then make predictions in the real world. At the other extreme, people without technical training might say they build automated (or partially automated) decision making systems. These people think about a finished product and how it operates in the world.

We will argue in this course that both of these views are incomplete. We take the view that data scientists and AI engineers do not build algorithms but rather they construct business processes that are used to build algorithms, which together we will call pipelines.

This broader view is more suitable for our purposes because it alllows for a more holistic appreciation of the stages where things can go wrong. If you take one thing away from this course is that your work output as a data scientist is a pipeline that generates a model not just the model itself. A sloppy pipeline without documentation and bits and pieces of manual work that no one recalls or can necessarily reproduce is a problematic work output, even if the resulting model has incredible accuracy. This mindset shift is urgently needed not just from data scientists but also from their managers and the public. Exclusive focus on the model and its accuracy invites complacency in the manufacturing process. Ask any chef or factory floor manager - a clean and organized working space underlies all quality work

Let's take a few steps back. What is a data science pipeline anyway? Let's start with the most classical use case for machine learning, that of *supervised learning*. This comes in two main 'flavours', which you may have met already.

1.2.1 Classification

[TODO: INSERT CAT PICTURE]

Here's a cat. Or is it a dog? Your brain is pretty sure it's a cat even though you don't quite know why. All you know is that when you see a cat some area within your brain fires up and returns the answer cat. This is known as *classification* in machine learning and it is the problem of attaching a label to an object based on a description or some *features* of that object. This is an example of *binary classification* because we only have two labels: cat and dog.

Of course, your brain can do much more than that. Had the picture been of a giraffe, you wouldn't have said "well this is neither cat nor dog... but it's closest to a cat so I'll have to go with that". You (hopefully) would have immediately reported that it is a giraffe and that it is ridiculous to call it a cat or a dog.

This is because your brain acts like a multi-class animal classifier, with hundreds of animals to choose from, not just two. Even such multi-class classifiers have their limitations - for example you may never have seen the first animal before and it is not clear whether the second is a chihuahua or a muffin. This sort of "out of sample" or misapplication error is a common mode of failure for classifiers that we will explore in greater detail later in the course.

[TODO: Add Bluefooted booby and chihuahua/muffin images]

Returning to binary classifiers, how might we construct one? We could try to write a function that describes a set of rules about how a cat picture looks different to a dog picture. In fact, early image classification techniques, known as expert systems, did just that. A long list of rules would be developed trying to exhaustively list the differences between dogs and cats. That's tedious work and it is also not how humans learn to tell dogs apart from cats.

If you have a child, you know that it takes a lot of pointing and trial and error for a child to learn how each animal looks. A lot of early childhood play is in fact about that about training children to classify objects: shapes, colours, animals and so on. WHen describing these everyday objects, we might sometimes revert to explicit rules, for example we might say to a toddler that dogs are generally bigger than cats. However, this is usually to fine tune a system that has been put in place by looking at a large set of *labelled examples* and asking children to extrapolate.

[TODO: Insert Model schematic]

Supervised learning follows the same logic. We we have a model with two supporting functions: a training function allows it to take as input a set of examples x and their associated labels y and use these update its internal state according to some optimization criterion. The model also has a predict function, which can provide a label (that may or may not be correct) to any example it is given. We can see from this that a predictive model cannot really be thought of as separate to its training data and training algorithm. They are in effect one and the same object.

Now let us look at that a little bit more abstractly. An algorithm is any pre-specified list of instructions that takes an input and produces an output. For example addition takes as input two numbers and gives back their sum.

A machine learning model is more complex, it has an algorithm that takes as input a training data set of labelled examples and outputs what is in reality another algorithm but in fact in software it is usually represented by a parameter θ. This parameter might be a single number, a vector of numbers, or something quite a bit more complex. For example in a linear or logistic regression θ would be a vector of regression coefficients. The fitted parameters θ can then be used as an input by the predictive model to produces a predicted label for a new example. We can think of the predictive model as just another algorithm with two inputs: the fitted parameters θ and the example we would like to classify.

