The Alan Turing Institute

Responsible Research and Innovation

Ensuring that Data Science and AI Contribute to the Social Good

A guidebook produced to support a series of online training workshops

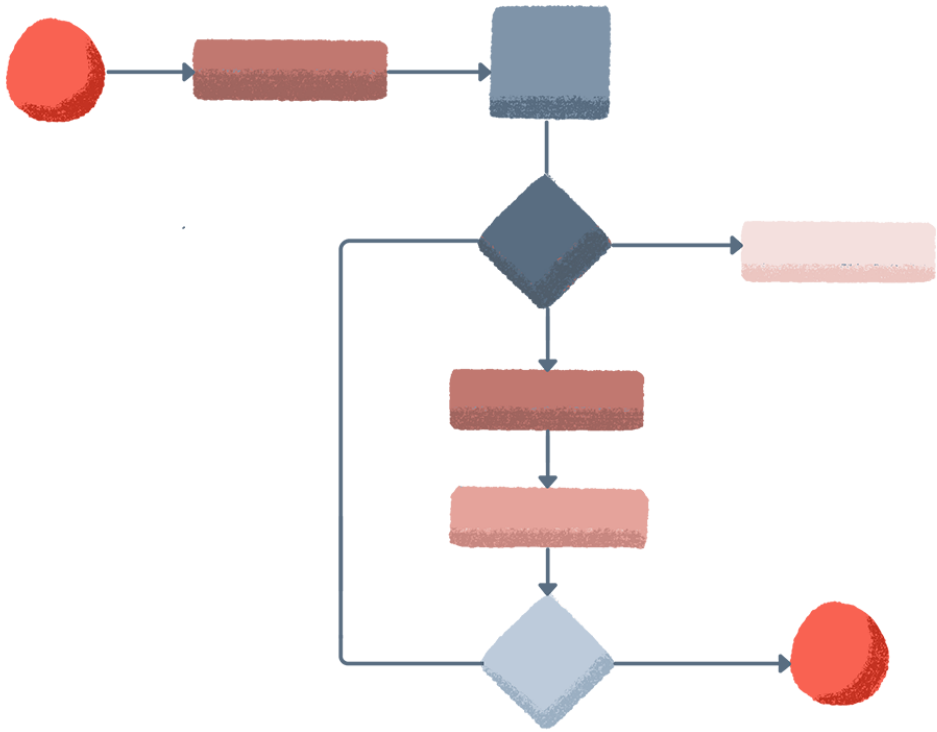
Table of contents

About this Guidebook

Responsible scientific research and technological innovation (RRI) is a vital component of a flourishing and fair society. As an area of study and mode of enquiry, RRI plays a central role within academic, public, private, and third-sector organisations. For example, the UKRI's Engineering and Physical Sciences Research Council (EPSRC) is increasingly making a commitment to RRI necessary for research funding, and also embedding RRI training into its Centres for Doctoral Training. Furthermore, the UK Government has highlighted the importance of RRI in both of its national data and national AI strategies.

Building on these commitments, this course will explore what it means to take (individual and collective) responsibility for (and over) the processes and outcomes of research and innovation in data science and AI.

The notion of 'responsibility' employed throughout this course will be grounded in an understanding of the moral relationship between science, technology, and society, exploring both historical and contemporary examples of RRI practices. As well as looking at the theoretical basis of RRI this course will also take a hands-on approach by exploring a variety of tools and procedures that can help operationalise and implement a robust notion of responsibility within research and innovation practices.



Who is this Guidebook For?

Primarily, this guidebook is for researchers with an active interest in RRI. This doesn't mean you have to be a data scientist, or a researcher using R or Python to analyse data. You could be an ethicist, sociologist, or someone with an interest in law and policy. However, the guide is oriented towards research issues and related topics.

In addition, while this course has practical, and sometimes hands-on activities, these activities are geared towards ethical reflection and deliberation. If you want to dive deeper into the specific day-to-day requirements of RRI for data science, we recommend heading on over to The Turing Way community, organised by our fantastic colleagues.

Learning Objectives

This guidebook has the following learning objectives:

Understand what is meant by the term 'responsible research and innovation', including the motivation and historical context for its increasing relevance.

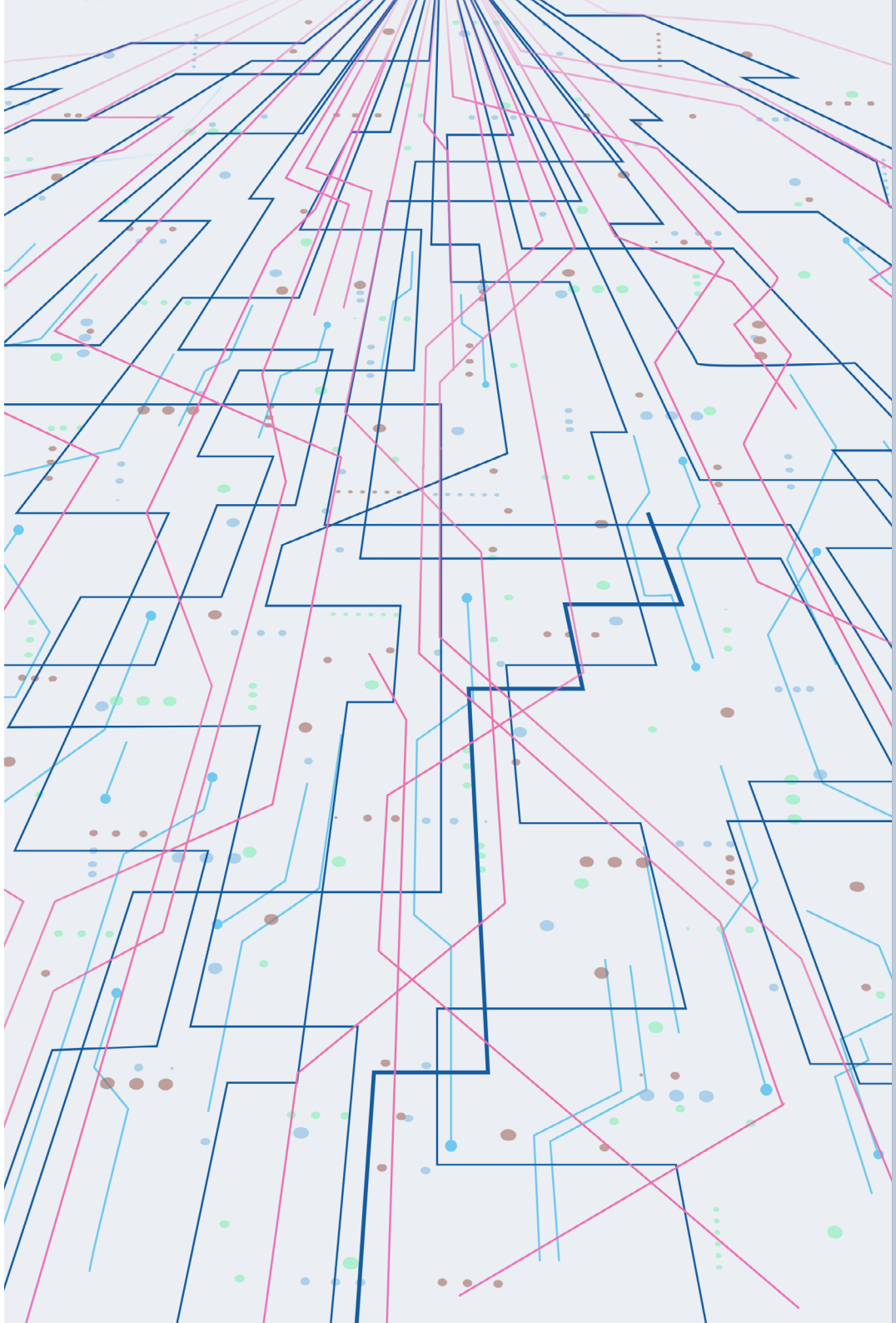
Identify and evaluate the ethical issues associated with the key stages of a typical data science or AI project lifecycle: (project) design, (model) development, (system) deployment.

Explore practical tools and mechanisms for operationalising the concept of 'responsibility' within the context of data science and AI research and innovation.

Gain an appreciation of shared goals and values across scientific disciplines and research domains through dialogue with other participants.

01

Introduction



Summary

> **This introduction serves to motivate the significance of some of the topics that will be explored throughout this guidebook. Although some topics and concepts are introduced, they are not explored fully in this introduction.**

Anticipating the consequences of scientific research and technological innovation can be an intricate and formidable task—one that is made all the more challenging when the target domain is a complex system such as the human brain or society.

In the early 1970s, a group of scientists working at Eli Lilly—a pharmaceutical company—were engaged in research that would go on to have profound effects on both of these complex systems. Primarily, the group were exploring the role of serotonin in depression, and their research led to the discovery of a drug known as fluoxetine hydrochloride, which affects the reuptake of serotonin in the brain (Wong et al., 2005)[1]. But the introduction of the drug also had a profound effect on society as it rapidly grew in popularity under its more common name, Prozac.

Prozac is a technology in the broad sense of the term, but it is not the sort of technology that this course guide will typically focus on. However, in spite of this, the discovery and development of Prozac is an interesting case to begin this guide with.

[1] A full list of references for this guide can be found online at https://turing-commons.netlify.app/

Unanticipated and Unintended Consequences

Although the scientists who were involved in the discovery and development of fluoxetine hydrochloride were confident that it could be used in the treatment of depression, the company Eli Lilly initially tested the drug as a treatment for high blood pressure and as an anti-obesity drug. Part of the reason for this was due to market analysis forecasting a limited demand for another antidepressant drug to rival the class of medications known as 'tricyclic antidepressants' that were already in use. So, as one of the discovers of Prozac admits,

> **"it would be presumptive to claim that we anticipated the wide acceptance of Prozac by both physicians and patients. Neither did we foresee that its pale-green and light-yellow capsule would appear on the cover of Newsweek (26 March 1990), which described it as: "A breakthrough drug for depression."**

As we will see throughout this course, anticipating the effects of your research or innovation project is a core part of what it means to take responsibility. But the discovery of Prozac shows us that there are limits to anticipation, and that sometimes unintended consequences may arise. For instance, antidepressants such as Prozac have received a large amount of critical scrutiny that focuses on their over-prescription and overuse in society, which can often lead to alternative (non-pharmaceutical treatments) being overlooked; the patchy state of our empirical understanding regarding how they affect and alter an individual's brain; and the range of side effects, including nausea, headaches, difficulty sleeping, dizziness, fatigue, and even increased suicidal ideation—the very thing that the drugs are often prescribed to help alleviate (Spence and Reid, 2013). Should the developers of Prozac have anticipated these consequences? Does taking responsibility for

research and innovation require researchers and innovators to have near-omniscient levels of anticipatory capabilities, in order to ensure no unintended consequences arise? Obviously, the answer is no. Any sensible moral theory that is intended to support practical decision-making must make room for the cognitive limitations of human individuals and teams, and we cannot ignore the many positive impacts that counteract the potentially negative and unintended consequences.

Returning to our example, it should be recognised that Prozac and its kin—a class of drugs called selective serotonin reuptake inhibitors (SSRIs)—have had a profound and positive impact on the lives of many people who suffer with depressive mood disorders. These positive (and intended) consequences can't be dismissed in any evaluation of the social and individual benefits of SSRIs. Let's take a look at some of these trade-offs as they pertain to the topic of RRI.

Dealing with Uncertainty

Perhaps the most important is the ongoing empirical uncertainty surrounding how antidepressants operate. Given the complexity of the brain and the relationship between an individual's mind and their culture and society, there is, unsurprisingly, a vast amount that is not known about the mechanisms and causal processes by which SSRIs operate. The rationale behind a lot of their use is that there is a typical (or, normal) level of serotonin reuptake and that depressive mood disorders are characterised by a deviation from this level—a homeostatic conception often referred to by the label 'chemical imbalance.' As such, SSRIs are intended to correct for this imbalance by returning (and holding) the levels to a set point. But, there are also a wide variety of other interventions that work for individuals, including cognitive behavioural therapy, herbal or dietary supplements (e.g., St John's Wort), sports and leisure activities, and also better sleep—something we could probably all benefit from.

In the course of assessing, diagnosing, and treating depression it is common for patients to be asked to evaluate whether the impact of their depression on their day-to-day activities and relationships is worse than the potential side effects of medication. This is a value-laden decision made under uncertainty, which cannot be made by the psychiatrist on behalf of their patient. Nor could it be fully accounted for in the initial course of developing the drug—individuals respond in a remarkably diverse number of ways to treatment. Rather, it can only be rationally decided on by the user of the treatment, following a process of informed consent and understanding.

This brings us to another significant aspect of RRI, which we will discuss in depth during later sections—reflecting on how a research or innovation project will impact upon the lives, rights, and freedoms of users or stakeholders must be a participatory activity that is inclusive of the users or stakeholders themselves. We can assume that the development and use of SSRIs was (and is) guided by noble and beneficent intentions to help people and improve public health outcomes. But as the well-known saying goes, 'the road to hell is paved with good intentions.'

Good intentions are not sufficient for acting responsibly. Let's look at an example. Cathy—a data scientist with an interest in machine learning—may believe that she is doing a good act by developing an automated system that uses natural language processing to secretly monitor the tweets or comments of her friends on social media and then alerts her if it detects negative language that could be indicative of depression or suicidal ideation. If it detects a message, it notifies Cathy so she can reach out to her friends and try to offer some help or support. But is she acting responsibly in doing so? Your immediate reaction is probably one of discomfort, leading you to think the answer to this question should be 'no.' Let's explore some of the relevant factors to see why this is likely the case.

An Example

First of all, you may think that in spite of the fact that Cathy may be trying to help, if she were to explain to her friend that she had reached out on the basis of an automated monitoring tool, they would probably feel as though she were violating an important expectation of trust by operating this tool in secrecy. Arguably, this is an outcome that she should have been able to anticipate had she reflected on the consequences of her project. Although Cathy's friend's messages and comments were made in public, they would have a reasonable expectation that Cathy would not use their posts or data for purposes that could be perceived as some form of surveillance. Had Cathy reached out to her friends and included them in her initial idea formation, they would probably have been able to save her a lot of wasted effort and prevented the unnecessary harms to their friendship.

Second, let's assume that Cathy has developed this tool and hosted the code on a public GitHub repository. Her intention is to allow others to help and support their friends in the same way, and is operating under the belief that by making her work open, accessible, and transparent—several principles that are commonly found in ethical frameworks—she is contributing to the public good. But what if the tool becomes incredibly popular and widely used? A possible unintended consequence of this is that many users of social media begin to feel unsafe posting on platforms that they had, hitherto, used for the purpose of seeking help and advice from a supportive community? Unfortunately, Cathy did not properly reflect on how these principles would operate within the context of her project, and falsely assumed they were unconditional goods.

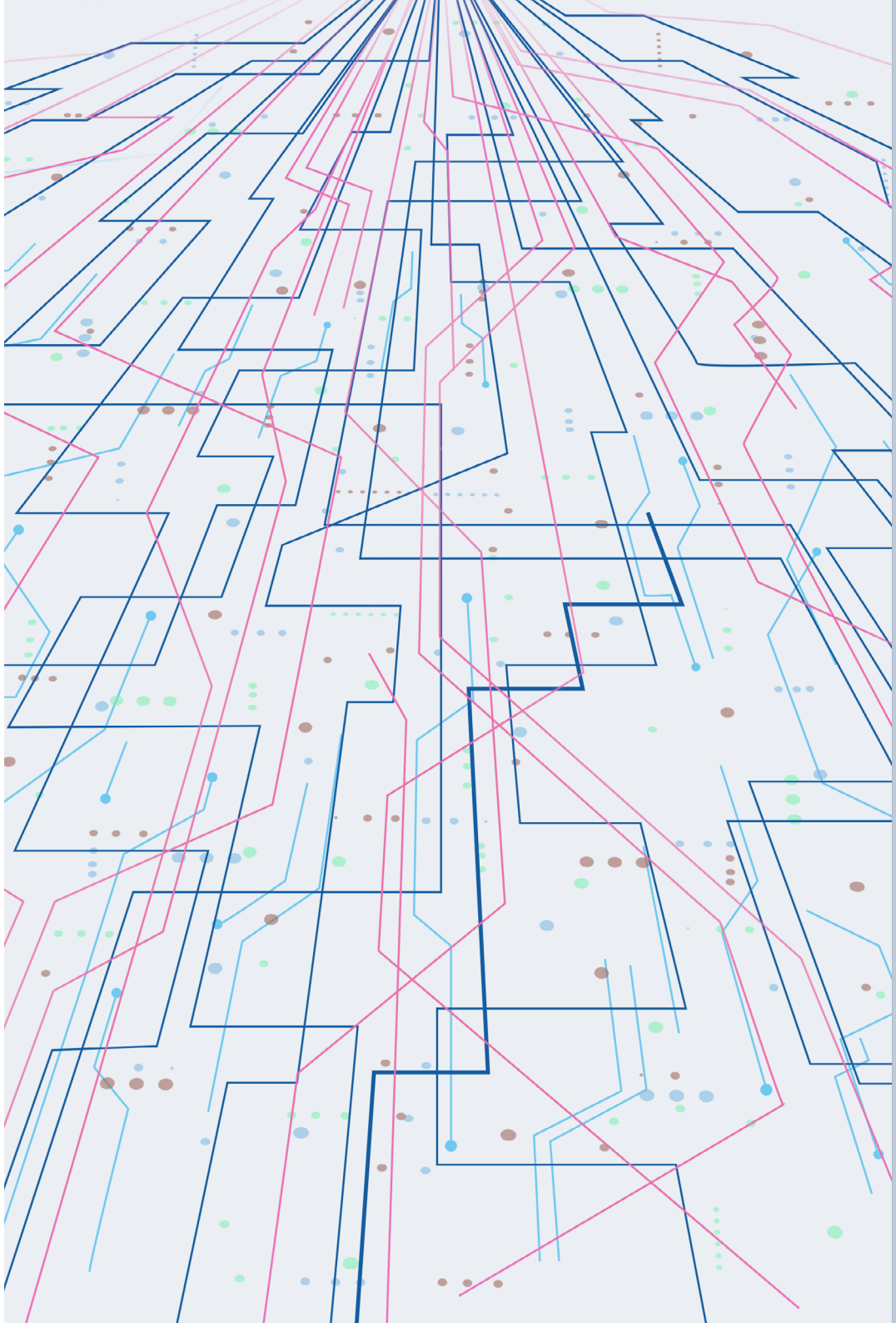
Finally, let's pretend that Cathy has genuinely overlooked both of these consequences and, therefore, has made no plans for rectifying the potential harm her tool has caused. By now, she is unable to respond to the emergence of the unintended consequences except by apologising and removing the tool, which could have already been copied to other user's computers.

It should be clear that this example is a case of irresponsible technological innovation that gives rise to several socially undesirable outcomes and harms. However, what is not necessarily clear at present is that Cathy's project can also be seen to contravene several principles of RRI: anticipation, reflexivity, inclusiveness, and responsiveness.

We will explore these principles in more detail in a later section. However, this introduction has already indirectly introduced you to most of the themes, topics, and principles that are covered in this guide. In the remainder of the guide we will approach each of them in a more systematic and structured manner, taking time to discuss and explore how they can help you conduct more responsible research and innovation in data science and AI.

02

What is Responsible Research and Innovation?



Summary

This section defines the term ‘responsible’, as it occurs in ‘responsible research and innovation’, and also explores a short history of RRI with reference to notable events and case studies that help illustrate central themes and issues. The chapter helps to set the foundation for the remainder of the course, which will rely on many of the core concepts that are introduced. However, the section also plays a motivating role by highlighting the interconnected nature of science, technology, and society.

Learning Objectives

In this chapter, you will:

- Learn what is meant by the term ‘responsibility.’
- Familiarise yourself with several principles that characterise RRI.

  Explore illustrative and historical case studies.
- Reflect on how the interactions between science, technology, and society give rise to many normative issues.
- Review pivotal moments in the history of science and technology studies (STS).

What is 'Responsibility'?

If we want a systematic and reliable method for evaluating research and innovation projects—whether past, present, or future instances—we need to have an operationalisable definition of 'responsibility.' Fortunately, we don't have to start from scratch.

In an oft-cited article, Von Schomberg defines RRI as follows:

"Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)."

(Von Schomberg, 2011)

This is a helpful definition, but it can be challenging to grasp its meaning on first glance, so let's break it down.

First, we have the phrase,

"transparent, interactive process by which societal actors and innovators become mutually responsive to each other".

This is straightforward enough. Without transparency and interaction between innovators, researchers, and society, associated outcomes cannot be meaningfully scrutinised or challenged. As such, possible harms or unintended consequences may go unnoticed. For example, if a medical research team failed to interact with and explain to their study participants the risks of a novel drug they are testing, the lack of transparency may prevent the participants from giving their meaningful and informed consent.

Second, we have the goal to which the transparent, interactive process is directed:

"with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process".

The process of (responsible) research and innovation is necessarily dialogical and inclusive because it acknowledges an inherent pluralism of values that are implicated within the research and innovation lifecycle. For instance, many would agree that technological advances in renewable energy are desirable because of their contribution to a more sustainable future. But there may be disagreement about spe-

cific projects, such as the development of a hydroelectric power plant that displaces downstream communities by disrupting the local ecology. Such a tension between values can only be observed and resolved through a transparent and interactive process with affected and impacted stakeholders.