Let's revisit exactly what we mean by a *label* here. In the example above we had two possibilities "cat" and "dog", that's binary classification. A classification tool could accept more than two labels, but there needs to be a finite set of them determined in advance. There are many real world variations of this problem with labels that are nested or hierarchical. We might also convert a numeric variable into a label for convenience, for example we might predict whether the value of a stock went up or down, if we are interested in only the direction and not the size of that change.

We will often want to predict the probability of each possible label, rather than returning the most likely label. This is known as probabilistic, soft or fuzzy classification, depending on who you ask. With a little bit of work, most modern classifiers are able to produce a score which can be interpreted as a probability, though the underlying algorithm might not be actually designed to do so. This can sometimes result in quite old-looking probabilities an issue that we will revisit it when we discuss explainability in detail.

1.2.2 Regression

In other settings, you might want to predict or describe a numeric variable, for example the yield of a crop or the price of a house. Usually this is referred to as a *regression* problem, in contrast to classification problem that we met earlier. Many algorithms that perform classification can also perform regression and vice-versa, requiring only minor modifications.

1.2.3 Unsupervised Learning

Finally, there are also types of machine learning that are unsupervised, where there is no obvious *target variable* whose value we want to predict using a number of other features. In these cases, we might still want to group examples together that look similar or spot ones that look out of the out of the ordinary. You have a full course dedicated to each supervised and unsupervised learning - we will see some examples of each in this module.

1.2.4 Modelling as part of a Pipeline

[TODO: Add pipeline image]

We know that the trained model is just one part of what the data scientist is responsible for. A more complete view is offered in Figure [TODO: REF]. A standard data science pipeline starts with some data being collected, either as part of an existing process or purposefully, say with a survey or an experiment. This data is then pre-processed and a problem is formulated which needs to be addressed using this data set including a choice of target variable. Ideally this would be done before collecting any data in the first place, but the world is a messy place and we do what we can with what is available to us.

Following this, features are then created (or *engineered*) out of the raw data and several models are trained on the resulting table of features and targets. A performance measure is then used to evaluate the models and to identify the best performing among those considered. If the performance of this best model is considered to be good enough then it is deployed - this could be a one-off use to produce some business insights, or increasingly a continuous use where the model is made available as a service to either other pieces of technology or to human users.

While in production the model is monitored, or at least should be monitored, so that the data scientists can stay abreast of any changes in model performance or other issues that might arise.

This breakdown of the data science workflow is also to guide our exploration of the ethical risks at each step in the pipeline.

1.2.4.1 Data Collection

For example, at the data collection stage we should be aware of any historical biases that are represented in the data. These biases are likely to bias our model's predictions if not addressed. For example, using historical data to see which employees were promoted rapidly in an organization to influence future hires will likely reproduce any biases present in these promotion practices. Historical data can also be biased in terms of what they do not record. For example it is well known that for certain demographic groups in the United States, access to health care is difficult. This means that the rates of diagnosis of certain diseases in these groups will either be smaller than the real prevalence of the disease or arrive much later in the patient history. This type of selection mechanism can poison all sorts of downstream analysis.

1.2.4.2 Data Pre-processing

After the data have been collected, it typically needs extensive processing to be brought into shape for analysis. This is known as *data pre-processing* or more recently it falls under the more general moniker of *data engineering*. The stage of the data science process is often dismissed as tedious but mathematically uninteresting work and yet it is in that stage that bias can easily creep in. Consider filling in missing values - this is usually done using off-the-shelf methods without much thought. This usually results in inappropriate techniques being used, like filling in a value with the average of that variable across the other respondents. This would result in overestimating the weight of women in a patient population, as women tend to weigh less than men. Even more challenging situations can arise when a value being missing or not is itself dependent on the value itself. For example, in a survey trying to estimate the prevalence of drug use the people who use drugs may less likely to respond than those who do not.

1.2.4.3 Formulating Your Research Question

We now come to the problem formulation stage, which is perhaps the most important of all. Einstein famously said “If I had an hour to solve a problem I would spend 55 minutes thinking about the problem and 5 minutes thinking about solutions”. Having a clear, correct and answerable question is important for two reasons. Firstly, asking the wrong question can only give you a wrong answer and secondly this step is very hard to revisit later - no model performance metric will ever alert you to having asked the wrong question in the first place. Just like Douglas Adams had said in the Hitchhiker's Guide to the Galaxy, it is much harder to build a computer that can tell you what the right question to ask is than it is to build a computer that answers a question for you.

The skill of defining well-formed research questions is particularly important and problematic in data science because our tools expect our problems to take a specific shape, say that of a classification or a regression. As a result it is tempting to immediately look for ways to convert a real world problem into a prediction task, and to do so in a hurry because your "real work" starts once this has been done. Another temptation is to assume that it is the job of a business person or a domain expert to do that translation and it is their fault if you can't answer their vaguely defined question. In fact forming good research questions can only be done well in a multi-disciplinary fashion. This is because it requires a deep understanding of the domain of application, the business problem, and the modelling approach.

Feature engineering carries similar challenges as problem formulation many domain-specific assumptions can come into all of the little decisions we need to make when we create a feature. For an unstructured or semi-structured data set, feature engineering can be a very manual and time-consuming process but it also gives us a chance to understand the problem domain. On the other hand, modern techniques like end-to-end deep learning give us a chance to completely automate that process using very sophisticated computational approaches to trial and error. This improves the reproducibility and speed of the pipeline, but it does increase the reliance on bias-free data as there's no opportunity to correct or self-reflect in this pipeline

1.2.4.4 Modelling and Inference

Once the set of predictive features and choice of target variable are all in place, we then usually rely on standard machine learning libraries like scikit-learn in Python to go through a few types of models, say a random forest classifier, an xgboost classifier and a logistic regression. We fit them to the data, assess their performance and we select the best performing one. Both the optimization algorithm used to fit the classifier and our subsequent selection of the best performing one rely on the choice of *performance metric.* This metric is very important, defining what the relative severity of different types of errors. For example, you would likely regard misclassifying a patient with cancer as healthy as a more severe error than the other way around. However, a false alarm also has real costs - both for obvious emotional reasons and also because it will require follow-up tests to confirm the diagnosis, which might not have been necessary had the initial classification been correct. Choosing the right performance metric is half the battle during model selection but here a huge number of checks and balances need to be kept to avoid overfitting and so on.

1.2.4.5 Deployment

You have now completed your task, the model has been fitted and has say 93% out of sample accuracy. Your job is done, right? Wrong.

For starers, a model has to be documented with the appropriate guard rails to ensure that it is used in the right way. Think back to the example of using our cat-dog classifier to a giraffe. A more serious example might be a smartphone app that predicts the risk of heart failure in the next four weeks given a number of tests or readings. You need to make sure that the app is indeed deployed on the right data and interpreted by users in the right way.

You must also think about access issues. If you have just built a great solution for a real world problem, what can you do to help make it available to more users? You want your heart failure app to be used by everyone just not just folks with expensive phones. You might think that this is not your job, but once again you can add more value in this conversation than you may think. Are there ways to make the algorithm computationally more efficient, so that it can run on older generation phones? Perhaps that drop in accuracy by switching to a faster algorithm is worth it, if in return the solution is available to many more people and more lives are saved in total.