Finally, von Schomberg's definition draws our attention to how RRI facilitates a,

> **"proper embedding of scientific and technological advances in our society".**

This is another way of saying that scientific research and technological innovation do not operate in a vacuum. It draws our attention to the fact that the practices and processes or research and innovation occur at a specific place and time, and also to an awareness of their consequences, which can vary in scope and impact. Obvious examples, such as the development of the atomic bomb, the discovery of penicillin, or the invention of the internet may spring to mind. But all research and innovation has the potential to reshape society and social norms or expectations. RRI takes this embedding as a starting point, taking seriously the moral duties and obligations that such an embedding inculcates.

Von Schomberg's definition helps to make clear the importance of social context for identifying the scope of societal responsibility, and subsequently helps emphasise the need for public and stakeholder engagement in delineating this scope—a point to which we will return. But other key characteristics and principles of RRI remain under the surface.

To identify these characteristics and principles, there are several frameworks for RRI that we could turn to and explore to identify the remaining principles. However, this would give rise to misleading impression that RRI is a top-down approach, in which we start with pre-determined principles that exist independently of the interconnected practices of science, technology, and society. In contrast, RRI is a normative framework that has been constructed slowly and iteratively in response to tangible social harms and felt injustices. The principles that have emerged have been shaped by such events. Therefore, although we will eventually discuss a set of principles for RRI, it is important to first explore some notable historical case studies.

Understanding RRI

The term ‘responsible research and innovation’ is most strongly associated with the European Commission’s Framework Programmes for Research and Technological Development—a set of funding programmes that support research in the European Union. Beginning with the seventh framework programme in 2010, and continuing on through Horizon 2020 (FP8), the term ‘responsible research and innovation’ became increasingly important for the European Commission’s policy.

Since then, other national funding bodies have also shown a commitment to RRI. For example, UKRI’s Engineering and Physical Sciences Research Council have developed the AREA framework, which sets out four principles for RRI: Anticipate, Reflect, Engage, and Act (AREA).

In almost all cases, two significant and motivating drivers behind these policies and principles are (a) an awareness of the impact that science and technology can have on society, and (b) an appreciation of the need to include the public in a dialogue about how science and technology should shape society. The following three case studies help to provide illustrations of these points, while also serving as useful examples that will be returned to in subsequent discussions.

FIG. 1
Doctor drawing blood from a patient as part of the Tuskegee Syphilis Study (Reprinted from Wikimedia Commons)

Case Study 1:

Tuskegee Syphilis Study

Starting in 1932, the U.S. Public Health Service ran a study of "untreated syphilis in the male Negro", which affected almost 400 African-American men with the disease (Reverby, 2001). As we know today, Syphilis can cause many symptoms including sores, blindness, hair loss, stroke, heart failure, and even death if left untreated. However, aside from the risk that these men were exposed to, a particularly abusive aspect of the study was that it was carried out on impoverished individuals, affected by the Great Depression, all while telling them they were being "treated" for their "bad blood" (Reverby, 2001).

The study created a massive outcry but, nevertheless, continued for 40 years until 1972. In this time, syphilis became treatable as a result of the increased availability of penicillin, and funding for the study was withdrawn. However, as one of the participants states,

"The thing that disturbs me now is that they found a cure," Shaw told the Baltimore Sun. "They found penicillin. And they never gave it to us. It vexed me awfully sadly."

(Duff-Brown, 2017)

FIG. 2
Participants of the Tuskegee Syphilis Study

By the time the study ended, 128 participants had died from syphilis or related complications. Moreover, 40 of the participants' spouses had been infected, and 19 children were born with congenital syphilis. It should be obvious why this case study is infamous and well-rehearsed in courses on research ethics or biomedical ethics (Beauchamp and Childress, 2013). However, there are many reasons for the study's continued infamy that have direct relevance to RRI, including the need to continually reflect on the structural biases that exist in society and create forms of racial discrimination that researchers should not ignore. One specific consideration is the fact that the study directly influenced a landmark event in the ethical importance of informed consent: the 1979 Belmont Report (Commission, 1979).

This report, produced by the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, set out a variety of guidelines for clinical research, including the requirement for participants to have an understanding of the research being conducted, in order to provide informed consent.

Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied.

The conditions that need to be satisfied are,

1. **Disclosure of sufficient information**
2. **Ensuring participant's comprehension**
3. **Voluntarily providing consent**

Neither (1) nor (2) were satisfied in the Tuskegee study, and it would be difficult to argue that (3) was satisfied given the socioeconomic deprivation which affected the participants.
Aside from the historical importance of this study, its inclusion here is also to help highlight why so much significance is placed on the responsible participation of stakeholder and affected users or individuals in present day research and activity. We will say more about this in the next section.

FIG. 3
DNA Double Helix

Case Study 2:

Human Genome Project

The Human Genome Project was proposed in 1986 following a project feasibility workshop in Santa Fe, New Mexico. At the time, the US Department of Energy's Health and Environmental Research Advisory Committee urged the department

> **to commit to a large, long-term, multidisciplinary, technological undertaking to order and sequence the human genome**
>
> (Barnhart, 1989)

Nearly 20 years and 2.7 billion dollars later, on April 14, 2003, the project was formally announced as complete. The project had been an international effort to identify all of our ~20,500 genes and determine the sequences of nearly 3 billion chemical base pairs make up our DNA. The scale of the project and the magnitude of the technological accomplishments should be praised in their own rights. But even more impressive is the continuing impact that this project has had on research projects, collaborative practices, and technological advancements (McGuire et al., 2020).

Even from the very beginning of the project, it was clear that expanding our knowledge of genetics and genomics would have profound impacts on society—not all of which would be

positive. For example, concerns were voiced about whether the knowledge could be used to further discriminate against certain individuals or sub-groups of the human population, raising the spectre of past injustices caused by the practice of eugenics [1]. Therefore, in 1990, the National Human Genome Research Institute founded a program to specifically oversee and study the project's ethical, legal, and social implications, known as the ELSI program (NHGRI, 2021).

Writing several years after the completion of the project, the manager of the ELSI program, Michael S. Yesley, offered some remarks that are worth quoting in full.

FIG. 4
DNA in space, conceptual illustration

The qualifications to do bioethics analysis are straightforward: familiarity with, and ability to analyze, the relevant facts and values. No discipline or profession has a monopoly on these skills or should dominate the process. Of the major disciplines engaged in bioethics, philosophy is useful in raising questions and providing rationale, but the actual resolution of bioethics issues – deciding which course of action to take or recommend – generally relies most heavily on factual analysis and seldom on philosophical insight alone. Law sometimes resolves bioethics issues but in most cases establishes only what is socially permissible, not what is most desirable, or merely imposes procedural requirements rather than a substantive result. The ethics traditions of medicine and science pervade bioethics and provide much guidance, but these professional perspectives have built-in conflicts that practitioners may not recognize when balancing the rights and interests of others. Social science is an obvious source of empirical information about both facts and values, but just as other fields play limited roles in bioethics, social science must be integrated in the broad policymaking process. To be useful in this process, social science must view bioethics policymaking as the goal, not the object, of its study. (Yesley [2008], emphasis mine)

Three things stand out about this quotation. First, Yesley is acutely aware of the complex relationship between both facts and values, but also between competing value perspectives when analysing and determining what is desirable from scientific research or technological innovation.

Second, as a lawyer, he also has an appreciation of the normative limitations that can be derived from legal precedents, acknowledging that the law tends to establish that which is socially permissible but not what is most desirable. To paraphrase, the law can help create guardrails but is often unable on its own to set the direction of travel.

And, finally, Yesley recognises that reflection on ethical, legal, and social implications counts for little without an ability to

influence and shape policy. This final point is important, as Yesley continues by acknowledging how the model employed

[1] This concern became the centrepiece for the 1997 dystopian sci-fi, Gattaca, which takes its title from the four letters of the nucleobases of DNA.

by ELSI reflected little consideration of whether it "would be useful or appropriate to develop public policy that could potentially question the direction of the [Human Genome Project's] scientific research". And, furthemore, that the project's scientist-administrators, "established ELSI simply by earmarking funds from their science budget, and they controlled the content of ELSI by determining the boundaries of the funded research. No one would represent the public interest in the administration of ELSI, which would lack at its core an independent, representative entity to analyze the issues, determine research needs, analyze research results and develop well-supported policy recommendations".

At first glance, therefore, the 5% of the annual budget of the Human Genome Project that was earmarked for ELSI research seems impressive, and certainly unprecedented in terms of ethical funding. But Yesley's comments should give us pause to ask how the language of ethics—and indeed, RRI—can be co-opted by commercial, scientific, and political interests, rather than being used to cast a critical perspective on the most pressing questions that face society. This is all the more important when one recognises how many of the questions raised by the Human Genome Project are also brought up in the context of data ethics and AI ethics, such as the possibility of discriminatory outcomes, or concerns about "ethics-washing" by those with vested interests in the advancement of science and technology. As we will see in the next section, it is important to ask what social goal is being served by research and innovation, and whether such a goal is desirable as well as permissible.

Case Study 3:

Cambridge Analytica

In 2013, three researchers at the University of Cambridge and Microsoft Research published a paper in the Proceedings of the National Academy of Sciences. The paper was titled, 'Private traits and attributes are predictable from digital records of human behavior', and it provided details of an application (MyPersonality) that allowed Facebook users to participate in a range of psychometric tests, including a personality test, an intelligence test, and a Satisfaction with Life survey.

Following these tests, users were asked if they were happy for their social media profile data to be collected for research purposes. This included, where available, the various "Likes" of the users; their age, gender, sexual orientation, relationship status, political views, religion, and social network information (e.g. network density); details of the users' consumption of alcohol, drugs, and cigarettes, and whether their parents

FIG. 5
Cambridge Analytica Logo

Cambridge Analytica

stayed together until the user was 21 years old; and also visual inspection of profile pictures, in order to assign ethnicity to a randomly selected subsample of users.

The purpose of gathering this information was to see whether psychological traits could be predicted from social media data. In short, could the results from the user's psychometric tests be inferred from the data gathered from their social media profiles. The results were mixed and gave rise to many questions about validity, reliability, and generalisability (Burr and Cristianini, 2019). However, the results or these related questions are not our present concern.

Almost four years after the publication of their research paper, in 2017, Donald Trump was inaugurated as President of the United States and the United Kingdom gave formal notice of its intent to withdraw from the EU following the Brexit referendum. In an attempt to make sense of these surprising events, a large number of investigative journalists started contacting the three researchers enquiring about their links to a political consulting firm known as Cambridge Analytica. Despite having never worked with Cambridge Analytica, and refusing their requests, the method the researchers described in their 2013 paper led the firm to work with another Cambridge University researcher to develop their own app (Weaver, 2018). It was later revealed that this app enabled Cambridge Analytica to access the data of up to 87 million Facebook users without their knowledge or permission due to poor data privacy and protection policies on the platform (Kang and Frenkel, 2018). The firm subsequently used this data to develop and sell predictive analytics services to the Trump administration, influencing the outcomes of the US election, and also attempt to court the Leave.EU Brexit campaign (Ball, 2020).

The events that surround the Cambridge Analytica data scandal read as though they were plucked from a novel about espionage, information warfare, and PSYOPS, carefully woven together and exposed by the tireless efforts of investigative journalists (Cadwalladr, 2017). Our interest in the case study is, unfortunately, less glamorous. It can be captured by a single question:

If you were one of the original researchers, back in 2013, investigating whether social media data could be used to infer otherwise private information about the psychological attitudes and beliefs of users, would you think you were behaving irresponsibly by publishing your research?

In our first activity, we'll reflect on this question and others related to these three case studies.

Activity 1:
Exploring Case Studies

In this activity you will answer several questions related to the three case studies and either discuss in a group or reflect on them on your own.

Please visit the course website to view the associated instructions.

Science, Technology, and Society

In April 1945, Michael Polanyi—a chemist and sociologist of science—and Bertrand Russell—a philosopher and logician—were speaking on a radio programme about the practical implications of the famous formula, $E=mc^2$. They were asked whether the formula had any practical applications for society, but neither could provide an answer. Three months later the Manhattan project dropped the first of their three atomic bombs!

(Bridgstock, 1998) draws attention to this story because it was used originally by Polanyi, in his essay, 'The Republic of Science' (Polanyi, 1962), to suggest that the practical and societal outcomes of pure scientific research are often unforeseen and unintended. The problem with this suggestion is that it implies that scientists cannot be held accountable or exercise any real responsibility for the consequences of their research—a troubling implication if true!

However, the definitions we have already encountered suggest that science, technology, and society are closely interconnected, and that RRI requires reflection upon the myriad ways that science and technology can impact and shape social norms and practices. Responsibility arises out of this relationship. But if the impacts or consequences are unforeseeable and often unanticipated, then the principles of RRI may be too demanding.

Fortunately, the implications of Polanyi's example are narrow in scope. The following thought experiment will help us illustrate why.

THE CARELESS CEO

Imagine a CEO of a large manufacturing company is approached by one of her scientific advisors and informed that a project that she has proposed will require an environmental impact assessment before it can proceed. The CEO dismisses this and orders that the project continue without the assessment. Furthermore, she callously proclaims that she does not care what the environmental consequences of the project may be. All she is interested in is making as much profit as possible. As it turns out, the project ends up causing vast amounts of pollution that cause irreparable harm to the nearby flora and fauna, and also affects the health of a community living downstream of the manufacturing plant. The CEO is, rightfully, held accountable, both morally and legally, and is prosecuted when her dismissal of the impact assessment is uncovered.

Few would take issue with these consequences for the CEO. She had a responsibility to ensure her company's processes operated in a safe and ethical manner, but chose to wilfully neglect this responsibility in spite of receiving advice from her staff. However, let's alter some of the details of this case while keeping the logical structure the same.

This time, everything about the thought experiment remains the same except for the outcomes of the project. Now, the project causes no harm. In fact, the efficiency of the new project actually reduces the company's emissions and leads to more sustainable operations.

There is no need to worry about accountability in this instance, as no harms occurred. But does the CEO deserve praise for her actions? This question is less likely to have a universal consensus among the answers.

Those that believe the CEO deserves no praise are likely to point to the fact that she did not carry out her due diligence or reflect upon the possible consequences of her project. She chose to ignore the suggestion of undertaking an environmental impact assessment, and, therefore, did not act in a responsible manner. She was unable to anticipate any harms or benefits because she did not gather the appropriate evidence. Neither did she act with deliberate intention, but instead acted in a careless manner. In short, she was lucky that the outcomes were positive, and because of a lack of deliberate or intentional action some may argue that there are no legitimate grounds for praise.

Others may disagree, and simply argue that the consequences are all that matter. We won't try to settle this debate here. Instead, it can be left for personal reflection.

Returning to the issue raised by Polanyi's statement, we can of course acknowledge that he is right to suggest that the societal impacts or consequences of some pure scientific research are hard to anticipate. This is especially true for the more distant effects—consider again the long-term impact of the Human Genome Project, which is arguably still affecting current research. But this is not the case for all research or innovation projects. Many consequences can in fact be more readily anticipated or predicted by carrying out careful activities of reflection and deliberation.

There is no question about whether scientists, researchers, and developers have some responsibility for the applications of their research. This must be true for them to be praised for the positive outcomes, which they often are, and also to be held accountable and blamed for the negative consequences when they occur. The question is, rather, when they should receive praise and blame. To help us address this question, we will look at some of the practical ways that scientists, researchers, and developers can take responsibility for the social impacts of their research, in order to maximise the potential opportunities and minimise the possible harms associated with their work. This is a primary objective for the entire course, but we can introduce some of the practical mechanisms now.

Risk and Impact Assessments

At the start of this chapter we looked at a definition of RRI from (Von Schomberg, 2011). Let's look at another one, this time from the European Commission:

European Commission

Responsible research and innovation is an approach that anticipates and assesses potential implications and societal expectations with regard to research and innovation, with the aim to foster the design of inclusive and sustainable research and innovation. (Commission, 2014)

René von Schomberg

Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society). (Von Schomberg, 2011)

There are many similarities between the two definitions, but they also emphasise different aspects through their choice of terminology. For instance, the definition from the European Commission mentions anticipation and assessment, as a means for highlighting the importance of associated activities such as risk or impact assessments. There are many types of risk and impact assessments that can be carried out, such as safety and risk assessments, equality impact assessments, human rights impact assessments, and, of course, data protection impact assessments.

The necessity of such assessments will be familiar to those who work in commercial or public sector organisations, but less so to those in academic institutions. Typically they are carried out for compliance reasons. However, the structured nature of such assessments can also support more ethical forms of reflection and anticipation.

It is not necessary to present an overview of all the different impact assessments that could be useful within the context of RRI.[1] Instead, we will focus on a process that is central to almost all forms of risk or impact assessments: stakeholder participation.

[1] See the 'Further Resources' section for links to guides for each of these activities.

Inclusive and Deliberative Stakeholder Participation

QUESTION

Why should stakeholders be included in a research or innovation project?

There are at least two answers that can be given to this question?

1. **To identify and meet stakeholder or user needs**

2. **To ensure that possible harms that could arise are identified and addressed**

The first of these answers is reminiscent of a typical stage in product design. For example, a design committee for a new product may reach out to possible consumers/users to identify what they think about a range of prototypes or to gather feedback about a possible feature. Such processes have what we can refer to as 'instrumental value.' That is, the purpose of including stakeholders or users is directed towards the benefit it brings to the project. Their participation serves an instrumental role in obtaining a goal, such as developing a product that is more likely to sell.

A similar claim could also be made for the second answer. However, in this instance, the focus is on mitigating risks or harms, as opposed to realising benefits. Here, the stakeholders or users still play an instrumental role in securing a goal of the project team (e.g., to ensure no harm is caused by their research or project).

However, stakeholder engagement has a further intrinsic value, which is more clearly exposed when science and technology are properly situated in a social context.

Framed as a third answer that appropriates the language of the above two definitions:

Stakeholder participation is a necessary component of responsible design, development, and deployment. It recognises the need for stakeholders and innovators to become mutually responsive to each other in an inclusive and deliberative process that aims at realising sustainable research and innovation practices that promote the social good.

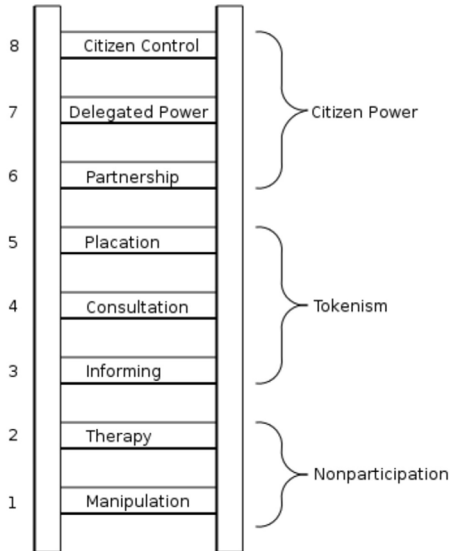
This answer helps to expose a limitation of the instrumental perspective. That is, stakeholder engagement is about more than the identification of potential risks and benefits; just as RRI is about more than the avoidance of gross misconduct (e.g., plagiarism, fabrication/falsification of results, developing obviously dangerous technologies). It is also about recognising the right that all members of society have to participate in science and innovation, especially insofar as it relates to how science pursues goals that shape and alter social norms and expectations. And, to be clear, it is a right—one which is captured in Article 27 of the Universal Declaration of Human Rights:

> **Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits.**

RRI relies on both the organic and creative nature of the human mind as much as it does the meticulous standards of the scientific method. Involving diverse stakeholders in creative forms of participation can prevent a sort of cognitive staleness or group-think arising within teams when it comes to problem solving. But more importantly, participation in research and innovation can improve trust, buy-in, and promote scientific understanding and literacy. This goes beyond a narrow instrumental value to a broader social and ethical value. It is also, arguably, part of the motivation behind famous ladder of public engagement (Arnstein, 1969).

FIG. 6

Sherry Arnstein's Ladder of Engagement (Reprinted from Wikimedia Commons)



To summarise, responsible participation goes beyond informing or consultation (i.e., monological forms of engagement). It aims at dialogical engagement that is representative of meaningful partnerships and, hopefully, sustainable forms of devolution of power where members of society are equipped with the capabilities necessary to have legitimate control over the outcomes of scientific research or technological innovation. Our next activity will help us widen our sphere of consideration for ethical reflection and deliberation, in order to support these forms of participation and engagement.

Science and Technology Studies (A Timeline)

This section provides a timeline for Science and Technology Studies (STS), which can be treated as a reference and an optional resource for the RRI course.

Science and technology studies (STS) is the interdisciplinary study of how science and technology shape, structure, affect and interact with society, politics, and cultural knowledge and values. Because of its inherent interdisciplinarity, it arose as the result of the convergence of different ideas and areas of interest, each of which have their own distinct histories. However, many of these intertwined ideas shared important themes, such as (a) viewing science as a social institution with a distinct normative structure (e.g., values of prioritising novel research; ethos and community of peers), (b)) treating science and technology as socially situated practices, and (c) the belief that technology and society co-constitute and mutually shape each other, including the way that scientific facts and technological artefacts are understood and conceptualised (Merton, 1973).

As Rohracher notes:

> **Facts and artifacts are but temporarily stable outcomes of heterogeneous activities of scientists and engineers and their entanglement in wider social and political relations.**
>
> —Rohracher (2015)

Because of these themes, the concerns of STS overlap closely with the concerns of RRI. Therefore, on the following page you will find an incomplete history (or, timeline) of STS that serves as a reference to some notable publications and events that you may find interesting to explore.[1]

A more detailed version of the timeline is available on the course website.

Activity 2:
Ethical Reflection and Deliberation

In this activity you will engage in a structured form of reflection and deliberation to identify the respective agents and subjects for a variety of moral decisions.

Please visit the course website to view the associated instructions.

[1] The following timelines is based on the account by (Rohracher, 2015), which goes back further than the 1960s.

STS Timeline

1960s:

Influential Precursors to STS

- **(1962)** Thomas Kuhn publishes, 'The Structure of Scientific Revolutions'
- **(Mid 1960s)** Women's studies grows as an academic field
- **(1968)** Garrett Hardin publishes an essay in Science that popularises the phrase "tragedy of the commons"

1970s:

STS Emerges

- **(1972)** First STS program developed by Elting E. Morison—a historian of technology at MIT
- **(1979)** Bruno Latour and Steve Woolgar publish the first edition of their study, 'Laboratory Life'

1980s:

The Social Construction of Technology

- **(1981)** The European Association for the Study of Science and Technology is formed
- **(1984)** Trevor Pinch and Wiebe Bijker publish, 'The social construction of facts and artifacts'
- **(Late 1980s)** Actor-Network Theory is developed at the Centre de Sociologie de l'Innovation

1990s:

Engaging Power and Feminist STS

- **(1991)** Donna Harraway publishes 'Situated knowledges: the science question in feminism and the privilege of partial perspective' and Judy Wajcman publishes 'Feminism Confronts Technology'
- **(1991)** Sandra Harding publishes 'Whose Science? Whose Knowledge? Thinking From Women's Lives.'
- **(1999)** Ian Hacking publishes the 'Social Construction of What?"

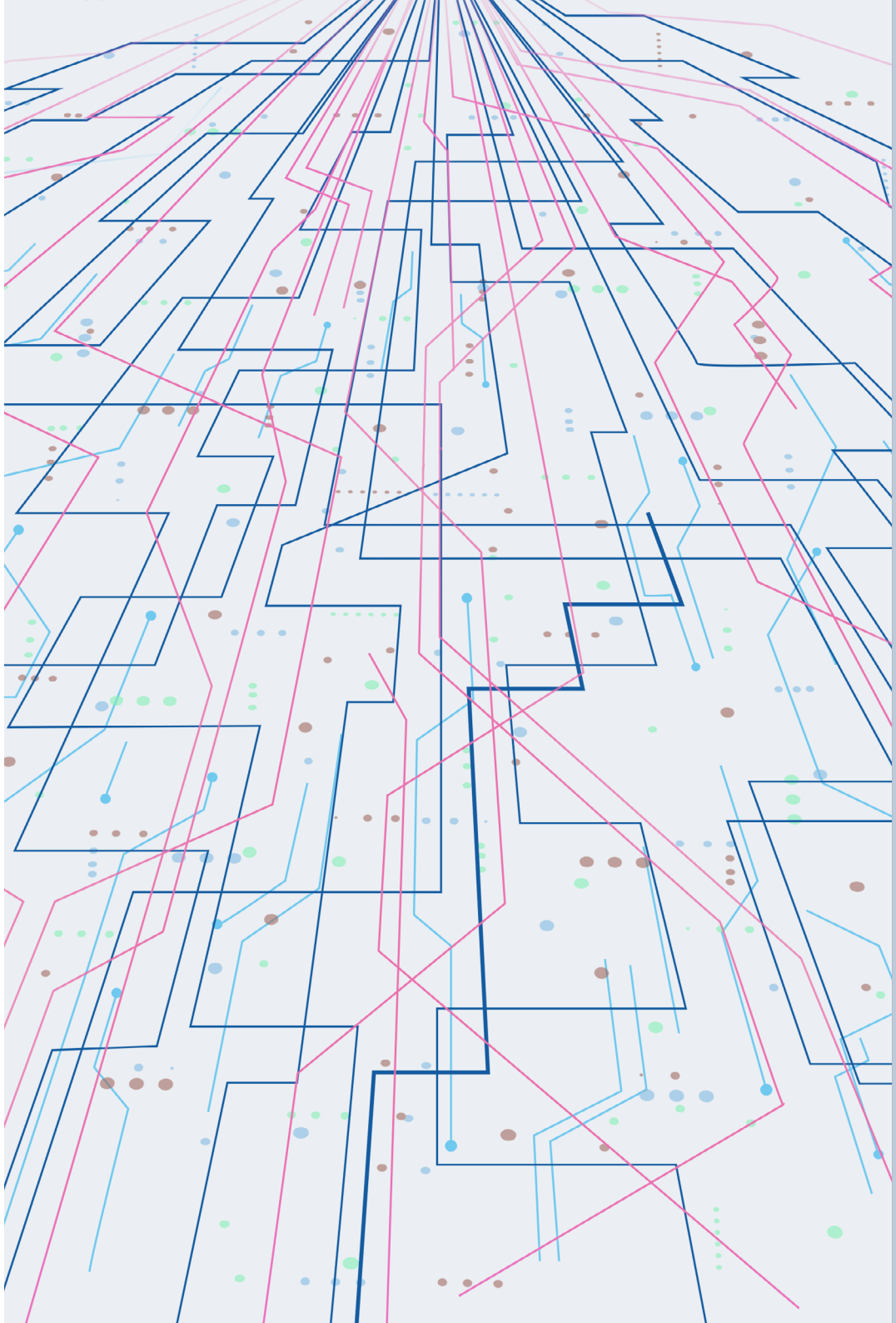
2000s:

Engaging Power and Feminist STS

- **(2000)** The UK's House of Lords Select Committee on Science and Technology publish a report titled 'Science and Society'
- **(2002)** H.M. Collins and Robert Evans explore the notion of 'scientific expertise' as it relates to public decision-making
- **(2009)** Sheila Jasonoff and Sang-Hyun Kim introduce the concept of "sociotechnical imaginaries"

03

Responsible Data Science and AI



Summary

This chapter builds on the content of the previous one by applying many of the central concepts to research and innovation in data science and AI. First, we will look at what being responsible means in the context of data science and AI, and explore several principles that can help get us started with operationalising the term 'responsibility.' Next, we explore a simple model that has been designed to help with reflection and deliberation throughout the project lifecycle, and also look at what this model means for individual roles within a project, as well as a broader notion of collective responsibility. Finally, we examine the concept of 'bias,' which will play an important role in the subsequent chapters.

Learning Objectives

In this chapter, you will:

- Explore what differentiates responsible data science and AI from responsible research and innovation more generally.
- Examine a model of a typical project lifecycle to better appreciate why individual responsibility is often insufficient in the context of data science and AI.
- Understand the differences between social, statistical, and cognitive biases, and why they all matter for responsible data science and AI.

Responsible Data Science and AI

What separates responsible data science and AI from responsible research and innovation more generally?

We saw in the previous chapter how RRI can be defined with reference to concepts that emphasise the need for ethical reflection on possible social harms and benefits, supported by inclusive participation of affected stakeholders. Responsible data science and AI shares this emphasis, but can be further refined by considering more specific principles that are geared towards the particular harms and benefits associated with data science and AI. These principles can help us identify what is unique to responsible data science and AI.

SAFE-D Principles

According to Mittelstadt (2019), in 2019 there were at least 84 statements that provided "high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI". By now there are surely many, many more!

In response to this proliferation of principles, some have attempted to distil and condense the myriad documents, in order to identify commonalities and extract a unified list of shared principles (Floridi and Cowls, 2019, Jobin et al., 2019). However, regardless of which set of principles we start with, one thing remains the same: good principles should support ongoing reflection and deliberation; they are not decision procedures in their own right.

This point is sometimes lost in the ensuing debate about which set of principles should be used or adhered to, or which set is best. However, what matters is that the set of principles should (a) be responsive to the actual harms and benefits that matter to the communities of affected individuals, (b) be underwritten by a set of shared values, which support and motivate dialogue between stakeholders1, and (c) serve as starting points in a wider process of reflection and deliberation.

With these points in mind, we will make use of the following set of principles known as the 'SAFE-D principles':

Safety

Accountability

Fairness

Explainability

Data Quality, Integrity, Protection and Privacy

These principles are grounded in comprehensive research and understanding of human rights and data protection law, as well as applied ethics of data and AI. (Leslie et al., 2021)

Examples

You can click through the following illustrative examples to get an idea of some of the social harms associated with data-driven technologies:

Predicting Risk

Avon and Somerset Police and Bristol City Council developed a sophisticated predictive risk tool that was used, among other things, to predict the risk of children suffering sexual abuse. But, the Bristol Cable reported that many children were falsely flagged as being at risk, and that the tool was developed using dozens of public sector databases, including schools, housing, NHS records, and even credit scores from Experian.

Facebook Discriminatory Job Adverts

The algorithmic system used by Facebook to automatically show job adverts to users it believes are most likely to engage with them was reported to perpetuate discriminatory gender norms. Thee BBC reported that

almost all Facebook users shown adverts for mechanics were men, while ads for nursery nurses were seen almost exclusively by women.

Racist Photo Cropping Tool

Twitter was forced to apologise after many users reported that the automated tool for cropping images on the social media platform showed a racial bias towards faces of white people over faces of black people. According to Twitter, one source of the issue was the use of a "saliency algorithm" that was trained on human eye-tracking data

Lethal Autonomous Weapons

Turkish company STM manufactures the Kargu-2—an attack drone that can operate autonomously by using machine learning and real-time image processing to identify targets. According to a UN security council report this drone was reported to have been used to "remotely engage" and "hunt down" logistics convoys and retreating forces in the Libyan civil war during 2019.

#TravelingWhileTrans

In their book, Design Justice, Sasha Costanza-Chock highlights how the design of sociotechnical systems reinforce and embed a variety of social norms and expectations that can be harmful to vulnerable or marginalised communities. For example, the impact of full-body scanners at airport security that require operators to select either 'male' or 'female', even when presented with non-binary or trans individuals whose bodies may not conform to the models embedded within the machine.

Any Others?

Do you know any other examples of social harms associated with data-driven technologies?

Each of the SAFE-D principles is either motivated by and captures a specific set of harms that have been uncovered and exposed, responds to a set or well-documented risks that arise in the context of data science and AI, or is oriented towards the sustainable, ethical, and responsible use of data-driven technologies. Let's look at each principle in turn.

FIG. 7
Good principles can help us make sense of a maze of competing values, interests, and features that affect moral decision-making.

Safety

Safety can mean a couple of things. From a technical perspective, safety requires the outputs of a project to be secure, robust, and reliable. For example, if an organisation is developing an autonomous vehicle, it should operate safely in the intended context of use. However, in the context of responsible data science and AI, safety also has a social sustainability component. This aspect of safety requires a project's practices to be informed by ongoing consideration of the risk of exposing individuals to harms even after the system has been deployed and the project completed—a long-term (or sustainable) safety.

Accountability

Accountability can refer to transparency of processes and associated outcomes that enable people to understand how a project was conducted (e.g., project documentation), or why a specific decision was reached. But it can also refer to broader processes of responsible project governance that seek to establish clear roles of responsibility where full transparency may be inappropriate (e.g., confidential projects).

Fairness

Fairness is inseparably connected with legal conceptions of equality and justice, which may emphasize a variety of features such as non-discrimination, equitable outcomes, or procedural fairness through bias mitigation. However, these notions serve as a subset of broader normative considerations pertaining to social justice, socioeconomic capabilities, diversity and inclusivity. For this reason, the term 'fairness' can be confusing due to the wide variety of ways it is employed, and the large number of more specific concepts that fall within its scope.

Explainability

Explainability is a key condition for autonomous and informed decision-making in situations where data-driven systems interact with or influence human judgement and choice behaviour. Explainability goes beyond the ability to merely interpret specific aspects of a project (e.g., interpreting the parameters of a model); it also depends on the ability to provide an accessible and relevant information base about the processes behind the outcome.

Data Quality, Integrity, Protection and Privacy

Data quality, integrity, protection and privacy must all be established to be confident that a research or innovation project has been designed, developed, and deployed in a responsible manner.

'Data Quality' captures the static properties of data, such as whether they are (a) relevant to and representative of the domain and use context, (b) balanced and complete in terms of how well the dataset represents the underlying datwa generating process, and (c) up-to-date and accurate as required by the project.

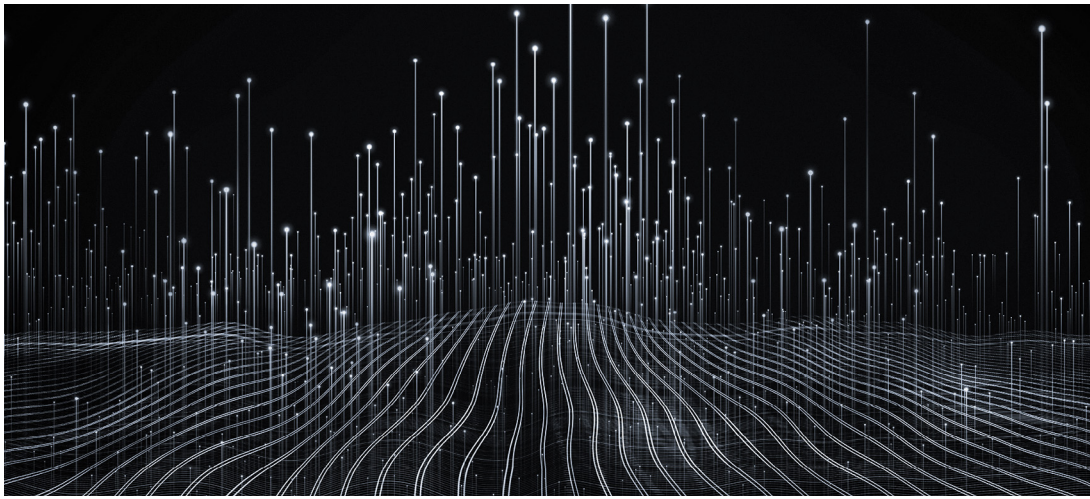
'Data Integrity' refers to more dynamic properties of data stewardship, such as how a dataset evolves over the course of a project lifecycle. In this manner, data integrity requires (a) contemporaneous and attributable records from the start of a project (e.g., process logs; research statements), (b) ensuring consistent and verifiable means of data analysis or processing during development, and (c) taking steps to establish findable, accessible, interoperable, and reusable records towards the end of a project's lifecycle.[2]

'Data protection and privacy' reflect ongoing developments and priorities as set out in relevant legislation and regulation of data practices as they pertain to fundamental rights and freedoms, democracy, and the rule of law. For example, the right for data subjects to have inaccurate personal data rectified or erased. (ICO, 2021)

Each of these principles can be treated as a goal to which responsible data science and AI ought to be directed.3 But, on their own they are insufficient for establishing what specific actions or decisions should be taken in any given project. For instance, what does it mean to develop a fair diagnostic model in healthcare?

- **Does it mean ensuring that all patients are exposed to the same level of risk with respect to the distribution of possible false negatives?**
- **What about false positives instead?**
- **What about the use of the decision support system in which thee model is implemented?**
- **Will it be used in all hospitals on all patients?**
- **Or, will only those wealthy enough to afford private healthcare receive this service?**
- **Should it instead be used for the most vulnerable and worse off in society?**

Questions such as these have no straightforward answer and are heavily context-dependant. Even if consensus were to be reached for a specific model used, say, in the diagnosis of lung cancer (Svoboda, 2020), this would be no guarantee of a similar answer in a different area of healthcare (e.g., paediatrics, mental healthcare), or even for another diagnostic model in radiology (e.g., MRI instead of CT scans). Therefore, starting in the next section we will look at a model for helping us get a clear grasp of the situated and sociotechnical context under consideration in research and innovation projects.



[1] We won't say much about ethical values in this course. However, the course on AI Ethics & Governance focuses on them directly.
[2] These are known as the FAIR principles (read more here).
[3] In our guidebook on public communication of science we formalise this notion of an ethical goal in relation to a method known as argument-based assurance. Here, the goals are supported by specific properties that must be established in a project, in order to provide justifiable assurance to stakeholders that the respective goal has been realised.

Introducing the Project Lifecycle (A Sociotechnical Approach)

There are many ways of carving up the lifecycle for a data science or AI project (hereafter, 'project lifecycle').[1] For instance, (Sweenor et al., 2020) break it into four stages: Build, Manage, Deploy & Integrate, Monitor.[2] Ashmore et al. (2019) also identify four stages, which have a more specific focus on data science: data management, model learning, model verification, and model deployment.

The multiplicity of approaches is likely a product of the evolution of diverse methods in data mining/analytics, the significant impact of ML on research and innovation, and the specific practices and considerations inherent to each of the various domains where ML techniques are applied. While there are many benefits of existing frameworks, they do not tend to focus on the wider social or ethical aspects that interweave throughout the various stages of a ML lifecycle.

Fig. 7, therefore, presents a model of a typical lifecycle for a project involving data science or the production of an ML/AI system. We have designed this model to support the ethical reflection and deliberation that is characteristic of responsible data science and AI, while remaining faithful to the technical details. However, it is important to note that the model is a heuristic for reflection and deliberation. Therefore, it is not intended to be perfectly capture of describe the processes for all data science or AI projects.

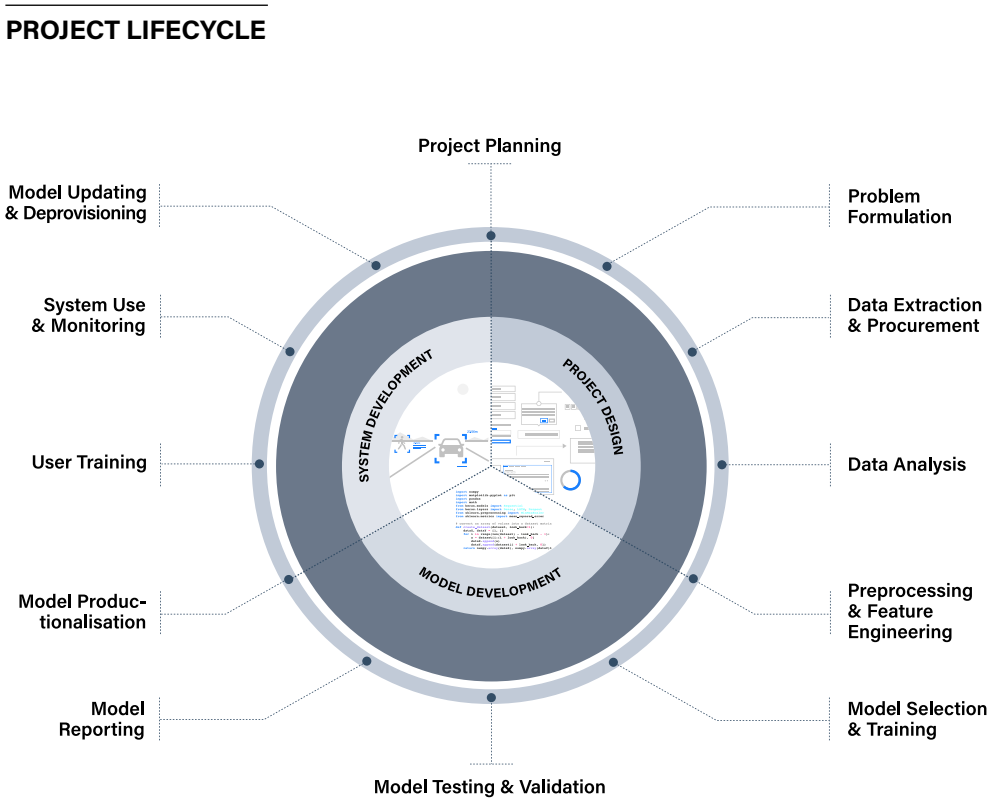


FIG. 8
The Project Lifecycle. The overarching stages of design, development, and deployment (for a typical data-driven project) can be split into indicative tasks and activities. In practice, both the stages and the tasks will overlap with their neighbours, and may be revisited where a particular task requires an iterative approach. The spiral indicates that this is a diachronic, macroscopic process that evolves and develops over time, and as the deployment stage finishes, a new iteration is likely to begin.

To begin, the inner circle breaks the project lifecycle into three processes:

1. (Project) Design
2. (Model) Development
3. (System) Deployment

These terms are intended to be maximally inclusive. For example, the design stage encompasses any project task or decision-making process that scaffolds or sets constraints on later project stages (i.e. design system constraints). Importantly, this includes ethical, social, and legal constraints, which we will discuss later.

Each of the stages shades into its neighbours, as there is no clear boundary that differentiates certain project design activities (e.g. data extraction and exploratory analysis) from model design activities (e.g. preprocessing and feature engineering, model selection). As such, the design stage overlaps with the development stage, but the latter extends to include the actual process of training, testing, and validating a ML model. Similarly, the process of productionalising a model within a system can be thought of as both a development and deployment activity. And, so, the deployment stage overlaps with the 'development' stage, and also overlaps with the 'design' stage because the deployment of a system should be thought of as an ongoing process (e.g. where new data are used to continuously train the ML model, or, the decision to de-provision a model may require the planning and design of a new model if the older (legacy) system becomes outdated). For these reasons, the project lifecycle is depicted as a spiral. However, despite the unidirectional nature of the arrows, we also acknowledge that ML/AI research and innovation is frequently an iterative process. Therefore, the singular direction is only present at a macroscopic level of abstraction (i.e., the overall direction of progress for a project), and allows for some inevitable back and forth between the stages at the microscopic level.

The three higher-level stages can be thought of as a useful heuristic for approaching the project lifecycle. However, each higher-level stage subsumes a wide variety of tasks and activities that are likely to be carried out by different individuals, teams, and organisations, depending on their specific roles and responsibilities (e.g. procurement of data). Therefore, it is important to break each of the three higher-level stages into their (typical) constituent parts, which are likely to vary to some extent between specific projects or within particular organisations. In doing so, we expose a wide range of diverse tasks, each of which give rise to a variety of ethical, social, and legal challenges.