1.2.4.6 Monitoring

No one steps in the same river twice said Heraclitus, or as it is sometimes abbreviated “everything changes”. Once your model is deployed, you want to make sure you keep tracking its performance - for example its speed, as well as a number of other dimensions that we will explore in this course. You wouls like to detect any changes that occur, for example if you tested your algorithm in a group of older patients then you might want to make sure that the same performance applies to your population of actual users, who might be younger. Even if the populations were the same when your solution was put into production, sometimes the world itself changes. The COVID-19 pandemic has really challenged many machine learning models that were built on data collected prior to its beginning. Patterns of human behaviour and interaction have changed and so many models had to be retrained. This effect is known as data set shift. One solution is to monitor the model and trigger reviews whenever there is an indication that something is going wrong.

It is unfortunately easy to find real world examples of all of the above risks going wrong. A recent landmark publication in the highly esteemed journal Science, the authors demonstrated that an algorithm used to manage population health issues used health cost as a proxy for health need. That is, to understand how much care someone suffering with a certain disease requires they looked at how much money their insurer spends on them per year. This proxy is an ingenious idea, to get around a difficult data collection problem the study used anonymised insurance claim data which is often available for millions of patients. However, the study omitted the fact that there are groups of people that do not have access to high quality medical insurance. These people will therefore have less care for the same level of need and a model trained on this data variable will unfortunately perpetuate this historical bias.

1.2.5 Notes on the term “bias”

It is worth commenting on the use of the word **bias** in a course. It will be used quite heavily because it has both an ethical meaning, that is discrimination against certain individuals or groups, but it also has a number of different technical meanings in statistics and data science.

A useful graphic to get us to think about this in a more structured way was presented in a paper by Mitchell et al [TODO: add citation] and is presented also here.

[TODO: Add Mitchell graphic]

We all have a view of the world as it should be free of discrimination and prejudice, however that differs from the world as it is, and as it has been, due to societal bias and injustices that have taken place in the past. Even the perfect capture of all human activity right now would inherit that bias. Unfortunately the situation is even worse than that. We only observe the world imperfectly with samples that are small or even non-representative, which introduces an additional source of bias that is statistical in nature. This graphic [TODO: ref Fig] is a useful reminder to tease apart the two sources of bias as they each require a different mitigation strategy

1.2.6 Conclusion

In this section, we have defined what it is we mean by data science work. We have highlighted how it is important to pay as much emphasis on the pipeline that generates a predictive model as to the predictive model itself. We broke a typical data science pipeline down to multiple stages and explained how each stage poses its own ethical risks, showcasing some examples along the way.

From here we will return to our discussion of foundations in order to understand what progress has already been made in establishing best practices and regulatory frameworks.

1.3 Moral Frameworks

In Section 1.1, we explained why we need to consider ethical matters as data science professionals. We next reviewed what a typical data science pipeline looks like and gave some examples of the risks that each stage in such a pipeline can pose.

We will now return to our discussion of moral frameworks. The objective today is to demystify the term “ethical”, at least in terms of how it’s used in this course.

1.3.1 How do we determine ethical behaviour?

Let's first start by reviewing how is it that we shape our moral opinions and attitudes as individuals. Morality sometimes is the result of a thoughtful rational exercise, but in many situations it is an almost instinctive response. It takes some effort to recognize precisely the origins of our moral attitudes because we don't always have the ability to directly introspect them.

Clearly, the society we live in influences what we think is right and wrong. That being said, societies can themselves often be pluralistic, divided or both. For every division of opinion in a large social group you can probably find a fair amount of shared moral attitudes as well, many of which people have often simply stopped noticing or paying attention to since they are not contented opinions.

In addition to norms and habits, there are philosophical viewpoints that can influence moral attitudes on big topics such as crime and punishment, tradition versus progress, and many others. Our community's special history and its relationship with its surrounding society is also an important source of moral views, along with our self-identity and broader cultural influences. Another clear source of moral attitudes is of course religion.

Finally, our professional lives can also play a big role in our moral attitudes. This can be especially true in the case of mission-driven professions like doctors, nurses and teachers or in the case of very strong corporate identities like the ones surrounding lifelong careers in a single corporation or more recently certain technology companies.

What makes this plurality of sources interesting is that, often, they can actually come into conflict. Something you're asked to do at work might for example be contrary to your personal morality or religious beliefs. It might therefore seem like we have a really hard problem in our hands: to be an ethical data scientist do we first need to all agree on the same definition of what it means to be an ethical person? Surely that would take us a few years of study at the very least. As we will see shortly, that's not exactly what is required.

1.3.2 Medical Codes of Conduct

Let us consider the example of doctors; they are a professional group that abides by a strict code of conduct, relying on four relatively simple principles:

- non-maleficence,
- beneficence,
- equity,
- autonomy

Non-maleficence, the idea that a doctor should "do no harm" we have already encountered. **Beneficence** is the other side of the same coin, a commitment to only intervene so as to optimize the health and welfare of the patient. The principle of **equity** reminds us that all patients should have access to equal care, regardless of race ,gender on any other attribute. Finally patient **autonomy**, which includes **privacy** and medical **confidentiality** asserts that the patient is the one that decides whether they want to receive treatment and the doctor must support them in this decision, even if they disagree.

There's a couple of things to observe here. First, just four fairly generic and simple to state principles are obviously not enough to tell a doctor what the right thing to do is in any given specific circumstance. This is done on a case-by-case basis through ethics committees and, in particularly thorny cases, debated in bioethics journals. However, the principles offer a robust framework against which any new situation must be analysed. Second, none of these principles mention social or religious considerations. They allow for some flexibility in defining what is good, but not an infinite amount wiggle-room. Looking specifically at the autonomy principle, this would precluded an authoritarian state from running experiments on their citizens without their consent. This emphasises that professional codes of conduct do not need to be a complete philosophical treaties, written from first principles on the nature of good and evil. Professional codes of conduct are instead a contract of trust between the professional and the society they live within.

If a doctor is responsible to do good but do good to whom and by whom the obvious answer is the patient but is that all there is to it the american medical association's code of medical ethics answers this question for us in the highlighted passage above. Physicians it says, must recognize responsibility to patients first and foremost, as well as to society, to other health professionals, and to themselves. There is a hierarchy of stakeholders with a patient coming first, society at large second, other doctors third and self last.

To consider a concrete example, a medical doctor is obligated to offer antibiotics to a patient who needs them. The doctor also needs to also keep in mind that over-prescription of antibiotics will, over time, lead to resistance to those drugs, which harms broader public health. There will be borderline cases where, absent that consideration about resistance and public health, the doctor might have prescribed antibiotics to a patient but didn't. These moral dilemmas can be tricky to solve but the important first step is to recognize them by thinking expansively about who might be affected by your work. It is rarely just the user of your product or service; most commonly it is their community, their family, and society at large that are indirectly affected by use of an AI-based service.

1.3.3 Codes of Conduct in Other Professions

We focused earlier on on the Hippocratic oath but broader professional codes of conduct that can be enforced (and that can incur some kind of penalty when not respected) are a fairly recent development. A pivotal moment in their development were the Nuremberg trials, which prosecuted the cruel, unthinkable human experiments run by the Nazi regime that unfortunately have also occurred under other authoritarian regimes. This trial triggered a need to formulate an explicit code of medical conduct that doctors can swear to and feel loyal to. This can then act as a counterbalance of their loyalty to their employer or government. Although the Nuremberg trials is an absolutely extreme example, the principle that professionals owe allegiance or loyalty to their profession over and above to their loyalty to their employer, government, social group, or religion is absolutely foundational.

Codes of conduct therefore act as a fail-safe. A chartered accountant is obliged to be the first one to report a fraud, even if that means reporting their own employer. A medical doctor must respect patient autonomy, even when in desperate circumstances from the patient's perspective. A chartered statistician cannot misrepresent data, even if the CEO of the start-up they work for asks them to do so. Codes of conduct are nowhere near as enforceable as law but they're broader and they're more agile, which is useful as a means for rapidly evolving technologies to self-regulate.