Chapter 4 is dedicated to exploring each specific stage and activity in detail. The remainder of this chapter covers some topics that apply to the project lifecycle as a whole.

[1] The following text is adapted from a publication titled, 'Ethical Assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies' (Burr and Leslie, 2021).
[2] These four stages are influenced by an 'ML OPs' perspective. The term 'MLOps' refers to the application of DevOps practices to ML pipelines. The term is often used in an inclusive manner to incorporate traditional statistical or data science practices that support the ML lifecycle, but are not themselves constitutive of machine learning (e.g. exploratory data analysis), as well as deployment practices that are important within business and operational contexts (e.g. monitoring key performance indicators).

Roles and Responsibilities

One thing should be clear from the wide variety of activities associated with the project lifecycle... designing, developing, and deploying an AI system is not a one-person task!

The activities and processes which comprise a typical AI project lifecycle involve a wide-ranging and far-reaching set of skills and capacities. Such skills are usually encapsulated within myriad roles and responsibilities: project commissioner, product manager, data protection officer, data scientist, system engineer, etc. However, the way such roles and responsibilities are often defined at an organisational or institutional level (e.g. in job specifications) tend to reflect the practical demands of organisational efficiency or HR management, rather than the normative demands of ethical and responsible governance.

In reality, the individual roles and responsibilities, which are implicated in the design, development, and deployment of complex AI projects, are interwoven to such an extent that they form an inextricable Gordian Knot of collective responsibility. While there will always remain a pragmatic need to have individual roles and responsibilities for complex projects, which may be undertaken by a single person or a team, due to the inescapable burden of time constraints and finite cognitive resources, such a pragmatic consideration does not provide the individual project member with a morally defensible reason for excusing themselves from the shared and collective responsibility.

What is needed, however, is a clear illustration of how this collective responsibility is instantiated throughout the lifecycle of an AI project—or a project involving the development of a data-driven technology. **FIG. 9** offers a stylised illustration of the project lifecycle, introduced in the previous section, that we can build on.

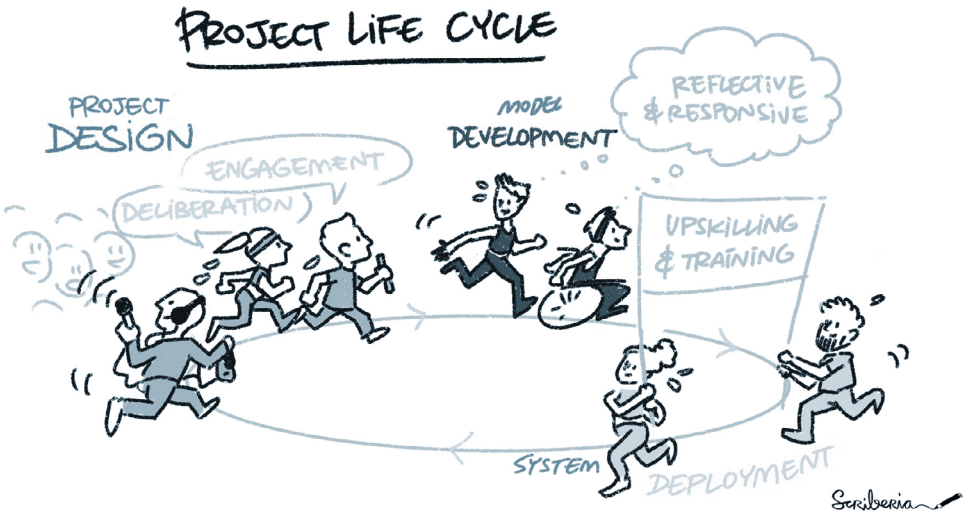


FIG. 9
A stylised version of the project lifecycle, emphasising the collective responsibility for its success. Image by Scriberia Zenodo

A key benefit of this illustration is that is helps us reflect on the way that individual roles and responsibilities within the project team interconnect. For instance, the developers or software engineers, in virtue of their expertise in implementing models and designing software tools may have a better understanding of how UX/UI elements will affect the decision-making of the end user, perhaps in virtue of their impact on cognitive biases (Burton et al., 2020). As such, they may recognise the need to ensure that any output of an automated decision support system should be accompanied by a graphical representation that helps explain how an algorithm reached any particular decision.

Therefore, reducing individual project members to their functional role devalues their creative and reflective potential. Taking our developer as an example, again, it is worth noting that bringing them more upstream into, say, the project design discussions with a product manager and stakeholders may allow them to hear things that the product manager would not otherwise know to ask them or tell them.

FIG. 7, therefore, serves as a starting point for reflecting on how a project team share a collective interest in realising a shared goal (i.e. the responsible design, development, and deployment of a data-driven technology), which requires all team members to reflect on how their own roles and responsibilities intersect or impact with the various stages of the project lifecycle beyond those that fall within their immediate remit of, say, a functional job description.

Activity 3:
Contributing to Collaborative Projects

If you are following this guide as part of an online course, there is an associated activity that is designed to help you collaborate with your team. Please visit the course website to view the associated instructions.

If you are reading this as a self-directed exercise, you can instead visit https://the-turing-way.netlify.app/collaboration to read more about collaborative research.

The Dynamics of Project Teams

There are, of course, a wide variety of ways that a project team can be established, and projects can come in all shapes and sizes. Some projects may be distributed across countries or continents. Some may be based within a single department in a university or research institute. Others may operate across multiple organisations, requiring complex management structures to ensure the project is governed effectively. Others may rely on the voluntary contribution of passionate individuals, who work as a decentralised cooperative, as is common in open source software.

Each structure comes with its own benefits and drawbacks. A common drawback though is the existence of power imbalances that can lead to internal rifts or complications within a project team.

For example, it is common within academic research projects for their to be a lead or principal investigator who is in charge of the project. Often, this individual is well-established in their academic profession, and (to some extent) protected from the precarious nature of academic employment. Although they will have a wide-range of faculty responsibilities to juggle, the security afforded by more senior positions, and the prestige and influence associated with them, creates many opportunities. This is not the case, however, for more junior members of the project, such as PhD students or research assistants. For these individuals, a core priority may be ensuring they obtain a key publication in a top-tier journal, prior to graduating and exploring post-doctoral positions or teaching roles within a university. Like their senior counterparts, these individuals have their own priorities and individual responsibilities to consider. Unfortunately, the resulting picture is not always a harmonious one, and when priorities clash the power imbalance often favours the more senior individual.

A similar picture emerges in technology companies, as is illustrated nicely by Tanya Reilly in her excellent blog article, 'Being Glue.' Rather than summarising the article, I'd encourage you to set aside the time to read it.

There are, of course, many more ways that project teams can lead to power imbalances or challenging dynamics. Varying levels of public and private funding between countries can create systematic forms of advantage for research teams at more prestigious institutions or citizens of wealthier countries. Language barriers (or disciplinary assumptions) can impede transparent and effective communication, both within and between research teams, and also between innovators and the public. Physical and Mental Disabilities, and socioeconomic disparities (e.g., educational background), contribute to an uneven (and often unjust) playing field. As do more general social or organisational norms or expectations, such as those associated with neurotypicality.

All of these considerations, and more, shape the dynamics of project teams. It is important, therefore, to take the time to reflect on our own position and standing within our teams, and to identify if there is anything we can do to support each other. The more that individual's take this personal responsibility, the easier it becomes to share in a collective responsibility.

Understanding Bias

FIG. 10
An artistic representation of social, statistical, and cognitive biases (by Johnny Lighthands)

Before we start exploring the project lifecycle and its associated activiyies there is a final topic that we need to explore. You will likely have some prior familiarity of the term 'bias', but your understanding of the concept may emphasise specific properties that reflect a specific focus of your research background. This section will present three common ways that 'bias' can be understood as it applies to and affects the research and innovation lifecycle. The three perspectives that we will look at are: social, statistical, and cognitive bias.

Amazon's recruitment tool that perpetuated bias in hiring against women.

This algorithmic system learned to perpetuate a bias to prefer male candidates to female candidates because this reflected past hiring decisions.

Predictive policing that use geospatial data to try to learn associations between places, events, and historical crime rates.

The attempt to predict where and when crimes are more likely to happen can create a positive feedback loop, which results in over-policing that may exacerbate tensions between communities and police.

Clinical decision support systems can contribute to existing forms of racial bias in access to healthcare.

A study conducted in the US found that an algorithm that used health costs as a proxy for health needs was "less likely to refer black people than white people who were equally sick to programmes that aim to improve care for patients with complex medical needs". (Obermeyer et al., 2019)

Social Bias

Outside of research and development communities, the term 'bias' is often associated with some form of prejudice or discrimination. For example, an inclination or disposition to treat an individual or organisation in a way that is considered to be unfair. This understanding of bias is necessary to draw attention to pre-existing or historical patterns of discrimination and social injustice that can be perpetuated, reinforced, or exacerbated through the development and deployment of data-driven technologies. There are numerous examples that illustrate this point.

A commonly heard response to such examples is the claim that the underlying problem is that the training data used to develop the algorithms or models were insufficiently representative. In other words,"it was the data that were biased". The assumption behind this claim is that better data collection would solve the problem. Unfortunately, at best this response is only partially true, but at worst it belies a commitment to a form of 'technological solutionism' [1] that often ignores how technology affects social practices and norms.

It is important to remember that most decisions that drive the project lifecycle are made by the project team. A choice to design a study in a particular way, or to deploy a system in a context that is characterised by patterns of historical discrimination, cannot simply be blamed on poor data.

Statistical Bias

If you are a data scientist, or use techniques from data science in your research or development, then it is likely that your understanding of bias is influenced by statistical concepts.

Jeff Aronson explores the etymology of the term 'bias' in a series of interesting blog articles, which emphasise the statistical understanding of bias [Aronson, 2018]. He begins by tracing it back to the game of bowls, where the curved trajectory of the bowl as it ran along the green reflected the asymmetric shape of the bowl (i.e., its bias). However, according to Aronson, the term was not used in statistics until around the start of the 20th Century where it was used to refer to a systematic deviation from an expected statistical result that arises due to the influence of some additional factor. This understanding is common in observational studies where bias can arise in the process of sampling or measurement.

On the basis of his historical review, Aronson identifies six features of definitions of bias that adopt a statistical perspective:

Systematicity: bias arises from a systematic process, rather than a random or chance process

- **Truth:** a realist assumption that the deviation is from a true state of the world
- **Error:** the bias reflects an error, perhaps due to sampling or measurement
- **Deviation (or Distortion):** a quantity in which the observed result is taken to differ from the actual result were there no bias.
- **Affected elements:** the study elements that may be affected by the bias include the conception, design, and conduct of the study, as well as the collection, analysis, interpretation, and representation of the data
- **Direction:** the deviation is directional, as it can be caused by both an under- or over-estimation

Some of these features are specific to a statistical framing of 'bias', but some also apply to the other two perspectives For instance, 'systematicity' is arguably a necessary property for social biases (i.e., a bias that systematically leads to discriminatory outcomes). And, 'error' is sometimes a property of our next perspective: cognitive biases.

Cognitive Bias

Our modern understanding of cognitive biases has been most heavily influenced by research conducted by Daniel Kahneman and Amos Tversky. A lot of their work exposed a wide variety of psychological vulnerabilities, which impact our judgement and decision-making capabilities. In short, their experiments showed how individuals rely on an assortment of heuristics or biases, which speed up judgement and decision-making but also lead us astray.

For example, consider the following example

The Linda Problem

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

Try to answer this question yourself, before you reveal the answer.

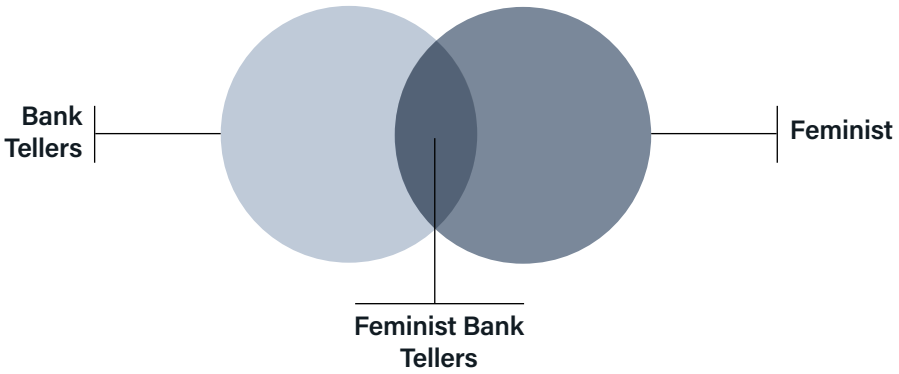
The correct answer is (1). Did you get it right?

If you got it wrong, you have just committed what is known as the 'conjunction fallacy.' But don't worry you're in very good company! When Tvserky and Kahneman posed this question to a group of 88 undergraduate students, only 15 got the correct answer (Tversky and Kahneman, 1983).

The reason it is (1) is because the probability of two events occurring *in conjunction*, such as Linda being both a 'bank teller' and 'active in the feminist movement' must be less than or equal to the probability of either event occurring on its own. Formally, for two events A and B:

$$Pr\ (A\ ^\wedge\ B) \leq Pr\ (A)\ and\ Pr\ (A\ ^\wedge\ B) \leq Pr\ (B)$$

Or, to put it more simply, someone cannot belong to the set of *feminist banktellers* without also belonging to the set of *banktellers*

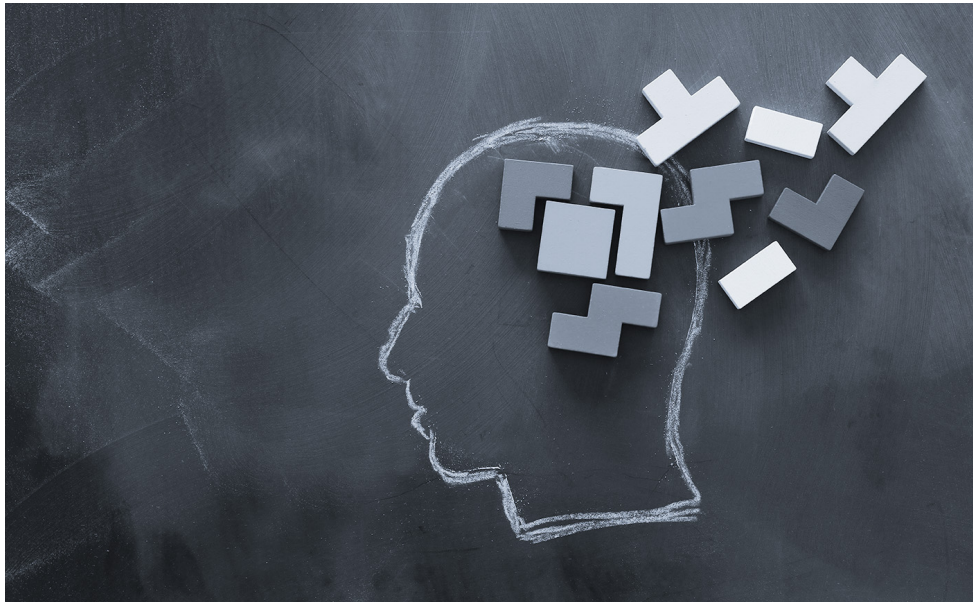


Tversky and Kahneman attributed this systematic error to what is known as the representativeness heuristic. In short, people don't think about the conjunction of events or consider probability theory when formulating an answer. Instead, their choice is based on which of the two options is most representative of the description of Linda. That is, they employ a mental shortcut (or, a heuristic) that in some instances lead to the right answer—hence, their efficiency. However, in in other cases their use lead to mistakes or errors in judgement.

A critical perspective on the view of judgement and decision-making put forward by Kahneman and Tversky would view it as an attempt to catalogue a variety of cognitive failures or irrationalities that stem from an individual's inability to perform rational calculations. However, those who adopt a view known as 'ecological rationality' argue that such a perspective judges human agents against a normative standard of rationality that is unsuitable for situated agents whose choice behaviour is constrained by myriad cognitive and environmental factors (e.g. temporal constraints that force decisions, limited information). This alternative account, made famous by Herbert Simon, and later developed by Gerd Gigerenzer reframes a lot of human judgement and decision-making as underpinned by "fast and frugal" heuristics, which are highly adaptive and support decision-making in complex and uncertain environments. It's not necessary to delve into this debate for the present purposes, but it is an interesting tangent for those interested in exploring how the choice to present statistical information in different ways (e.g., as probabilities versus frequencies) can affect comprehension and understanding.[2]

When carrying out research and innovation in data science and AI, cognitive biases can impact the processes and outcome of the project lifecycle in myriad ways. There is, after all, a large list of cognitive biases to consider. No one is expected to memorise this list as a prerequisite for responsible action. However, there are some key cognitive biases that it can be helpful to consider.

The next activity will explore a small handful of these biases, but we will look at others in more detail when we starting exploring the different stages of the project lifecycle in the next chapter. This way, we can anchor our understanding of biases—social, statistical, and cognitive—in the parts of the project lifecycle where we can hopefully mitigate their (potentially) negative impact.



Activity 4:
Bias Cards

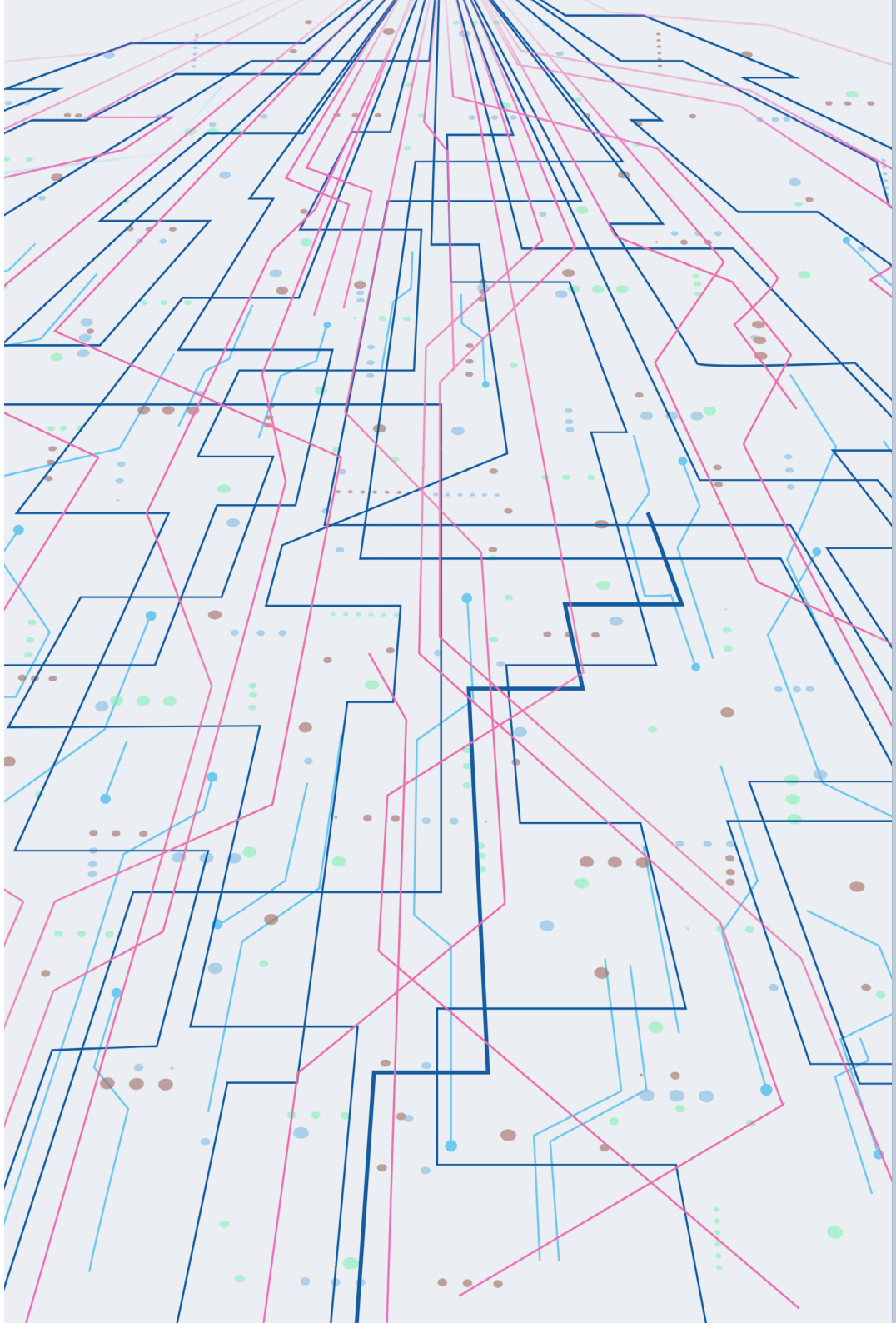
This activity introduces 9 key biases (3 for each of the different types), and explores which of the stages and activities for the project lifecycle is most impacted by their effects.

Please visit the course website to view the associated instructions.

[1] The term 'technological solutionism' is often used to refer to the belief (or, "faith") that technology can be used to solve a problem that was likely caused by technology in the first place. (Morozov, 2013)
[2 For those who want to reconstruct the debate between Kahneman, Tversky, and Gigerenzer, the following papers can be read in order: (1) (Tversky and Kahneman, 1974), (2) (Gigerenzer, 1991), (3) (Kahneman and Tversky, 1996), (4) (Gigerenzer, 1996)

04

The Project Lifecycle



Summary

In this chapter we will work our way through the key stages of the project lifecycle: (project) design, (model) development, and (system) deployment. For each stage, we will describe the accompanying activities, and highlight some of the salient ethical, social, and legal issues. However, this is presented at a relatively high-level of abstraction in the guide, because the material relies heavily on the use of case studies and accompanying activities to flesh out some of the context-specific issues. Therefore, if you are reading this chapter as part of an individual, self-directed study, you may have to adapt the activities a little.

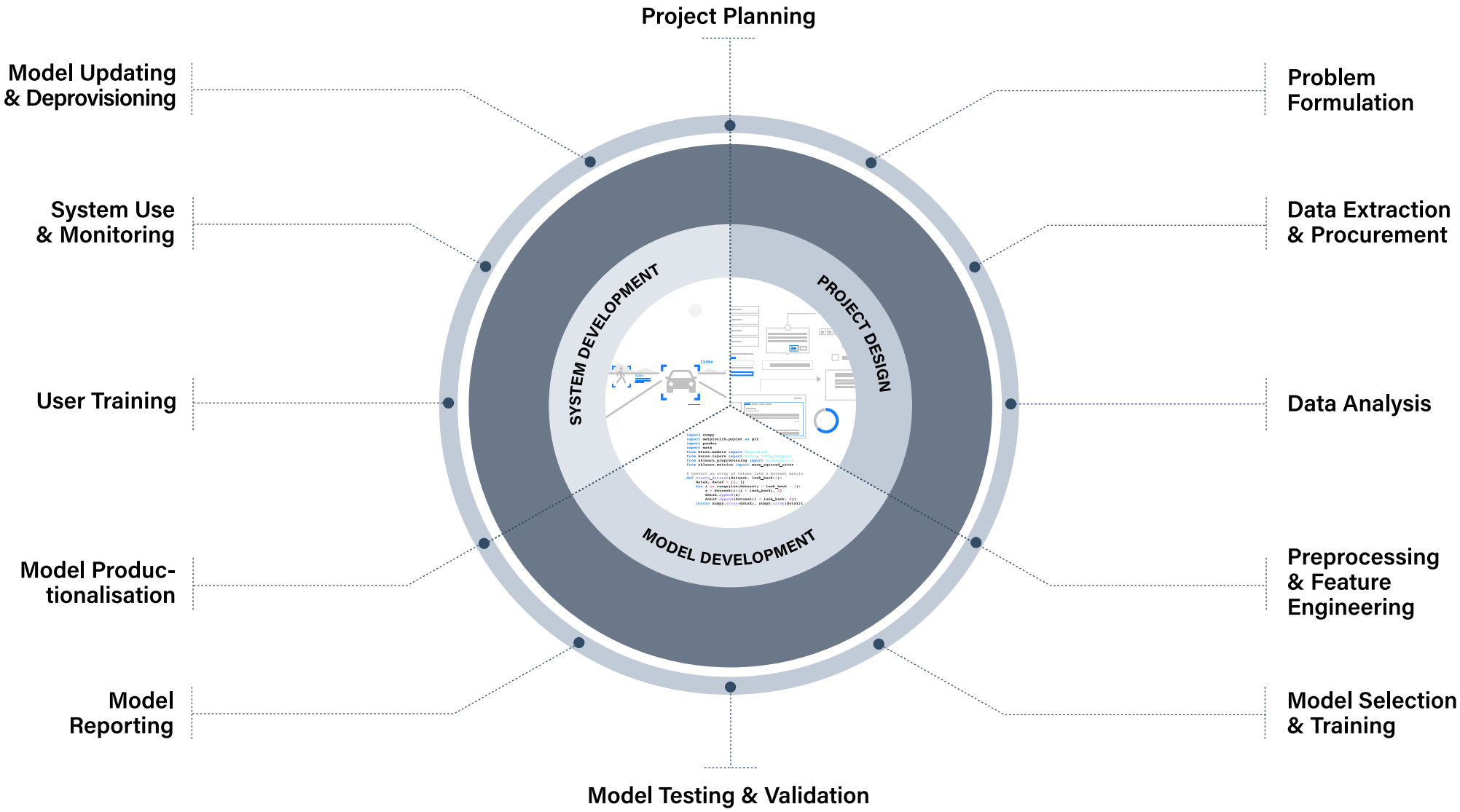
Learning Objectives

In this chapter, you will:

- Gain a high-level understanding of the central stages of the project lifecycle: (project) design, (model) development, and (system) deployment.
- Explore the activities that are associated with each of the three stages, focusing on salient ethical, social, and legal issues.
- If you are following this material as part of a live course, you will also engage in practical discussions and activities, using several illustrative case studies to help you better understand how to conduct responsible data science and AI.

PROJECT LIFECYCLE

In this chapter, we will explore the project lifecycle and the respective stages in more detail



Case Studies

The following case studies have been designed to work alongside a series of practical activities that cover the various stages of the project lifecycle. They offer only *basic* information to guide reflective and deliberative activities. If you find that you do not have sufficient information to address an issue that arises during deliberation, you should try to come up with something reasonable that fits the context of your chosen case study and then reflect on what this would mean for the project.

For example, when you reach the model selection section, you will have to consider what the benefits and risks could be for choosing a particular model over another, given the type of technology being developed. Because the focus of this course is on normative issues associated with data science and AI, we will not be training any models, so the actual outcome of choosing one model over another will require some informed speculation and reflection.

Predicting Risk of Reoffending

You are a project team responsible for developing a predictive risk assessment tool that can support sentencing decisions by judges in criminal courts. The tool will take data about a defendant and feed this into an algorithm that predicts a *risk score*, between 1 and 5 that is presented to the judge alongside additional case evidence (e.g., witness testimony). This score will represent the likelihood of reoffending, and, therefore, inform the sentencing decision made by the judge. For those that are discharged, the system will also receive feedback about whether the defendant goes on to reoffend.

Details

Category	Details
Type of technology:	Decision Support Tool
Context of use:	Sentencing Decisions in Criminal Courts
Outcome:	A Risk Score
Project team:	Data scientists working for the courts
Data types:	Age, Gender, Crime Committed, Postcode, Housing Status, Level of education, Occupation, Past offences, Feedback from police and parole officers (i.e., whether the defendant goes on to reoffend)
Data source(s):	Criminal Court Data

Recommending Courses

You work for an EdTech company and need to develop a recommender system that will be sold to schools to augment careers advice for students considering university courses. The system will ask each student to answer a series of questions, and will then provide an ordered list of recommended courses (linked to the respective university) that it "believes" are good options for the student. The system will also use satisfaction survey results and obtained degree results from those students who used the system previously as a way of calibrating and adjusting its recommendations (i.e., learning).

Details

Category	Details
Type of technology:	Recommendation System
Context of use:	Secondary Schools or Colleges
Outcome:	A ranking of possible degrees and career paths
Project team:	Private EdTech Organisation
Data types:	Courses currently being studied at school, Current grades, Gender, Age, Postcode, Extracurricular activities, Interests/Hobbies (from list of pre-selected options), Satisfaction survey results from previous users of system, University grades of previous users of system
Data source(s):	Input by Student, Gathered from Partner Universities

Classifying Hate Speech

You are a team of social data scientists employed as consultants for a social media company. You have been tasked with reducing the levels of hate speech on the company's platform by developing a classifier that can flag potential instances of hate speech for review by human moderators. The tool will automatically review every post submitted to the platform, but will only flag those that are likely to represent an instance of hate speech, based on whether they exceed some likelihood threshold. In addition to the textual content contained within the post, your tool can also use a variety of other input sources to improve its decision-making, including feedback from the human moderators that help improve the accuracy of the model over time.

Details

Category	Details
Type of technology:	Hate Speech Classifier
Context of use:	Social Media Platform
Outcome:	Binary variable ('hate speech' or 'not hate speech') with confidence rating
Project team:	Social Media Consultants and Platform
Data types:	Text content of post, Links or URLs, Network or connections of user, Tags, Images, Use of emojis, Liked communities, Stored cookies, Moderator feedback
Data source(s):	Social Media Company Data

Project Planning

FIG. 11
An illustration of automated facial recognition (by Johnny Lighthands)

In October 2021, the Financial Times reported that facial recognition cameras were being used in UK schools to scan the faces of "thousands of British pupils in school canteens" to automate the process of taking payment for lunches. The managing director of CRB Cunninghams—the company that developed the system sold to schools—claimed that "In a secondary school you have around about a 25-minute period to serve potentially 1,000 pupils. So we need fast throughput at the point of sale."

QUESTION

Does this seem like a plausible justification for the design, development, and deployment of an automated facial recognition system? Or, does the use of controversial and possibly biased technology run the risk of normalising invasive surveillance practices?

Addressing questions such as the one above should be one of the first activities in a responsible project lifecycle. Unfortunately, vested interests often prevent them from being discussed in an open and responsible manner. The result can be the treatment of data-driven technologies as a "hammer" with which to go looking for nails!

To prevent this, it is best to have a clear idea in mind of what the project's goals are at the outset, and what problem is being addressed. This can help to avoid a myopic focus on a narrow class of technology-based "solutions", and also helps create space for a diversity of approaches—some of which may not require data-driven technology at all.

Project planning, therefore, can comprise a wide variety of tasks, including, but not limited to:

- an assessment of whether developing or using data-driven technology is the right approach given available resources and data, existing technologies and processes already in place, the complexity of the use-contexts involved, and the nature of the policy or social problem that needs to be solved (Leslie et al 2021a);
- a discussion with an *ethics committee* or *internal review board* to help evaluate the ethical credentials of the project;
- an analysis of *user needs* in relation to the prospective model and whether a solution involving the latter provides appropriate affordances in keeping with user needs and related functional desiderata;
- identification and mapping of key stages in the project to support project governance and business tasks (e.g. scenario planning);
- an assessment of *resources and capabilities* within a team, which is necessary for identifying any *skills gaps*;
- a contextual assessment of the target domain and of the expectations, norms, and requirements that derive therefrom;
- a *stakeholder impact assessment*, supported by affected people and communities, to identify and evaluate possible harms and benefits associated with the project (e.g. socioeconomic inequalities that may be exacerbated as a result of carrying out the project), to gain social license and public trust, and also feed into the process of problem formulation in the next stage;
- an analysis of *team positionality* to determine the appropriate level and scope of community engagement activities (Leslie et al 2021b);
- wider *impact assessments*—both where required by statute and done voluntarily for transparency and best practice (e.g. equality impact assessments, data protection impact assessments, human rights impact assessment, bias assessment).

This can be a lot to undertake at the outset of a project, but the upfront cost can help offset the large technical and ethical debt that may otherwise accumulate from a failure to anticipate or foresee possible challenges. Many of the above activities have been tried and tested in a wide variety of domains, and there are many templates or frameworks that can be used to help speed up the process (see Further Resources).

Activity 6:
Privilege Walk

This is the first activity that will involve the case studies introduced at the start of the chapter. It is a known as a "privilege walk" and is designed to help you reflect on how disparate forms of social privilege interact and how different harms or benefits associated with your research or innovation project may have disparate impacts on affected stakeholders.

Please visit the course website to view the associated instructions.

Problem Formulation

In the previous section we acknowledge that it is important to clearly define and delineate the problem that your project is intended to address. Here, 'problem' is used in two interrelated ways:

1. **To refer to a *well-defined computational process* (or a higher-level abstraction of the process) that is carried out by the algorithm to map inputs to outputs.**

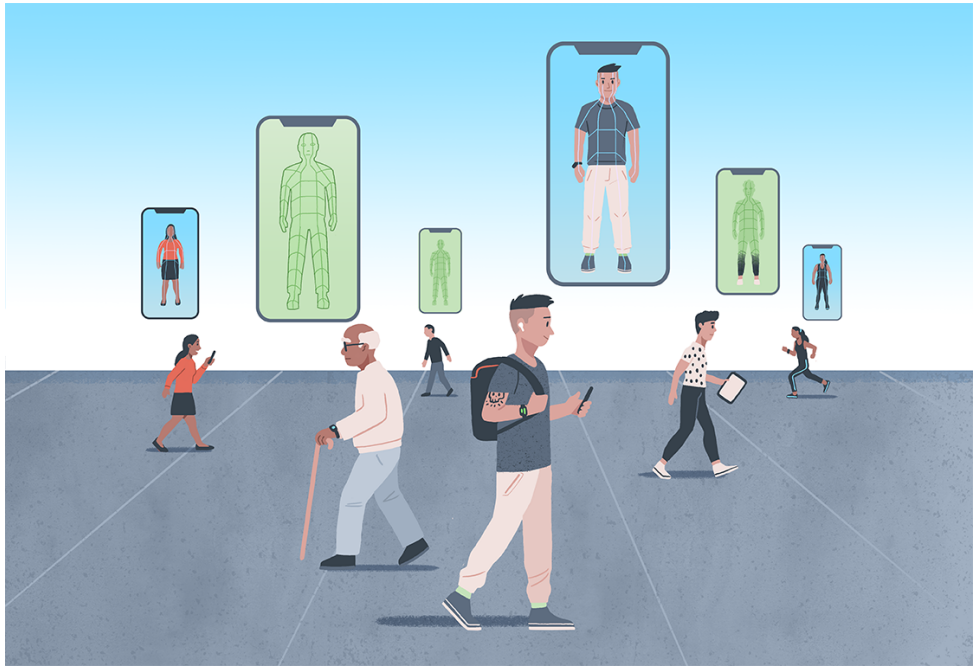
2. **To refer to the wider practical, social, or policy issue that will be addressed through the translation of that issue into the aforementioned mathematical or computational framing.**

On the computational side, we can think of how a convolutional neural network carries out a series of successive transformations by taking (as input) an image, encoded as an array, in order to produce (as output) a decision about whether some object is present in the image. On the practical, social, and policy side, there will be a need to define the computational "problem" being solved in terms of the algorithmic system's embeddedness in the social environment and to explain how it contributes to (or affects) the wider sociotechnical issue being considered. In the convolutional neural network example, the system being produced may be a facial recognition technology that responds to a perceived need for the biometric identification of criminal suspects by matching facial images in a police database. The social issue of wanting to identify suspects is, in this case, translated into the computational mechanism of the computer vision system. But, beyond this, diligent consideration of the practical, social, or policy issue being addressed by the system will also trigger, *inter alia*, reflection on the complex intersection of potential algorithmic bias, the cascading effects of sociohistorical patterns of racism and discrimination, wider societal and community impacts, and the potential effects of the use of the model on the actors in the criminal justice systems who will become implementers and subjects of the technology.

Sociotechnical considerations are also important for determining and evaluating the choice of target variables used by the algorithm, which may ultimately be implemented within a larger automated decision-making system (e.g. in a risk prediction system). The task of formulating the problem allows the project team to get clear on what input data will be needed, for what purpose, and whether there exists any representational issues in, for example, how the target variables are defined. It also allows for a project team (and impacted stakeholders) to reflect on the reasonableness of the measurable proxy that is used as a mathematical expression of the target variable, for instance, whether moving a patient into an intensive care unit (ICU) is a reasonable action to take after being classified as "at risk" by a model that uses a set of demographic and physiological input variables.

The fact that formulating problems and defining target variables in ML/AI innovation lifecycles can often be a biased and contested process is why stakeholder engagement, which helps bring a diversity of perspectives to project design, is so vital. It is also why this stage is so closely connected with the interpretive burdens of the project planning activities seen in the previous section (e.g. discussion about legal and ethical concerns regarding permissible uses of personal or sensitive information).

Data Extraction and Procurement



Ideally, the project team should have a clear idea in mind (from the planning and problem formulation stages) of what data are needed prior to collection, extraction, or procurement. This can help mitigate risks associated with over-collection of data (e.g. increased privacy or security concerns) and help align the project with values such as *data minimisation* [ICO and Institute, 2020]. Of course, this stage may need to be revisited after carrying out subsequent tasks (e.g. data analysis, preprocessing, model testing) if it is clear that insufficient or imbalanced data were collected to achieve the project's goals.

Where data is procured, questions about provenance arise (e.g. legal issues, concerns about informed consent of human data subjects). For instance, what information about the dataset is necessary to provide sufficient assurance to the team that they are procuring data that has been reliably collected by another party. Or, will they be able to reuse and repurpose the data for their intended project. The procured data will need to be sufficiently representative of the intended target domain of the project if it is to be useful.

Generally, addressing these issues and ensuring responsible data extraction and procurement requires the incorporation of myriad forms of expertise into decision-making. This can include,

- legal expertise (e.g., data protection officer) who is able to inform the project team of the relevant lawful bases for data collection and help;
- ethical expertise (e.g., whether the various rights and freedoms of data subjects have been properly respected);
- domain expertise (e.g., whether the method of original data collection, the expected quantity of data, and the variety of features, will be sufficient based on the problem being addressed, as formulated in the previous stage); and
- personal expertise (e.g., whether the data subjects are likely to be willing to provide access to all the data being requested)

The FAIR Principles

At this stage in a research project, it is also important to address the long-term sustainability of your work. One component of research sustainability is the active support of reproducibility of research. To this end, the FAIR guiding principles for scientific data management and stewardship were developed, as a means to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of research data and digital assets.

In summary, data should be,

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.

Accessible

Once the user finds the required data, she/he/they need to know how can they be accessed, possibly including authentication and authorisation.

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

We won't delve into these principles in any further detail, as it is beyond the scope of this course. However, there is more information on the Go-FAIR website and a great guide over on the Turing Way [Community *et al.*, 2019].

Data analysis

Exploratory data analysis is a crucial stage in the project lifecycle. It is where a number of techniques are employed for the purpose of gaining a better understanding of the dataset and any relationships that exist between the relevant variables. Among other things, this could mean,

- Describing the dataset and important variables
- Cleaning the dataset
- Identifying missing data and outliers, and deciding how to handle them
- Provisional analysis of any relationships between variables
- Uncovering possible limitations of the dataset (e.g. class imbalances) that could affect the project

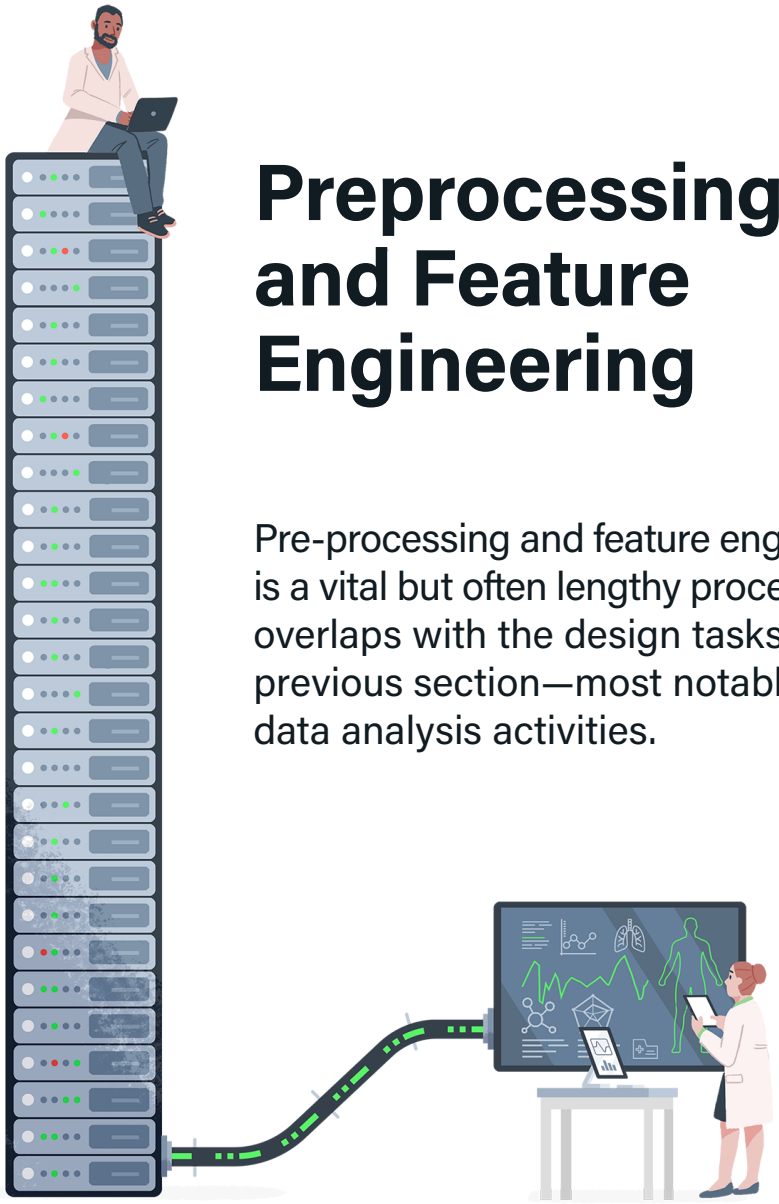
In the online version of this guidebook we cover each of these sub-stages from a bias-aware perspective, using Jupyter notebooks—a popular tool in data science—to aid our exploratory data analysis by visualising some data.

Please head over to https://bit.ly/3nF5L7h for further details for this section.

Activity 7:
Developing Case Studies

In this activity we will further develop the case studies that were selected for the previous activity through group discussion.

Please visit the course website to view the associated instructions.



Preprocessing and Feature Engineering

Pre-processing and feature engineering is a vital but often lengthy process, which overlaps with the design tasks in the previous section—most notably the data analysis activities.