Here is another example: the Royal Statistical Society's code of conduct specifically asks Fellows of the Society to seek to counter false or misleading statements which are detrimental to statistical science,the profession, or society. In the age of fake news and misinformation, does that oblige a fellow of the RSS to respond to say a Twitter thread that misrepresents data about the pandemic? For example what if your employer has a strict policy against doing something like that; against entering public debates on social media? We do not need to answer these questions now, we just need to recognize that it is important that they get asked.

The picture we are landing on looks as follows. Different professions have their codes of conduct: statisticians, doctors, accountants, and so on. As these folks do their daily work they also have loyalties to their companies, for example a technology or pharmaceutical company, each of which has its own mission. The result is a creative and healthy tension between corporate objectives, including but not limited to profit making ones, and professional ethics. This tension helps the companies to organize themselves in healthy ethical practices. At the same time, professionals have their own personal identity, are influenced by societal norms and by a set of fairly universal moral values such as the right to privacy or to non-discrimination. Within a specific individual exists a similar amount of creative healthy tension with the code of conduct again acting as a fail-safe against the specific beliefs of a given professional, that may be in conflict occasionally with the welfare of the user.

The general medical council in the UK specifically describes what to do if you're a physician who is asked to conduct a procedure that goes against you personal beliefs, morality or religion. The physician in such circumstances must explain to patients if they have a conscientious objection to a particular procedure, they must tell the patient about their right to see another doctor, and make sure they have enough information to exercise that right. In providing this information they may not imply or express disapproval of the patient's lifestyle choices or beliefs. This gives space to the physician to hold individual beliefs while at the same time making sure that a patient is not harmed or misinformed as a result. This again emphasizes the need to seek codes of conduct as contracts of trust that can co-exist with other sources of moral attitudes, rather than as absolutist statements of truth.

1.3.4 Conclusion

In Section 1.1 we understood what this course will be about, in this section we have instead covered what it will not about. This course, and ethical AI as a whole, not an attempt to exhaustively establish what is right. This is a question touching upon philosophy, religion, politics, social sciences and history; none of which will be taught in this course.

We clarified that codes of conduct are obligations that arise if we are to expect the trust of our fellow citizens. We also emphasized, once again, that many modern codes of conduct and pieces of regulation came about after humanitarian disasters. Our ambition and hope is that AI and data science will self-regulate in an effective manner before that happens. Finally, we used examples from the medical profession to illustrate how professional codes of conduct can act as fail-safes against individual, corporate or government misconduct.

Throughout this course, we have come to the question of principles again and again. A simple set of concepts that create a scaffold against which we can analyse the ethical implications of our work. Finally we have reached the point where we can list our own set of principles and that is what we will do in the following section.

Chapter 2

Privacy and Autonomy

Cross-references make it easier for your readers to find and link to elements in your book.

2.1 Captioned figures and tables

Figures and tables *with captions* can also be cross-referenced from elsewhere in your book using `\@ref(fig:chunk-label)` and `\@ref(tab:chunk-label)`, respectively.