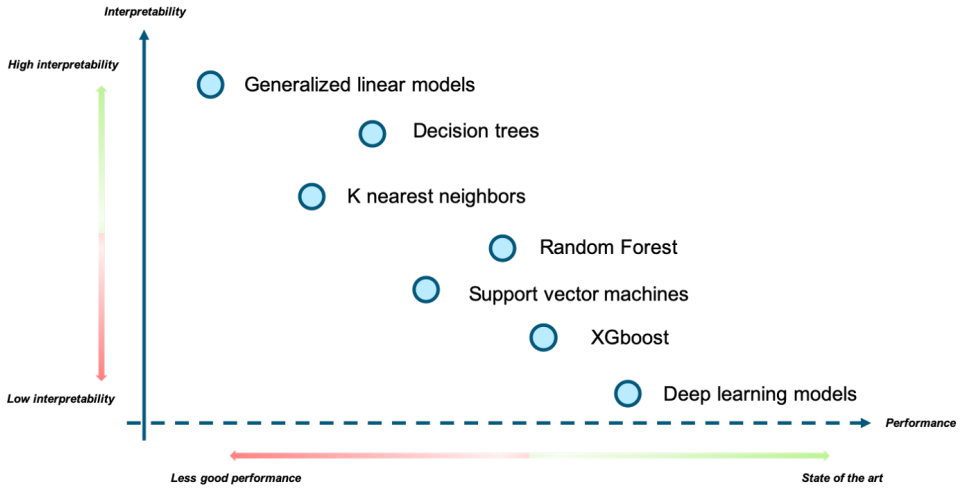
It also shares with them the potential for human choices to introduce biases and discriminatory patterns into the ML/AI workflow. Tasks at this stage include (additional) data cleaning, data wrangling or normalisation, and data reduction or augmentation. Whereas data analysis is oriented towards a provisional understanding the dataset (e.g., analysing possible relationships between variables), preprocessing is directed towards the model development tasks.

It is well understood that the methods employed for each of these tasks can have a significant impact on the model's performance (e.g. removing rows or columns can affect the predictive power of the model). As Ashmore et al. [Ashmore *et al.*, 2019] note, there are also various desiderata that motivate the tasks, such as the need to ensure the dataset that will feed into the subsequent stages is relevant, complete, balanced, and accurate.

Furthermore, at this stage, human decisions about how to group or disaggregate input features (e.g. how to carve up categories of gender or ethnic groups) or about which input features to exclude altogether (e.g. leaving out deprivation indicators in a predictive model for clinical diagnostics) can have significant downstream influences on the fairness and equity of an ML/AI system.

Model Selection

This stage determines the model type and structure that will be produced in the next stages. In some projects, model selection will result in multiple models for the purpose of comparison based on some performance metric (e.g. accuracy). In other projects, there may be a need to first of all implement a pre-existing set of formal models into code. The class of relevant learning algorithms used to train the model is likely to have been highly constrained by many of the previous stages (e.g. available resources and skills, problem formulation). For example, where the problem demands a supervised learning algorithm, instead of an unsupervised learning algorithm, to help develop a model that can predict likely values for future instances not contained within the original dataset.



Interpretability

While accuracy or predictive power may be typical goals that motivate the selection of specific learning algorithms and models (see below), there are also additional factors that can influence decision-making. One notable factor is the inherent interpretability of the model.

Certain learning algorithms produce models that are inherently less interpretable than others. For instance, a linear regression model is easy to understand because of the straightforward connection between input variables and the learned weights that alter how much influence the individual variables have on the outputs. However, a neural network, at the other end of the extreme is often described as a "black box model" because the relationship between the input features and the output is often too difficult to interpret without the use of ad hoc methods [ICO and Institute, 2020]. The trade-off for this lower interpretability can be greater performance in terms of accuracy or predictive power.

This trade-off has important normative considerations though. For instance, consider the decision to deploy an algorithmic decision support system in criminal courts to help a judge decide on a sentence. A more accurate model could reduce the number of unfair decisions (e.g., someone being given a prison sentence rather than community service), but the judge may not be able to understand why a

FIG. 12
A schematic showing the trade-off between model interpretability and model performance. Reprinted from [Diop, 2019]

particular decision is recommend by the model and thus be unable to explain their decision to the defendant. As transparency and accessibility are vital parts of judicial decision-making and the rule of law, the use of a black-box model, in spite of the greater accuracy, may be deemed unlawful and unjust [Bingham, 2011]. The trade-off is in almost all non-trivial cases, unavoidable. Therefore, such a decision is inescapably value-laden and inherently about exercising ethical reflection and responsible deliberation—likely in conjunction with affected stakeholders.

Some Common Algorithms

Supervised Learning

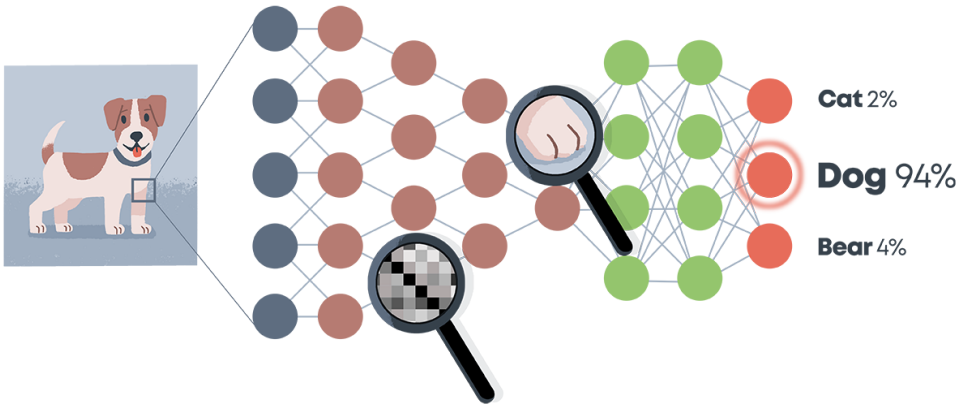


FIG. 13
An illustration of a neural network classifying an image of a dog correctly (by Johnny Lighthands).

Supervised learning involves training a model using a set of examples, which are pairs of input data and corresponding labels. For example, learning to classify labelled images as pictures of 'cats' or 'dogs', in order to then classify new (unlabelled) instances. Formally, the algorithm takes a set of n pairs, $(x_1, y_1),..., (x_n,y_n)$, where x_i is the feature vector of the i-th example and y_i is its label. The task of the learning algorithm is to learn a function that maps members of the input space X onto the output space Y.

Supervised learning algorithms can be used to perform classification or regression tasks. Commonly used learning algorithms include:

- Linear Regression
- Logistic Regression
- Naive Bayes
- Decision Trees
- Support-Vector Machines
- Neural Networks

Unsupervised Learning

In contrast, unsupervised learning algorithms try to find patterns in unlabelled data, typically by clustering the data into similar groups or reducing the dimensionality of the variables (or, features) under consideration. For example, an algorithm may try to cluster shoppers into groups based on their purchasing habits. A human would then need to interpret the meaning behind this clustering.

Commonly used unsupervised learning algorithms include:

- K-Means Clustering
- Hierarchical Clustering
- Support Vector Clustering
- Principal Component Analysis

—Reinforcement Learning
Finally, we have reinforcement learning (RL). RL algorithms try to learn an optimal policy that has the goal of maximising some value function when interacting within a particular environment. For example, an intelligent agent that has the goal of scoring the highest number of points in a video game by learning what actions to perform in response to visual feedback from a screen.

RL algorithms can be split into *model-free* or *model-based* methods, where the latter tries to build a model of its environment on which to choose the optimal policy. [AI, 2019]

Activity 8:
Identifying Missing data

In this activity you will try to identify likely sources of missing data for your hypothetical projects, and consider the related biases that could result from them.

Please visit the course website to view the associated instructions.

Model Training, Testing and Validation

Before training a model, the data need to be split into 'training' and 'testing' sets to avoid model overfitting.[^overfitting^]

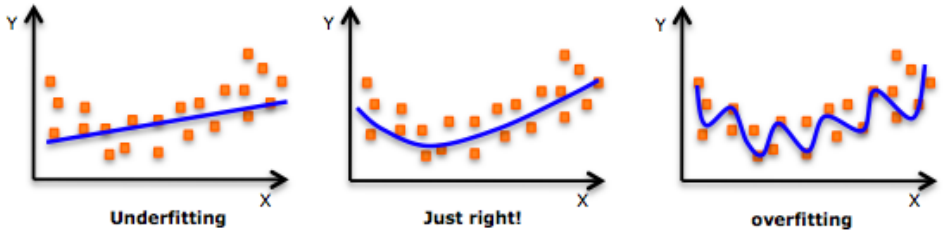


FIG. 14
An example of overfitting, underfitting and the appropriate balance [Bronshtein, 2020]

The *training* set is the one used, as the name suggests, to train the model, whereas the *testing* set is a hold-out sample that is used to evaluate the fit of the ML model to the underlying data distribution. The testing set is kept separate while training the model to provide an less biased evaluation of the model once it has been fit to the training dataset.

The human input at this stage is about deciding on the training-testing split and about how this shapes desiderata for model validation—a subsequent process where the model is validated either internally or in wholly new environments (i.e. external validation). As such, the decision can be very consequential for the trustworthiness and reasonableness of the development phase of an ML/AI system. For instance, what if the training/testing split is not random? What if, hypothetically, the training data contain only examples from one class of objects and the testing data contain instances of an entirely different class? We would be unlikely to get a very useful model.

There are various methods to help reduce the chance of this issue occurring, which are widely available in popular package libraries (e.g. the scikit-learn library for the Python programming language). However, a common method is to use something known as 'cross validation.'

One of the most popular forms of cross validation is K-Folds Cross Validation. Here, the dataset is first split into training and testing sets, and then the training set is further split into *k* different subsets (e.g. 10 subsets). The model training process then occurs *k* times, using a different subset as the validation set on each round. At the end, an average is taken from the *k* models and this is tested against the original hold out testing set. The following graphic should help you visualise this.

This type of validation is also known as 'internal validation,' to distinguish it from external validation, and, in a similar way to choices made about the original training-testing split, the manner in which it is approached can have critical consequences for how the performance of a system is measured against the real-world conditions that it will face when operating "in the wild."

Therefore, external validation can also be performed, either using entirely novel datasets—perhaps from different sites or using different collection methods. Or, separate external teams may even be able to externally validate a research model by attempting to reproduce similar results. To support this it is vital to ensure that the steps taken during these stages are properly documented and reported, as we will see in the next section.

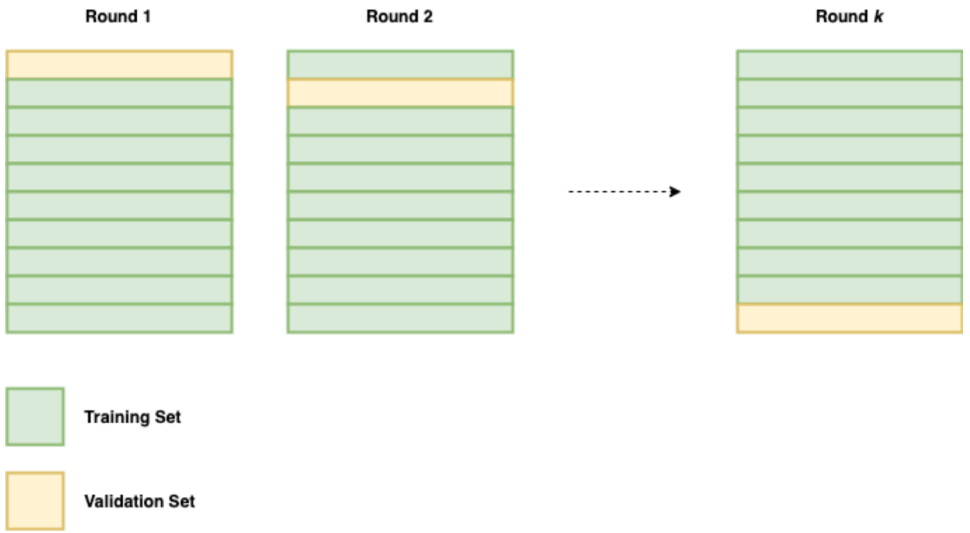


FIG. 15
A simple representation of the K-Folds Cross Validation Process.

Model Reporting

Over the course of the previous activites, your project team will have created many diverse artefacts and forms of documentation. For example, during project planning you may have created a series of risk and impact assessments, or during exploratory data analysis you will have developed a set of notebooks explaining how you imported, cleaned, and analysed your data. In some cases, the artefacts will be byproducts (e.g., system logs). In other cases, they will be the specific goal of the associated activity. Model reporting is an example of this latter set.

In short, model reporting is a process of developing and integrating documentation and evidence about the processes by which your model was designed and developed (e.g., trained, tested, and evaluated), as well as how it ought to be used or deployed.

What information you need to include is going to vary based on the type of project you are undertaking, and the intended use context. For example, a model developed by a private company that is sold to the NHS in England for the purpose of supporting radiologists when carrying out diagnosis in a hospital will need to be more thoroughly documented than a NLP model used as in a simple chatbot UI to support customer queries on an e-commerce website.

In general, however, a model report is likely to include information about the data (e.g., size, source, method of collection, any sensitive attributes), the datasets used to train, test, and validate the model (e.g. training-testing distributions), the performance measures selected for evaluating the model (e.g. decision thresholds for classifiers, accuracy metrics), the intended use of the model, and any legal or ethical considerations associated with the model's use (e.g. fairness constraints, use of politically sensitive demographic features).

There are several templates (and tools) that exist to help researchers and developers identify what information ought to be documented, but there are typically advantages and disadvantages associated with each of them. To give just a few examples:

- Datasheets for Datasets
- TRIPOD statement
- Model Cards for Model Reporting
- Factsheets

When evaluating whether to use or repurposes tools such as these, it is important to consider the context in which you are developing. A model produced within the context of a publicly-funded research project may have certain disclosure obligations that are not mandatory for models developed in a commercial context. Alternatively, it may just be best practice to support reproducibility by considering how your model report adheres to principles such as the FAIR guidelines.

In the next activity, we will reflect on the scope and content of some hypothetical model reports and grapple with ethical questions, such as the trade-off between transparency and privacy.

Activity 9:
Designing Model Reports

In this activity you will design hypothetical model reports for the case studies that you have been considering throughout the chapter.

Please visit the course website to view the associated instructions.

Model Productionalisation

Unless the end result of the project is the model itself, which is perhaps more common in scientific research, it is likely that the model will need to be implemented within a larger system. This process, sometimes known as 'model operationalisation', requires understanding (a) how the model is intended to function in the proximate system (e.g. within an agricultural decision support system used to predict crop yield and quality) and (b) how the model will impact—and be impacted by—the functioning of the wider sociotechnical environment that the tool is embedded within (e.g. a decision support tool used in healthcare for patient triaging that may exacerbate existing health inequalities within the wider community). Ensuring the model works within the proximate system can be a complex programming and software engineering task, especially if it is expected that the model will be updated continuously in its runtime environment. But, more importantly, understanding how to ensure the model's sustainability given its embeddedness in complex and changing sociotechnical environments requires active and contextually-informed monitoring, situational awareness, and vigilant responsiveness.

As noted, this stage of the product lifecycle involves often complex forms of software engineering, which in many cases are just about ensuring appropriate software dependencies and packages are installed that allow users of the system to access or interface with the model as intended (e.g. building and coding an API). This is not to deny that there are important ethical issues involved with model productionalisation, especially with how an API either enables inhibits its access to the benefits of a model. However, we can consider these issues elsewhere (e.g. during 'project planning' or 'user training' and in a way that doesn't presuppose familiarity with technical concepts. Therefore, we will keep this section as a brief note.User

User Training

Consider the following scenario:

Example

You are in charge of deploying an automated facial recognition system that is used to verify that people who are attempting to enter a secure building are using the appropriate identity card. Upon swiping their identity card, the system scans the face of the user and matches it to the expected image from a database of authorised people. If the person matches their card and is also allowed access to the building, they are automatically granted access.

After a week or so of deploying the system, you find out that the security guard in charge of the building has been overriding the system. You go to speak with the security guard, and he claims that the system has been refusing entry to people who clearly match their identity badge.

> When you investigate the issue, you find out that the system is functioning as expected and that no errors with the automated facial recognition system have been logged. However, every one of the attempts that the security guard has overridden are for people with expired identity cards. Although they match their cards, they should not have access to the building.

This scenario highlights an important, but sometimes overlooked part of system deployment and evaluation: **human factors**.

> Human factors is a field in which researchers and practitioners are interested in both understanding the interaction of people and technology and in making that interaction more efficient, safer, and pleasant. [Durso *et al.*, 2014]

Research into human factors considers both perceptual, cognitive, social, and physical characteristics of the human operator, as well as how a technological system has been designed. This is important, because if there is an issue, such as the one in our above scenario, the *sociotechnical* system can be improved by changing either a) the human or b) the technology.

For instance, in our scenario, you could provide training for the security guard that shows him how to identify the cause of an automated rejection. Or, you could design a simple prompt or notification that explains why an individual is being denied entry to avoid the possibility of the security guard overriding the system unless there is a legitimate false negative (i.e. an error with the facial recognition system for a valid identity card and matched person).

There is a huge amount of quantitative and qualitative research related to human factors in general (see [Durso *et al.*, 2014]), and a growing source of research for human interaction with algorithmic systems more specifically.

For present purposes, it is sufficient to note that although the performance of a model is evaluated in earlier stages, the model's **impact** cannot be fully evaluated without consideration of the human factors that affect its performance in real-world settings. For instance, the impact of human cognitive biases, such as *algorithmic aversion* [^aversion] must also be considered, as these biases can lead to over- and under-reliance on the model (or system), in turn negating any potential benefits that may arise from its use. Understanding the social and environmental context is also vital, as sociocultural norms may contribute to how training is received, and how the system itself is evaluated (see [Burton *et al.*, 2020]).

In the context of RRI, user training related to how an algorithmic system should be operated may include: a) conveying basic knowledge about the nature of machine learning (e.g. probabilistic results or outcomes), b) explaining the limitations of the system, c) educating users about the risks of AI-related biases, such as decision-automation bias or automation-distrust bias, and d) encouraging users to view the benefits and risks of deploying these systems in terms of their role in helping humans to come to judgements, rather than replacing that judgement.

System Use and Monitoring

The Ancient Greek philosopher, Heraclitus believed that all things are in a constant state of change or flux—a doctrine made famous by the following statement:

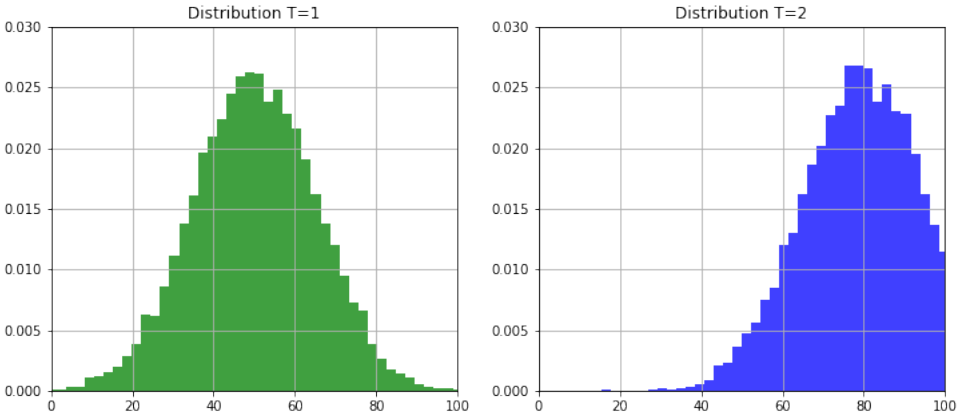
You cannot step twice into the same rivers; for fresh waters are ever flowing in upon you.

—Heraclitus, Fragment 12

Metaphysics aside, it is certainly true that our world is dynamic and that many things change over time. As a result of this, and depending on the context of deployment, the performance of a model that aims to predict, classify, or recommend, is likely to degrade as the populations or objects it represents change or evolve. This process of *model drift* can be said to occur when there is increasing variation between how representative the training dataset was at the time of development and how representative it is at later stages.

There are many context- and domain-specific causes of model drift. However, we can consider two main types.
On the one hand, drift can occur when the statistical properties of an input variable change (i.e. there is a shift in the underlying data distribution). For example, perhaps house prices start increasing and a model becomes more and more inaccurate at predicting them. [1]

The following offers a graphical representation of this problem, with a distributional shift between two different points in time (T=1, T=2).



[1] Sometimes this type of drift may not be because of any underlying change in the properties of the relevant variable, but because of issues with the original data collection in the first place (e.g. selection bias).

On the other hand, there could be a more nuanced reason related to the conceptual or social meaning of the input variables. An example of this could be a machine learning algorithm used in finance that aims to predict whether someone is likely to default on a loan using variables with social meaning, such as occupation. If the model detects a relationship between specific values for occupation and the employee's ability to pay back a loan in a timely manner, the system may recommend more loans to people in this occupation. However, if something happens that has a global impact on these jobs (e.g. increased investment in the sector creating a rise in average wages), this association will change. The result is that people who could otherwise afford a loan may still be denied one due to inaccurate models.

These two examples are an important reason why ongoing monitoring of a system is so important. How model drift is handled, however, is dealt with in the next section.

Model Updating or Deprovisioning

We come now to the end of the project lifecycle!

In the previous section we saw that model drift can affect the accuracy and overall performance of a model. One way to address this is to update the model by retraining on more timely and up-to-date data.

Model updating can occur continuously if the architecture of the system and context of its use allows for it. This can help prevent model drift from occurring in the first place. However, this type of online learning is a challenging task, and is not without its limitations. For instance, it introduces a further source of uncertainty into how the model and system will perform, given their close coupling with the environment.

Alternatively, model updating can occur periodically. Perhaps, there is known seasonal variation in the performance of a system (e.g. a recommender system for an e-commerce site that has to adjust for varied shopping patterns). As such, the re-training of the model may be planned around these times to help maintain a high-level of performance.

These types of updating can use the original model as a starting point, in order to just retune the model's parameters or maybe drop certain features that are no longer predictive. However, there is also the option of entirely de-provisioning (i.e., stopping use) the model and system if performance simply drops too low to be addressed by mere re-training.

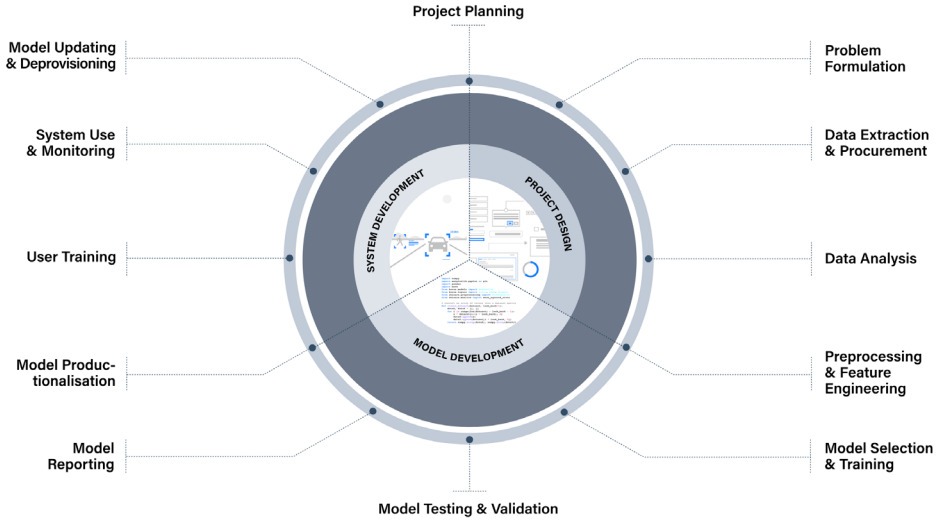


FIG. 16
The Project Lifecycle.

Full Circle

As you may recall, the project lifecycle image showed this final stage connected to the first stage of 'Project Planning' in a circular manner.

The reason for this should now be clearer. Depending on the choices made at this stage, it may be necessary to start planning for a new project. For instance, a project team may not be able to simply *remove* a system that serves a business critical function. Instead, an existing project may serve as a foundational input or constraint into the planning stages of a new project—starting the cycle once more.

Next Steps

The overview and summary of the project lifecycle presented in this chapter, by necessity, skips over a lot of detail—both practical and theoretical. However, the framework has been designed with 'sustainability' in mind, and so it should be able to accommodate subsequent additions. For instance, perhaps there is a method of bias mitigation or explainability that has not been considered. Hopefully, the framework as it stands should be able to integrate this into the unifying reflective and deliberative procedure without needing to alter any of the overarching considerations.

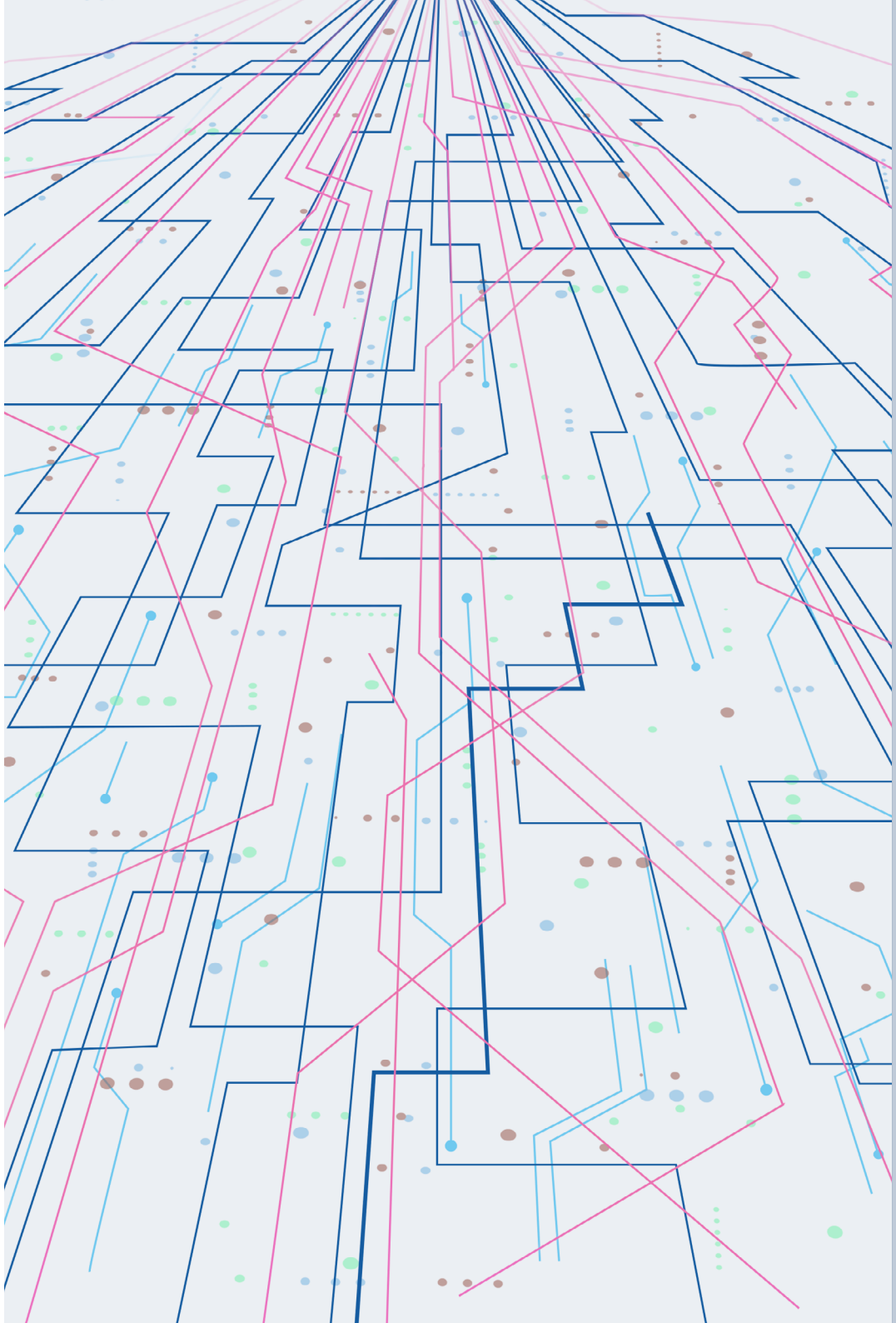
This remains to be seen. For the time being though, the project lifecycle framework provides a useful anchor for subsequent discussion, and serves to motivate the following question:

> How do you **responsibly communicate** the actions and decisions undertaken throughout the project lifecycle stages to an audience with diverse needs and expectations?

This can be challenging, because there may be a range of ethical values that are implicated in your project, which in turn affect a plurality of goals and needs for the various stakeholder groups, including ensuring that the system being deployed is safe, secure, fair, trustworthy, explainable, sustainable, or respectful of human agency and autonomy. How do you provide assurance that the interconnected project processes and activities individually and collectively support the relevant goals? This is the topic of our final chapter.

05

Responsible Communication



Summary

In this chapter we will look at what happens when a project reaches the stage where it is necessary to communicate research findings or present the output of the innovation lifecycle to a broader audience. We will consider a method known as argument-based assurance, which has been designed to help developers and project members engage their audience in a trustworthy and transparent manner.

Learning Objectives

In this chapter, you will:

- Consider the basics of the argument-based assurance methodology
- Understand when and how it can be used to facilitate responsible communication
- Use the method to identify broader normative goals that may not have been covered in this course, and determine which properties need to be assured to help demonstrate that the respective goal has been obtained

Engaging, Communicating, Assuring

In this final chapter, we will consider what it means to communicate the processes and outputs of your research or innovation project in a responsible manner. Although we have already looked at several relevant mechanisms for communication, such as stakeholder engagement or model reporting, these mechanisms are only two pieces within a larger puzzle. They play a significant role in supporting the early parts of a project's lifecycle, but some aspects of responsible communication go beyond any single stage or even the overall lifespan of a project.

Take a typical scientific research project. It is quite common for a research project to publish findings and then receive ongoing public interest for some time after the scope of the project itself. This may come in the form of engagement with policy-makers, journalists, other researchers, commerical organisations looking to apply the findings, or perhaps even citizen interest groups. Each of these groups will have a different objective or rationale that motivates their interest and engagement with a project's research process or findings. It is important to take these factors into consideration, as they can have a big impact on how a research or innovation project will be perceived.

There is, of course, a limit on how much responsibility (and accountability) a project team can take for the use (or misuse) of their research by others. For example, if a commercial organisation ignores a project team's clear advice and thorough documentation about the applicability of a model to a novel context, the project team should not be blamed. But this depends on whether the original communication and engagement was exercised in a responsible manner. The commercial organisation, for instance, should not be entirely blamed if the project team gave misleading or incomplete details about their work.

It can be difficult to identify when sufficient responsibility has been exercised though. After all, there is a big space between, on the one hand, publishing scientific findings in the public sphere and then assuming an entirely *laissez faire* approach to how the research is used and interpreted, and on the other hand, attempting to monitor every use or application of the findings to ensure that no harms arise. In this chapter, however, we will seek to provide an intuitive (albeit partial) method (or, procedure) to help address this challenge of delineating or demarcating the scope and substance of a project team's responsibility for ongoing communication and engagement.

Additional Guides

Before we dicuss the method, however, we should note that there is significant overlap here with the content of our two other courses:

1. Public Engagement and Communication of Science (PES)

2. AI Ethics and Governance (AEG)

The first of these two courses (PES) goes into more detail about the scientific and ethical rationale for public engagement, and also looks at how novel technologies are shaping the communication of science. The second course explores the normative foundations of public discourse, drawing upon work in discourse ethics and jurisprudence and extending their central themes to help address challenges of governing AI technologies that operate in a global and intercultural context.

The topics and themes explored in these courses develop the notion of 'responsible communication' in important ways. And, arguably, an understanding of 'responsible communication' would be incomplete without reflecting on some of the conceptual issues that the above two courses consider. However, there is a limit to how much can be covered in any single course, so we won't delve into these issues here. We encourage you to look into these courses if you're interested.

We will end this guide, therefore, by exploring a unifying framework that helps draw together the themes and activities we have already explored in this course, and which also provides a foundation for responsible communication. Unlike the topics and scope of the other courses, our present concern has a more (direct) pragmatic goal: the development of an assurance case that helps communicate and justify that you have exercised responsibility throughout a research or innovation project.

What is Argument-Based Assurance?

Reference

This section is based on [Burr and Leslie, 2021], which provides a more thorough account of the argument-based assurance methodology and how it applies to responsible research and innovation in data science and AI.

The method we will explore that serves the role of facilitating responsible communication is known as argument-based assurance (ABA). We can define ABA as:

> a process of using structured argumentation to provide assurance to another party (or parties) that a particular claim (or set of related claims) about a property of a system is warranted given the available evidence.

As a structured method for communication, ABA is already widely used in safety-critical domains or industries where manufacturing and development processes are required to comply with strict regulatory standards and support industry-recognised best practices [Hawkins *et al.*, 2021].

However, ABA is useful for more than just demonstrating regulatory compliance. It can also

- assist internal **reflection** and **deliberation** by providing a systematic and structured means for evaluating how the development of systems or products can fulfil certain normative goals (e.g. safety or robustness), according to certain well-defined properties (e.g. software hazards identified) and criteria (e.g. risk reduction thresholds met)
- provide a deliberate means for the **anticipation** and **pre-emption** of potential risks and adverse impacts through mechanisms of end-to-end assessment and redress;
- facilitate **transparent communication** between developers and affected stakeholders;
- support mechanisms and processes of **documentation** (or, reporting) to **ensure accountability** (e.g. audits, compliance);
- and build **trust and confidence** by promoting the adoption of best practices (e.g. standards for warranted evidence) and by conveying the integration of these into design, development, and deployment lifecycles to impacted stakeholders.

For these reasons, it is a useful foundation upon which to build a framework for responsible communication.

Assurance Cases

When a barrister stands in a court room, in front of the judge, jury, and defendant, they are tasked with presenting a case. If they are part of the prosecution, their role is to convince the jury, beyond all reasonable doubt, that the defendant is guilty of committing the crimes of which they stand accused, based on the admissible evidence agreed upon by all parties. The case they present, therefore, is an argument that attempts to justify their position or standpoint. Although the format and goals may be different, the purpose of **argument-based** assurance is also to develop and present a case. This is not an argument in the antagonistic sense of the term, but rather a structured and justifiable case supported by evidence.

The scope and content of what we can call an 'assurance case' is determined by the relevant details of the project in question, and what the project team need to provide assurance for. For example, if a project team needs to communicate the processes by which they have ensured the *interpretability* of their model, they may need to develop an *interpretability* case, which could look something like the following:

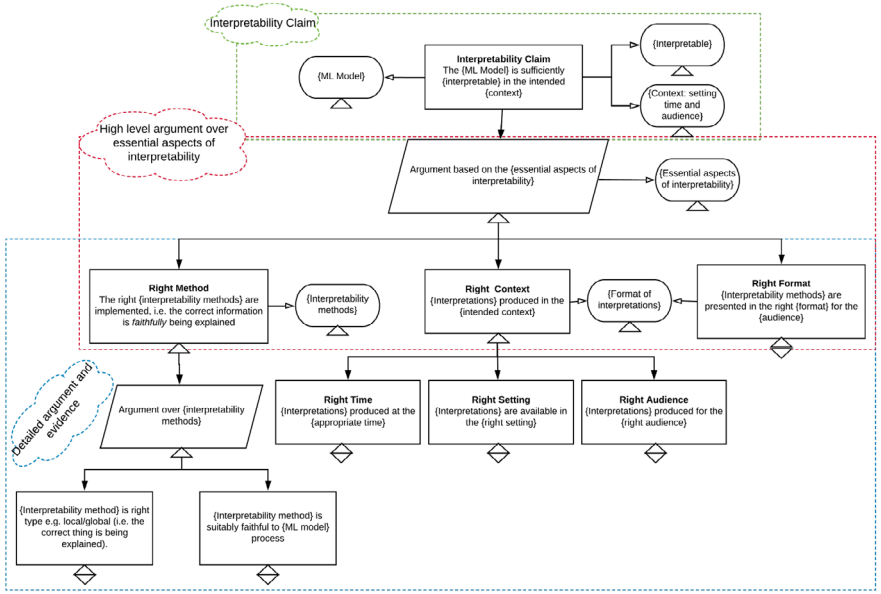


FIG. 17
A template of an assurance case that focuses on providing assurance for the interpretability of a machine learning model. Reprinted from [Ward and Habli, 2020]

Here, the assurance case is represented in a graphical format. The top-level claim is a goal that is supported by the lower level claims, which in turn further specify what it means to say "The {ML model} is sufficiently {interpretable} in the intended {context}". At the lowest level is the evidence that supports and establishes the relevant basis for making the specific claims. Overall, the case is a *structured argument* that is oriented towards the top-level goal.

This is achieved by first reflecting on what the goal claim means. For instance, what are the parameters of *sufficiently interpretable*? Or, what is the *intended context*?

Next, the project team consider the actions they have taken that can be referred to as supporting evidence. This evidence provides the inferential support for the higher-level claims.

Finally, once all the pieces are together in a structured manner, the entire case is used as the basis for *justifying* the validity of the top-level goal.

Elements of an Assurance Case

The interpretability example helps us identify a *minimal set* of elements that need to be established in an assurance case:

1. **A top-level normative goal**
2. **Claims about the project or system**
3. **Supporting evidence**

The top-level goal orients and delineates the case by setting a direction and helping to establish the scope of what claims need to be included. For instance, a particular claim about the privacy or security of a project's data management policy many be important but unnecessary to include in an assurance case that justifies why a model does not cause any discriminatory harm.

In addition to any clarificatory claims that pertain to the goal (e.g., what type of model is being assured; the context of use for the system), the lower-level claims should further specify the goal by a) addressing the specific activities that have been carried out during the project, or b) identifying properties of the system that help ensure the goal claim is legitimate. We can, therefore, separate the type of lower-level claim being established as either a project or system property claim.

Unless the claim is self-evidential, there will need to be a final element that points to some supporting evidence. This evidence establishes the foundation upon which the justifiability of the overall argument depends. The following two figures show the relationship between these elements and also provide an example of a partial assurance case that focuses on a goal of safety.

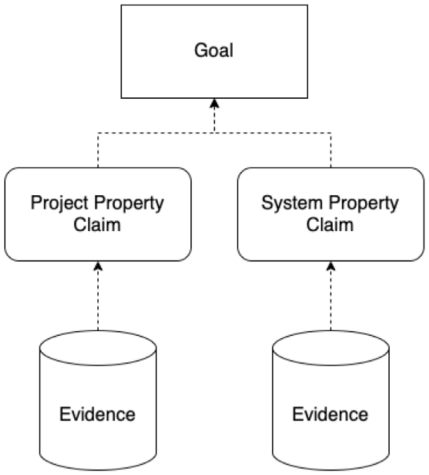


Figure A

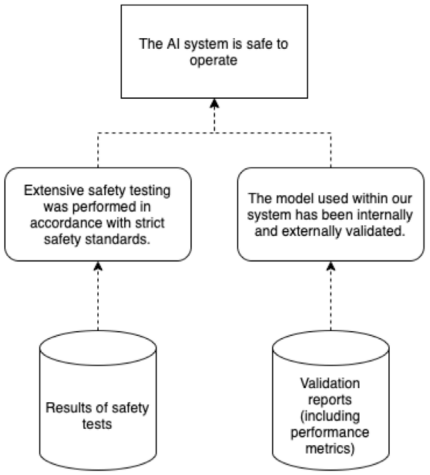


Figure B

Who is the Target Audience?

As you can probably imagine, all of these elements help play a vital role in the effective communication of an assurance case. A clear goal helps signal to stakeholders what values underwrite and motivate your project, as well as providing the means for more critical evaluation and engagement (i.e. an assessment by the stakeholders of whether your goal has been obtained, conditional on the evidence provided). The set of claims collectively establishes the scope and content of your argument, enabling stakeholders to identify whether there are any gaps (i.e. whether your argument is complete). And, the evidential base allows stakeholders to determine whether there is a legitimate reason for accepting your argument.

However, the justifiability and acceptability of an assurance case, in part, depends on the target audience. Sticking with the interpretability example, we can note that what is interpretable to an expert in statistical learning theory may be completely uninterpretable for a policy-maker tasked with evaluating whether a particular model is suitable to deploy in their own project. Therefore, prior to building an assurance case, it is important to identify the target audience and understand their needs. In some cases, this may be determined on the behalf of the project team (e.g. where an external auditor requests assurance for compliance objectives). In other cases, identifying relevant stakeholders may have been performed through stakeholder engagement processes carried out as part of the 'Project Planning' activities.

Reflect, Act, Justify

We now have sufficient background information to present an intuitive (high-level) procedure for developing and communicating an assurance case. The procedure is broken down into three steps, which also complement the goals of the project lifecycle:

1. Reflect
2. Act
3. Justify

Reflection is an anticipatory and deliberative process in which questions such as the following are asked of the project and its governance:

1. What are the goals of your system?
2. How are these goals defined?
3. Which stakeholders have participated in the identification and defining of these goals?
4. What properties need to be implemented in the project or system to ensure that these goals are achieved?
5. Which actions ought to be taken to establish these properties within the project or system?

These are inherently normative and value-laden questions, which is one reason why diverse and inclusive stakeholder engagement is so crucial.

Action occurs throughout all of the stages of the project lifecycle, and the output of many of these actions are likely to serve as the evidence for the claims of the assurance case. These actions and evidential artefacts can also help you identify what claims may be relevant in your argument. As such, the following questions serve only to provide some additional supporting structure to this process:

1. What actions have been undertaken during **(project) design** that have generated salient evidence for your goals and claims?

2. What actions have been undertaken during **(model) development** that have generated salient evidence for your goals and claims?

3. What actions have been undertaken during **(system) deployment** that have generated salient evidence for your goals and claims?

Using the project lifecycle model as a scaffold or guide is, therefore, a useful tool for both a) reflectively planning the necessary activities that ought to be undertaken, prior to the project's commencement, and b) evaluating and assessing whether there are any gaps as a project evolves.

The final step is to **justify** that your evidence base is sufficient to warrant the claims that are being made about the properties of your project or system. This does not mean that the assurance case is the final activity that needs to be done at the very end of a project. Rather, it's development should be seen as iterative and ongoing as the project evolves. Identifying the relevant evidence and determining whether the evidence base is sufficient and complete can be challenging. To help this process, deliberative prompts such as the following can be instructive:

1. Which stakeholders, identified in your stakeholder engagement plan, can support the evaluation of your evidence and overall case?
2. Is any evidence missing from your case?
3. Are the collection of property claims jointly sufficient to support your top-level goal?

However, in general stakeholder engagement—especially with domain experts—is essential.

In the next section, we will discuss some possible orienting goals and principles, to help you identify the sorts of properties that may comprise an ethical and responsible assurance case.

Goals, Properties, and Evidence

In the previous section, we looked at the reflect, act, and justify procedure, which was used to help develop and draw together the constitutive elements of an assurance case. To summarise, A schematic that depicts the reflect, act and justify process. presents a simple illustration of the above process with reference to the three core elements.