See Figure 2.1.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Don't miss Table 2.1.

```
knitr::kable(
  head(pressure, 10), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

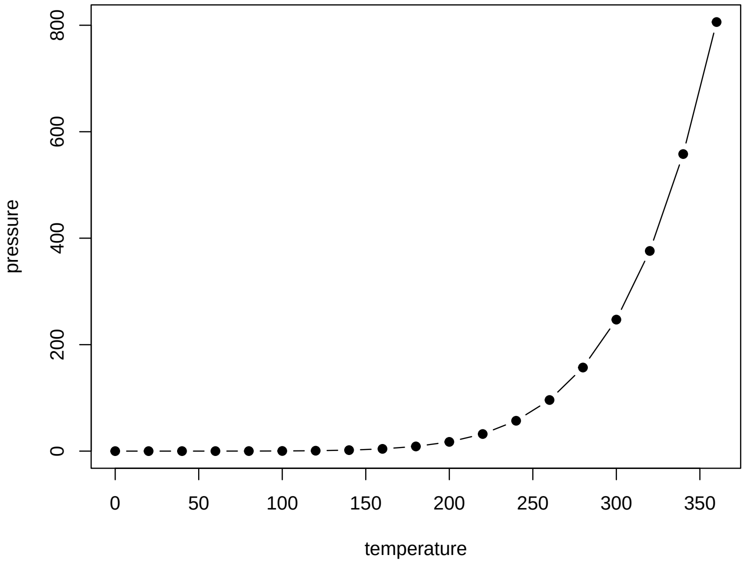

Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

temperature	pressure
0	0.0002
20	0.0012
40	0.0060
60	0.0300
80	0.0900
100	0.2700
120	0.7500
140	1.8500
160	4.2000
180	8.8000

Chapter 3

Fairness

You can add parts to organize one or more book chapters together. Parts can be inserted at the top of an .Rmd file, before the first-level chapter heading in that same file.

Add a numbered part: `# (PART) Act one {-}` (followed by `# A chapter`)

Add an unnumbered part: `# (PART\*) Act one {-}` (followed by `# A chapter`)

Add an appendix as a special kind of un-numbered part: `# (APPENDIX) Other stuff {-}` (followed by `# A chapter`). Chapters in an appendix are prepended with letters instead of numbers.

Chapter 4

Alignment and Control

4.1 Footnotes

Footnotes are put inside the square brackets after a caret `^[]`. Like this one [1].

4.2 Citations

Reference items in your bibliography file(s) using `@key`.

For example, we are using the **bookdown** package (Xie, 2022) (check out the last code chunk in index.Rmd to see how this citation key was added) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015) (this citation was added manually in an external file book.bib). Note that the `.bib` files need to be listed in the index.Rmd with the YAML `bibliography` key.

The `bs4_book` theme makes footnotes appear inline when you click on them. In this example book, we added `csl: chicago-fullnote-bibliography.csl` to the `index.Rmd` YAML, and include the `.csl` file. To download a new style, we recommend: https://www.zotero.org/styles/

The RStudio Visual Markdown Editor can also make it easier to insert citations: https://rstudio.github.io/visual-markdown-editing/#/citations

[1] This is a footnote.

Chapter 5

Explainability and Interpretability

5.1 Equations

Here is an equation.

$$f\left(k\right) = \binom{n}{k} p^k\left(1-p\right)^{n-k} \tag{5.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (5.1).

5.2 Theorems and proofs

Theorem 5.1. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 5.1.

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

5.3 Callout blocks

The `bs4_book` theme also includes special callout blocks, like this `.rmdnote`.

You can use **markdown** inside a block.

```
head(beaver1, n = 5)
#>   day time  temp activ
#> 1 346  840 36.33     0
#> 2 346  850 36.34     0
#> 3 346  900 36.35     0
#> 4 346  910 36.42     0
#> 5 346  920 36.55     0
```

It is up to the user to define the appearance of these blocks for LaTeX output.

You may also use: `.rmdcaution`, `.rmdimportant`, `.rmdtip`, or `.rmdwarning` as the block name.

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Chapter 6

Safety, Security and Accountability

6.1 Publishing

HTML books can be published online, see: https://bookdown.org/yihui/bookdown/publishing.html

6.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

6.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `bs4_book` provides enhanced metadata for social sharing, so that each chapter shared will have a unique description, auto-generated based on the content.

Specify your book's source repository on GitHub as the `repo` in the `_output.yml` file, which allows users to view each chapter's source file or suggest an edit. Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/bs4_book.html

Or use:

```
?bookdown::bs4_book
```

Build Information

This book was written in bookdown inside RStudio. The website ethics-1.zakvarty.com is hosted with Netlify. The complete source is available from GitHub.

The course logo was designed by Zak Varty.

This version of the book was built with:

```
#>  setting  value
#>  version  R version 4.2.0 (2022-04-22)
#>  os       macOS Big Sur/Monterey 10.16
#>  system   x86_64, darwin17.0
#>  ui       X11
#>  language (EN)
#>  collate  en_GB.UTF-8
#>  ctype    en_GB.UTF-8
#>  tz       Europe/London
#>  date     2022-10-05
#>  pandoc   2.18 @ /Applications/RStudio.app/Contents/MacOS/quarto/bin/tools/ (via rmarkdown)
```

Along with these packages:

Package	Version	Date	Source
bookdown	0.26	2022-04-15	CRAN (R 4.2.0)
brio	1.1.3	2021-11-30	CRAN (R 4.2.0)
cachem	1.0.6	2021-08-19	CRAN (R 4.2.0)
callr	3.7.0	2021-04-20	CRAN (R 4.2.0)
cli	3.3.0	2022-04-25	CRAN (R 4.2.0)
crayon	1.5.1	2022-03-26	CRAN (R 4.2.0)
desc	1.4.1	2022-03-06	CRAN (R 4.2.0)
devtools	2.4.3	2021-11-30	CRAN (R 4.2.0)
digest	0.6.29	2021-12-01	CRAN (R 4.2.0)
ellipsis	0.3.2	2021-04-29	CRAN (R 4.2.0)
evaluate	0.15	2022-02-18	CRAN (R 4.2.0)
fastmap	1.1.0	2021-01-25	CRAN (R 4.2.0)
fs	1.5.2	2021-12-08	CRAN (R 4.2.0)
glue	1.6.2	2022-02-24	CRAN (R 4.2.0)
htmltools	0.5.2	2021-08-25	CRAN (R 4.2.0)
knitr	1.39	2022-04-26	CRAN (R 4.2.0)
lifecycle	1.0.1	2021-09-24	CRAN (R 4.2.0)
magrittr	2.0.3	2022-03-30	CRAN (R 4.2.0)
memoise	2.0.1	2021-11-26	CRAN (R 4.2.0)
pkgbuild	1.3.1	2021-12-20	CRAN (R 4.2.0)
pkgload	1.2.4	2021-11-30	CRAN (R 4.2.0)
prettyunits	1.1.1	2020-01-24	CRAN (R 4.2.0)
processx	3.5.3	2022-03-25	CRAN (R 4.2.0)
ps	1.7.0	2022-04-23	CRAN (R 4.2.0)
purrr	0.3.4	2020-04-17	CRAN (R 4.2.0)
R6	2.5.1	2021-08-19	CRAN (R 4.2.0)
remotes	2.4.2	2021-11-30	CRAN (R 4.2.0)
rlang	1.0.5	2022-08-31	CRAN (R 4.2.0)
rmarkdown	2.14	2022-04-25	CRAN (R 4.2.0)
rprojroot	2.0.3	2022-04-02	CRAN (R 4.2.0)
rstudioapi	0.13	2020-11-12	CRAN (R 4.2.0)
sessioninfo	1.2.2	2021-12-06	CRAN (R 4.2.0)
stringi	1.7.8	2022-07-11	CRAN (R 4.2.0)
stringr	1.4.1	2022-08-20	CRAN (R 4.2.0)
testthat	3.1.4	2022-04-26	CRAN (R 4.2.0)
usethis	2.1.6	2022-05-25	CRAN (R 4.2.0)
withr	2.5.0	2022-03-03	CRAN (R 4.2.0)
xfun	0.31	2022-05-10	CRAN (R 4.2.0)
yaml	2.3.5	2022-02-21	CRAN (R 4.2.0)

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2022). *bookdown: Authoring Books and Technical Documents with R Markdown.* R package version 0.26.