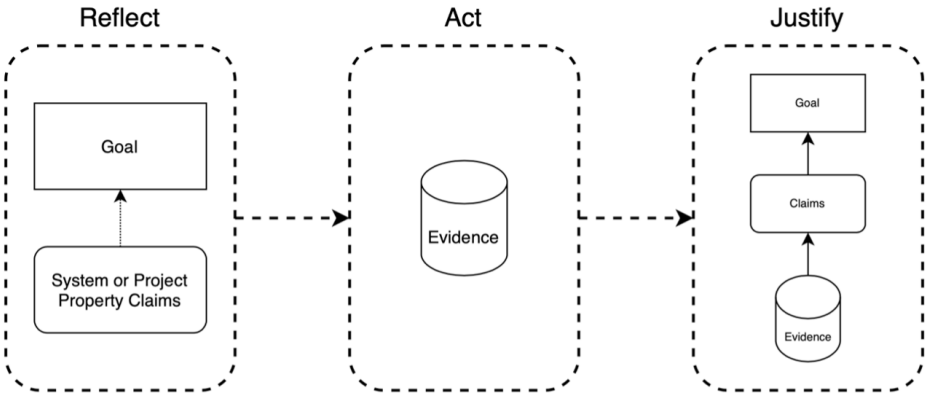


FIG. 18
A schematic that depicts the reflect, act and justify process.

But how do you know which **goals** are relevant for specific projects? How do you use these goals to identify salient **properties**? And, how do you evaluate and determine if you have sufficient **evidence** to support your claims and overall case?

By now, you will probably find it unsurprising to hear that a reflective, inclusive, and socially embedded process of stakeholder engagement is key. Consider again the original definition of responsible research and innovation presented in Chapter 2:

> **"Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)." [Von Schomberg, 2011]**

But even though stakeholder engagement is necessary, a project team still need a springboard to help facilitate communication and deliberation with stakeholders. That is, they need a starting point, grounded in an informed understanding of the potential harms that can arise from the misuse of data science and AI. This is why we introduced the SAFE-D principles in Chapter 3. These principles were not plucked from thin air. They are rooted in and intercon

nected with wide-ranging and multi-disciplinary methods of enquiry that explore such social and ethical harms and benefits (e.g., moral philosophy, human rights law, and science and technology studies), and are also enhanced through diverse modes of knowledge production (e.g., citizen science, investigative journalism, quality assurance, and activism).
As such, the SAFE-D principles represent significant normative goals and valuable starting points for a process of reflection, action and justification. While it is not possible to show how they can be operationalised in every possible data science or AI project, it is possible to expand upon the goals by discussing some of their core attributes.

Sustainability

Core Attribute	Description
Safety	Safety is core to sustainability but goes beyond the mere operational safety of the system. It also includes an understanding of the long-term use context and impact of the system, and the resources needed to ensure the system continues to operate safely over time within its environment (i.e. is sustainable). For instance, safety may depend upon sufficient change monitoring processes that establish whether there has been any substantive drift in the underlying data distributions or social operating environment. Or, it could also involve engaging and involving users and stakeholders in the design and assessment of AI systems that could impact their human rights and fundamental freedoms.
Security	Security encompasses the protection of several operational dimensions of an AI system when confronted with possible adversarial attack. A secure system is capable of maintaining the integrity of its constitutive information. This includes protecting its architecture from the unauthorised modification or damage of any of its component parts. A secure system also remains continuously functional and accessible to its authorised users and keeps confidential and private information secure even under hostile or adversarial conditions.
Robustness	The objective of robustness can be thought of as the goal that an AI system functions reliably and accurately under harsh or uncertain conditions. These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is, therefore, the strength of a system's functional integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, or undesirable reinforcement learning behaviour.
Reliability	The objective of reliability is that an AI system behaves exactly as its designers intended and anticipated. A reliable system adheres to the specifications it was programmed to carry out. Reliability is therefore a measure of consistency and can establish confidence in the safety of a system based upon the dependability with which it conforms to its intended functionality.
Accuracy and Performance	The accuracy of a model is the proportion of examples for which it generates a correct output. This performance measure is also sometimes characterised conversely as an error rate or the fraction of cases for which the model produces an incorrect output. Specifying a reasonable performance level for the system may also require refining or exchanging the measure of accuracy. For instance, if certain errors are more significant or costly than others, a metric for total cost can be integrated into the model so that the cost of one class of errors can be weighed against that of another.

Accountability

Core Attribute	Description
Traceability	Traceability refers to the process by which all stages of the data lifecycle from collection to deployment to system updating or deprovisioning are documented in a way that is accessible and easily understood. This may include not only the parties within the organisation involved but also the actions taken at each stage that may impact the individuals who use the system.
Answerability	Answerability depends upon a human chain of responsibility. Answerability responds to the question of who is accountable for an automation supported outcome.
Auditability	Whereas the property of answerability responds to the question of who is accountable for an automation supported outcome, the notion of auditability answers the question of how the designers and implementers of AI systems are to be held accountable. This aspect of accountability has to do with demonstrating and evidencing both the responsibility of design and use practices and the justifiability of outcomes.
Clear Data Provenance and Lineage	Clear provenance and data lineage consists of records that are accessible and simultaneously detail how data was collected and how it has been used and altered throughout the stages of pre-processing, modelling, training, testing, and deploying.
Accessibility	Accessibility involves ensuring that information about the processes that took place to design, develop, and deploy an AI system are easily accessible by individuals. This not only refers to suitable means of explanation (clear, understandable, and accessible language) but also the mediums for delivery.
Reproducibility	Related to and dependant on the above four properties, reproducibility refers to the ability for others to reproduce the steps you have taken throughout your project to achieve the desired outcomes and where necessary to replicate the same outcomes by following the same procedure.

Fairness

Core Attribute	Description
Bias Mitigation	It is not possible to eliminate bias entirely. However, effective bias mitigation processes can minimise the unwanted and undesirable impact of systematic deviations, distortions, or disparate outcomes that arise to a project governance problem, interfering factor, or from insufficient reflection on historical social or structural discrimination.
Diversity and Inclusiveness	A significant component of fairness aware design is ensuring the inclusion of diverse voices and opinions in the design and development process through the participation of a more representative range of stakeholders. This includes considering whether values of civic participation, inclusion, and diversity been adequately considered in articulating the purpose and setting the goals of the project. Consulting with internal organisational stakeholders is also necessary to strengthen the openness, inclusiveness, and diversity of the project.
Non-Discrimination	A system or model should not create or contribute to circumstances whereby members of protected groups are treated differently or less favourably than other groups because of their respective protected characteristic.
Equality	the outcome or impact of a system should either maintain or promote a state of affairs in which every individual has equal rights and liberties, and equal access or opportunities to whatever good or service the AI system brings about.

Explainability

Core Attribute	Description
Interpretability	Interpretability consists of the ability to know how and why a model performed the way it did in a specific context and, therefore, to understand the rationale behind its decision or behaviour.
Responsible Model Selection	The normal expectations of intelligibility and accessibility that accompany the function of the system, as fulfilled in the sector or domain in which it will operate. This can also necessitate the availability of more interpretable algorithmic models or techniques in cases where the selection of an opaque model poses risks to the physical, psychological, or moral integrity of rights-holders or to their human rights and fundamental freedoms. The availability of the resources and capacity that will be needed to provide responsible, supplementary methods of explanation (e.g. simpler surrogate models, sensitivity analysis, or relative feature important) in cases where an opaque model is deemed appropriate and selected.
Accessible Rationale Explanation	The reasons that led to a decision—especially one that is automated—delivered in an accessible and non-technical way.
Implementation and User Training	Training users to operate the AI system may include: a) conveying basic knowledge about the nature of machine learning, b) explaining the limitations of the system, c) educating users about the risks of AI-related biases, such as decision-automation bias or automation-distrust bias, and d) encouraging users to view the benefits and risks of deploying these systems in terms of their role in helping humans to come to judgements, rather than replacing that judgement.

Data Quality, Integrity, Protection and Privacy

Core Attribute	Description
Source Integrity and Measurement Accuracy	Effective bias mitigation begins at the very commencement of data extraction and collection processes. Both the sources and instruments of measurement may introduce discriminatory factors into a dataset. When incorporated as inputs in the training data, biased prior human decisions and judgments—such as prejudiced scoring, ranking, interview-data or evaluation—will become the 'ground truth' of the model and replicate the bias in the outputs of the system in order to secure discriminatory non-harm, as well as ensuring that the data sample has optimal source integrity. This involves securing or confirming that the data gathering processes involved suitable, reliable, and impartial sources of measurement and sound methods of collection.
Timeliness and Recency	If datasets include outdated data then changes in the underlying data distribution may adversely affect the generalisability of the trained model. Provided these distributional drifts reflect changing social relationship or group dynamics, this loss of accuracy with regard to the actual characteristics of the underlying population may introduce bias into an AI system. In preventing discriminatory outcomes, timeliness and recency of all elements of the data that constitute the datasets must be scrutinised.
Relevance, Appropriateness, and Domain Knowledge	The understanding and utilisation of the most appropriate sources and types of data are crucial for building a robust and unbiased AI system. Solid domain knowledge of the underlying population distribution and of the predictive or classificatory goal of the project is instrumental for choosing optimally relevant measurement inputs that contribute to the reasonable determination of the defined solution. Domain experts should collaborate closely with the technical team to assist in the determination of the optimally appropriate categories and sources of measurement.
Adequacy of Quantity and Quality	This property involves assessing whether the data available is comprehensive enough to address the problem set at hand, as determined by the use case, domain, function, and purpose of the system. Adequate quantity and quality should address sample size, representativeness, and availability of features relevant to problem.
Balance and Representativeness	A balanced and representative dataset is one in which the distribution of features that are included, and the number of samples within each class is similar to the underlying distribution that exists in the overall population.

Attributable	Data should clearly demonstrate who observed and recorded it, when it was observed and recorded, and who it is about.
Consistent, Legible and Accurate	Data should be easy to understand, recorded permanently and original entities should be preserved. Data should be free from errors and conform with the protocol. Consistency includes ensuring data is chronological (e.g., has a date and time stamp that is in the expected sequence).
Complete	All recorded data requires an audit trail to show nothing has been deleted or lost.
Contemporaneous	Data should be recorded as it was observed, and at the time it was executed.
Responsible Data Management	Responsible data management ensures that the team has been trained on how to manage data responsibly and securely, identifying possible risks and threats to the system and assigning roles and responsibilities for how to deal with these risks if they were to occur. Policies on data storage and public dissemination of results should be discussed within the team and with stakeholders, as well as being clearly documented.
Data Traceability and Auditability	Any changes or revisions to the dataset (e.g., additions, augmentations, normalisation) that occur after the original collection should be clearly traceable and well-documented to support any auditing.
Consent (or legitimate basis) for processing	There must be demonstrable grounds that data processing can be carried out on the basis of the free, specific, informed and unambiguous consent of the data subject or of some other legitimate basis laid down by law. The data subject must be informed of risks that could arise in the absence of appropriate safeguards. Such consent must represent the free expression of an intentional choice, given either by a statement (which can be written, including by electronic means, or oral) or by a clear affirmative action and which clearly indicates in this specific context the acceptance of the proposed processing of personal data. Mere silence, inactivity or pre-validated forms or boxes should not, therefore, constitute consent. No undue influence or pressure (which can be of an economic or other nature) whether direct or indirect, may be exercised on the data subject and consent should not be regarded as freely given where the data subject has no genuine or free choice or is unable to refuse or withdraw consent without prejudice. The data subject has the right to withdraw the consent he or she gave at any time (which is to be distinguished from the separate right to object to processing).

Data Security	Each Party shall provide that the controller, and, where applicable the processor, takes appropriate security measures against risks such as accidental or unauthorised access to, destruction, loss, use, modification or disclosure of personal data. Each Party shall provide that the controller notifies, without delay, at least the competent supervisory authority within the meaning of Article 15 of this Convention, of those data breaches which may seriously interfere with the rights and fundamental freedoms of data subjects.
Data Minimisation	Personal data being processed is adequate (sufficient to properly fulfil the stated purpose), relevant (has a rational link to that purpose), and limited to what is necessary do not hold more data than needed for that purpose).
Transparency	The transparency of AI systems can refer to several features, both of their inner workings and behaviours, as well as the systems and processes that support them. An AI system is transparent when it is possible to determine how it was designed, developed, and deployed. This can include, among other things, a record of the data that were used to train the system, or the parameters of the model that transforms the input (e.g., an image) into an output (e.g, a description of the objects in the image). However, it can also refer to wider processes, such as whether there are legal barriers that prevent individuals from accessing information that may be necessary to understand fully how the system functions (e.g., intellectual property restrictions).
Proportionality	delivering a just outcome in ways that are proportionate to the cost, complexity, and resources available. In a similar vein, the term 'proportionality' can also be used as an evaluative notion, such as in the case of a data protection principle that states only personal data that are necessary and adequate for the purposes of the task are collected.
Purpose Limitation	The purposes for data processing must be outlined and documented from the beginning and made available to all individuals through privacy information. Personal data must adhere to the original purpose unless it is compatible with the original purpose, additional consent is received, or there is an obligation or function set out in law.

The above goals and corresponding list of attributes may seem daunting. However, they do not represent a checklist that all research and innovation projects must provide documentation for (e.g. an assurance case). Instead, it is best to think of them as deliberative prompts—a reflect-list rather than a checklist. If, in discussion as a team and in conjunction with stakeholders, it is determined that certain attributes are irrelevant to the project for justifiable reasons, then no activities may be necessary. A principle of proportionality can certianly be adopted here as a meta-principle of sorts, helping direct project governance decisions.

But, if a goal and attribute is deemed important, then the next step is to consider what system or project properties need to be established, and what evidence must be gathered and documented, in order to demonstrate that the top-level goal (and corresponding attributes) have been obtained.

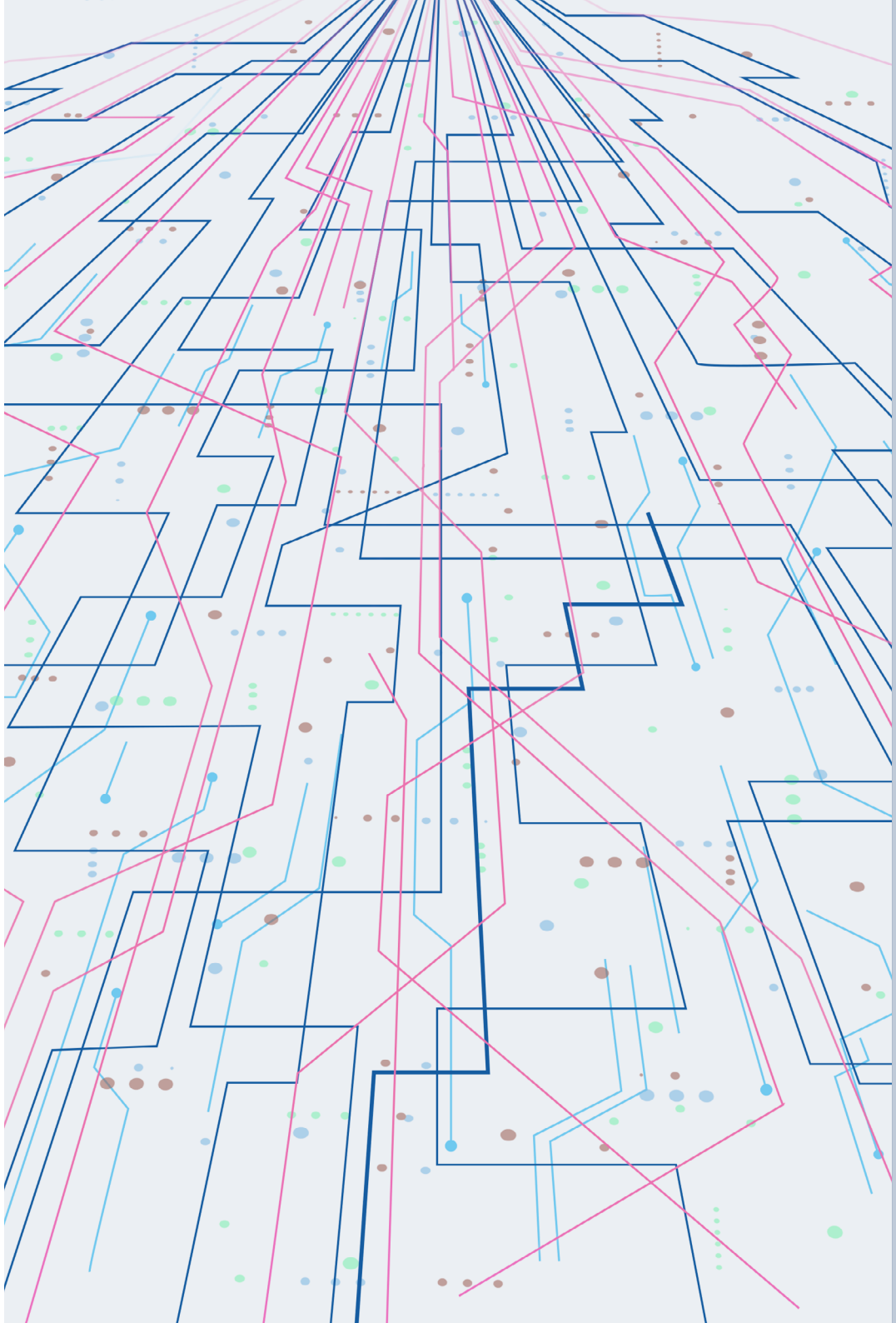
Here, unfortunately, we can only provide some illustrative examples that link a goal and attribute with a possible system or property claim, and also suggest where in the project lifecycle such an activity would be undertaken. As these are examples, no evidential artefact is provided.

Core Attribute	Description	
Sustainability (Robustness)	*The model used in our system has been internally and externally validated. The external validation has been carried out across several varied environments to ensure robustness of the system.*	Model Training, Testing and Validation
Accountability (Accessibility)	*All identified stakeholders were consulted prior to the development of our system to help critically evaluate our project plans and ensure they were intelligible.*	Project Planning and Problem Formulation
Fairness (Equality)	*Persons affected by use of the system have avenues of recourse, ability to contest system outputs and demand human intervention.*	System Use & Monitoring
Explainability (Responsible Model Selection)	*Features were hand-selected in conjunction with domain experts to optimise for both interpretability and predictive power.*	Preprocessing & Feature Engineering and Model
Data Quality (Timeliness & Recency)	*Only data that were collected within the previous 3 months were used to ensure the training data were up-to-date.*	Data Extraction or Procurement

In the next activity, however, we will consider what properties and evidence ought to be included for a hypothetical assurance case for the projects we have been considering.

06

Conclusion (Looking Forward)



Conclusion (Looking Forward)

At the start of this guidebook, several learning objectives were presented. They were to:

- Understand what is meant by the term 'responsible research and innovation', including the motivation and historical context for its increasing relevance.
- Identify and evaluate the ethical issues associated with the key stages of a typical data science or AI project lifecycle: (project) design, (model) development, (system) deployment.
- Explore practical tools and mechanisms for operationalising the concept of 'responsibility' within the context of data science and AI research and innovation.
- Gain an appreciation of shared goals and values across scientific disciplines and research domains through dialogue with other participants.

By now, you should be in a position to determine whether these objectives have been met. You should also be in a better position to critically evaluate the material that has been presented, and to ask how well it meets these objectives. I encourage you to be critical!

There is no shrinking away from the fact that this guidebook, unsurprisingly, has gaps and limitations. As such, there is plenty of room for improvement. This is one of the reasons why it is linked to a GitHub repository. Hopefully, others can contribute to its development and a community of interested participants can help improve the content to make it more useful for future visitors—a true commons for those that care about ensuring data science and AI work for the benefit of society. There is one limitation that is worth highlighting though.

At the time of writing this (November 2021), there is a lot of research underway in an area

known as 'intercultural ethics' [Evanoff, 2020], with a specific focus on how it applies to data science and AI. [1] Much of this research explores how different normative values are represented and accommodated—or, conversely poorly represented and accommodated—in the design, development, and deployment of data-driven technologies. Because many technological systems or models are deployed in global and multicultural environments, there is a significant concern about the extent to which those values embedded into a system by its developers come into tension or conflict with different communities across the globe. Moreover, as these technologies are "data-driven", there are also concerns about the degree to which marginalised or vulnerable communities across the globe are empowered to use or take control of how their data is managed. These issues are not well represented in the current version of this guidebook.

One reason for this is simply that there are two other guidebooks—one on AI Ethics & Governance, and one on Public Engagement and Communication of Science—for which these topics are more directly relevant. However, while these guidebooks are better suited to dealing with these topics directly, that does not preclude this guidebook from integrating some of the lessons about intercultural ethics or data justice into the current chapters. Therefore, subsequent iterations of this guidebook will aim to engage this literature more directly. In the meantime, the 'Further Reading' section has some starting points and references for those who may be interested.

[1] See the special issue by [Aggarwal, 2020].

Taking Responsibility

A core focus of this guidebook has been the idea that science and technology are inextricably interconnected with society, and help shape its norms and practices. In this context, the anticipatory and reflective elements of UKRI's AREA framework, are about looking forward into the future and trying to take responsibility for the impact of the research or innovation that you have some direct control over. Doing this, presupposes a vital ethical value that is overlooked by the AREA framework: an attitude or disposition to care for others and the society in which you arte situated. As Leslie [2020] notes, this changes the AREA framework into a CARE and Act framework.

Taking responsibility, therefore, is a reflection of your values and a reflection of what you choose to care about. By taking responsibility for your research and innovation you are helping to care for society and the future we will share.

The Alan Turing Institute